Developing Speech Rhythm Analysis for Forensic Voice Comparison

Luke Adam Carroll
BA (Hons), MA

Submitted for the degree of
Doctor of Philosophy

Department of Linguistics
and English Language

Lancaster University

March 2024

Abstract

Forensic voice comparison (FVC) involves the comparison of a criminal recording (e.g., a threatening phone call), and a known suspect sample (e.g., a police interview). It is the role of an expert forensic analyst to advise the trier of fact (e.g., judge or jury) on the likelihood that the two samples include the same or different speakers. To do this, the expert will carry out an assessment of the similarity of the speech characteristics in the criminal recording and the suspect sample.

Speech rhythm has been proposed as a feature that could contribute to FVC, but there is not yet a structured analysis framework that practitioners can exploit in forensic casework. When an analyst suspects a speaker's speech rhythm is relevant to an analysis, it is usually only described at an impressionistic level.

Using both production and perception experiments, the present research explores whether there are acoustic and auditory cues that could capture speech rhythm and subsequently be used to discriminate between speakers in forensic casework. The production experiments revealed that there was very little discriminatory power in syllabic duration, intensity and f_0 measurements across spontaneous, content-mismatched utterances. However, there does appear to be some speaker discriminatory value in applying these same measurements to, so-called, "frequently occurring speech units" (e.g., "*er*", "*erm*", "*yes*" and "*no*").

The perception experiments aimed to determine whether listeners (expert and non-expert) can make meaningful speaker identification assessments when presented with delexicalised speech samples that foreground the rhythmic attributes of speech. Results revealed that expert listeners were better than non-expert listeners in making correct speaker identification assessments, with those who had expertise in forensic phonetics generally performing better than those who did not.

The findings from these experiments give promise to the prospect of developing a perceptual (auditory) rhythm framework which can used in forensic casework.

Contents

List of Tables

List of Figures

Acknowledgements

First and foremost, I am eternally grateful to my supervisor, George Brown, for her wisdom, advice and encouragement over the last five and a half years. Words quite simply do not do justice to how truly thankful I am to have had you by my side throughout this PhD odyssey. Your patience, understanding and kindness really means the world to me, and I am so very lucky to have the opportunity to continue working with such a lovely human. Without you, I would not have reached this milestone, a milestone which I have long sought after, and for which I am truly proud of (eventually) reaching – thank you for being the bestest 'boss' and an ace friend.

Secondly, I am also extremely thankful to Christin Kirchhübel for sharing her forensic expertise with me over the past five years. The time you have selflessly put aside to enlighten me to the day-to-day workings of a forensic practitioner, along with the invaluable training you have provided, has been instrumental in both the direction and completion of this thesis. Thank you for your maintained belief in me and for providing me with the opportunity to develop a career doing something that I am truly passionate about. I know how lucky I am to have you as my mentor, and I will strive to make you proud as we move forward.

I would also like to extend my gratitude to all of the participants who took part in the perception experiments for this thesis, and to the Department of Linguistics and English Language for making it possible to offer a financial incentive to participate.

A sincere thanks is offered to the friends and colleagues I have met within the Department of Linguistics and English Language, especially those within the Phonetics Lab group. Being able to share the trials and tribulations of PhD-life with you all has helped in keeping me motivated to get the bloody thing finished. A special shout-out here to Lois Fairclough for her unrelenting encouragement over the years, but mainly thanks for being my partner in nonsensical ramblings, balance board spinnings, stress-busting shoulder massages, maniacal laughter, and general hilarity. Fanks, love.

Also worthy of mention here are all of the friends I have met since moving up to Lancaster in 2018. Whether these be friends I have met through study, through working in Grad Bar, through working in Barkers, or throughout my time playing with the mighty Grad F.C., a big fat cheers to you all.

Last but not least, special thanks to my Dad and Lisa for constantly supporting my study and putting up with me being a perpetual student bum. If I don't make it as a rock and roll star when I grow up, at least I've got a doctorate to fall back on now.

This research was supported by the Economic and Social Research Council North West Social Science Doctoral Training Partnership as part of a 1+3 studentship award in the Linguistics pathway.

Declaration

I declare that this thesis is my own work and has not been submitted elsewhere for the award of any other degree.

Luke Adam Carroll

04/03/2024

CHAPTER 1

Introduction

Forensic voice comparison (FVC) is the primary task carried out by forensic practitioners (Foulkes & French, 2012; French et al., 2017). The task often entails comparing a criminal recording, such as a threatening phone call, with a known suspect sample, such as a police interview. The responsibility of an expert forensic analyst is to provide guidance to the trier of fact, whether it be a judge or jury, regarding the likelihood of the two samples originating from the same speaker or different speakers. The methods which are used to achieve this goal are subject to variation amongst experts, however, in a survey of international practices in FVC conducted by Gold and French (2011), 64.9% of respondents (23/34) reported that they used the auditory-phonetic and acoustic approach. Other approaches include analysts making use of automatic speaker recognition systems (usually alongside some degree of human analysis), with the use of such systems being shown to be on the increase (Gold & French, 2019). Nevertheless, it remains that the auditory-phonetic and acoustic approach is still the most widely used.

In this approach, the assessment of voice and speech characteristics involves both auditory judgments and acoustic analysis. Whilst some voice and speech features are predominantly analysed on an auditory level (e.g., voice quality), others are mainly assessed through acoustic measurement (e.g., fundamental frequency (f_0) – the acoustic correlate of a speaker's voice pitch). It is worth noting that certain voice and speech characteristics are examined through both auditory and acoustic analysis. For instance, vowel realisations involve an auditory assessment and description of

vowel quality, and acoustic measurements of vowel formants. As such, the auditory-phonetic and acoustic approach aims to comprehensively analyse the voice and speech patterns of the individuals being compared by considering a wide range of linguistic parameters.

The parameters chosen for analysis are determined individually on a case-by-case basis, with ideal features typically exhibiting low intra-speaker variability and high inter-speaker variability (Rose, 2002, p.10). Features which are commonly analysed include segmental features (e.g., vowels and consonants), suprasegmental features (e.g., fundamental frequency, voice quality, intonation, etc.), non-linguistic features (e.g., filled pauses, tongue clicking, etc.), discourse features, and/or conversational behaviours (e.g., discourse markers, opening and closing behaviours, etc.), as well as lexical, grammatical, and morphological features (for further detail on speaker characteristics frequently used by practitioners, see Jessen, 2018, pp. 227-229). The selected features are compared across the recordings, and an evaluation is made regarding the overall similarities and differences between them, taking into account that there is expected to be variation within an individual. That is, it will never be possible to achieve a perfect 'match' between samples as every utterance possesses its own unique and intricate details (e.g., Rose, 1996).

For certain features, there are commonly agreed-upon casework practices amongst forensic experts such as measuring fundamental frequency and measuring vowel formants. Consequently, these are the features that have received the majority of attention in the forensic phonetics literature, coupled with the fact that features such as these are comparatively ‘easy’ to measure. However, there are other features for which there are no formalised analysis practices and/or frameworks in the context of forensic casework, and for which there is a lacuna in the research literature, despite their potential in aiding speaker discrimination. Speech rhythm is one such feature.

It is, however, perhaps not surprising that speech rhythm is one of the features that remains comparatively under researched within the forensic domain and for which there is presently no common structured methodology used by practitioners for its analysis. This is owing to the notoriously complex nature of speech rhythm in that it is typically regarded as a manifestation of a range of different speech parameters that

overlap and interact with each other (e.g., duration, intensity, f_0, etc.). Given the numerous speech components and their complex interrelationships, it comes as no surprise that arriving at a single concrete definition of speech rhythm is somewhat problematic. However, what most definitions of speech rhythm do incorporate is the concept of their being some form of perceived regularity in relation to prominent speech units (e.g., syllables), with the perceived prominence of these units being attributed to the different parameters and their interrelationships (e.g., duration, intensity, f_0, etc.). In terms regularity, for a stress-timed language such as English, this relates to the perception that these prominent (or stressed) syllables occur at regular intervals from one another. This idea of regularity, or *isochrony* as it is often referred to within the speech rhythm literature, is, however, something which a wealth of empirical speech rhythm research has failed to substantiate. As such, arguments have been put forward that any such regularity attributed to speech rhythm is perceptual as opposed to being measurable regularities within the speech signal (see Chapter 2 for detailed discussion pertaining to the history of speech rhythm research along with further elaboration pertaining to the intricacies of defining speech rhythm).

Returning to the focus of the present thesis, irrespective of whether speech rhythm is governed by regularities (perceptual or not), it stands that the analysis of speech rhythm has potential in assisting within speaker discrimination tasks. Previous research has demonstrated that the three main parameters associated with speech rhythm – duration, intensity and f_0 – all have the capacity (to greater or lesser extents) at distinguishing between speakers (e.g., He & Dellwo, 2017; Leemann et al., 2014; Lindh & Eriksson, 2007; Zhang et al., 2021; see Chapter 2, Sections 2.5.1 – 2.5.3 for further discussion). That is, the findings from such research indicate that speakers may exhibit idiosyncratic behaviour in relation to their speech rhythm patterns (n.b., the terms ‘speech rhythm’, ‘speech rhythm patterns’, ‘speech patterns’ and ‘rhythm patterns’ are used interchangeably throughout this thesis). Additionally, the analysis of speech rhythm within the forensic context carries potential as some rhythmic attributes (such as durational characteristics) are realised in the temporal domain as opposed to the spectral domain. Where speech features which are realised in the spectral domain (such as vowel formant frequencies) will be affected by speech material which is degraded in quality (as the majority of forensic material is), features

realised in the temporal domain are typically less affected by such issues. Furthermore, the analysis of speech rhythm patterns in relation to intensity characteristics could hold forensic potential given that intensity patterns may not be as easily manipulated by speakers (e.g., as a disguise strategy) due to lack of possible auditory feedback as opposed to other features such as voice quality and vowel pronunciations (see Chapter 2, Section 2.5.2 for further discussion on the speaker discriminatory potential of intensity patterns).

In consideration of the forensic potential which the analysis of speech rhythm may have to offer, the present thesis explores whether there are acoustic and auditory cues that could capture spontaneous speech rhythm patterns and whether these can subsequently be used to discriminate between speakers in forensic casework. The remainder of this chapter is dedicated to highlighting the discrepancies that exist between forensic phonetics research and FVC practice, the purpose of which being to demonstrate the steps taken by the present work to alleviate this research-practice disparity.

1.1. Forensic voice comparison: research vs. practice

Forensic speech science research does not always neatly align with the practical realities of FVC. Vowel formants provide a fitting example. Gold and French (2011), in their survey of international practices in FVC, found that 97% of experts conducted some form of vowel formant analysis within FVC casework, with this likely being a contributing factor to this feature being readily studied academically. Fairclough et al. (2023) carried out a meta-analysis which reviewed the performance of formants for FVC and in doing so found over 100 forensic speech science research papers which focused on the analysis of vowel formants. Focusing on 37 studies in particular (for which there were 277 results), Fairclough et al.'s meta-analysis tested vowel formants as a parameter for speaker discrimination. They found that the results across the studies were highly variable and that some expected performance trends were not evidenced through their analysis. In discussing their findings, Fairclough et al. draw attention to three key analytical techniques used to examine vowel formants across

the studies: measuring the midpoint of the vowel, taking dynamic measurements across the trajectory of the vowel, and long-term formant analysis. Fairclough et al.'s meta-analytical approach revealed that relatively few studies measured the midpoint of the vowels for speaker discriminatory purposes, despite results showing midpoint measurements performed better than dynamic measurements and long-term formant analysis.

More surprising here, however, is that the relative scarcity in studies analysing vowel midpoints runs contrary to the practices of forensic practitioners within FVC casework, as Gold and French (2011) report that 94% of experts analyse vowels in terms of their midpoints. Fairclough et al. take this opportunity to further highlight the disparities between forensic phonetics research and FVC practice by reporting that the majority of the data used across the studies was laboratory-based speech which is not representative of the speech material found within FVC casework. The analysis of vowel formants is not a pertinent part of the present thesis, however, the observations made by Fairclough and colleagues highlight two key areas in which forensic speech science research is unsatisfactory, and which the present work aims to bolster.

The first key area of disparity is that research has focussed on aspects of analysis which are not necessarily implemented in casework practice. With regards to the findings obtained by Fairclough et al., this corresponds to the proportion of work being directed towards dynamic formant measurements as opposed to midpoint measurements. A further example can be found with regards to the analysis of voice quality. For example, within the auditory-phonetic and acoustic approach to FVC, it is unlikely that, when analysing aspects of a speaker's voice quality, an expert will conduct an acoustic analysis relating to spectral tilt and additive noise parameters (the acoustic correlates associated with voice quality). Reasons for this include that some of these voice quality parameters have been shown to be highly sensitive to degradations in recording quality (e.g., Kakouros et al., 2018) which is commonly found in FVC material. Instead, the forensic analyst will examine a speaker's voice quality from an auditory perspective taking into account acoustic observations (not measurements) where possible. This auditory analysis will often be carried out using a recognised methodological approach such as the Vocal Profile Analysis (VPA)

scheme (Laver, 1980). The VPA scheme is one which has been subjected to modifications through forensic research which has sought to optimise its use within FVC casework (e.g., San Segundo et al., 2019; San Segundo & Mompean, 2017; San Segundo & Skarnitzl, 2021). Such research serves as a useful exemplar as to how FVC practices can be effectively targeted through empirical study.

This is not to say that research which considers vowel formant measurements or the acoustics of voice quality for FVC should be abandoned, as results from such studies contribute to our understanding of voices (e.g., Chan, 2023; Hughes et al., 2019). Rather, it is a point of proportionality and a message of encouragement to not shy away from those research efforts which are more likely to have tangible outputs, and which can be directly applied to support FVC tasks.

A further area of research-practice disparity, directly corresponding to the above, is that research often does not incorporate aspects of analysis that are pertinent in casework. Ensuring that the methodologies and analytical procedures used in research are readily accessible, practically feasible, and geared towards the types of analysis which are commonplace within FVC casework is therefore a key consideration. The acoustic analysis of vowel formants was shown to be one area to which a good deal of forensic phonetics research has been directed. Although most experts will take acoustic measurements of formants as part of their overall analysis of a speaker's vowels (provided the material is suitable for formant analysis), it remains that this is only one part of the vowel analysis (and, of course, it is only a very small part of the overall FVC analysis). Importantly, the estimated numerical values that are produced through acoustic analysis do not exist in isolation but are always interpreted by the analyst with reference to the context from which they were taken (e.g., quality and comparability of the data). As such, it is the analyst who determines the probative value of the acoustic analysis of vowel formants per se; further, it is the analyst who evaluates the probative value of any similarities and/or differences observed in the comparison of these acoustic values. Research rarely incorporates 'the analyst' into the methodological design. Instead of just considering the acoustic values in isolation, research which considers findings from the acoustic analysis in tandem with the analyst's perceptual evaluations would take us closer to casework reality. It is

therefore important to carefully consider the importance placed on the acoustic analysis of vowels within research, with an expert's perceptual examination potentially being a more appropriate modus operandi. It is also worthy of note here that generally within FVC casework the analysis of the first vowel formant (F_1) and second vowel formant (F_2) rarely has probative value on its own. If there is a forensically significant difference in these values between recordings, then this is apparent from auditory analysis – that is, there will be an auditory difference in vowel quality. Although the acoustic analysis of F_1 and F_2 can be beneficial with regards to furthering what auditory analysis can achieve for controlled laboratory recordings, this is generally not the case for the recordings found within FVC casework due to the lack of comparability.

Returning to one of the areas of research-practice discrepancy highlighted by Fairclough et al. (2023) above, the vast majority of forensic speech research has used data which is not comparable to casework data. Research which makes use of content-controlled (e.g., read), laboratory-based speech is unlikely to yield results which are relevant to the spontaneous, content-mismatched speech material found in FVC casework. Within forensic casework, the speech material which practitioners must analyse will frequently be of suboptimal quality. Whether this be due to factors such as poor recording quality, signal degradation through telephone transmission, background noise, multiple speakers and overlapping speech, emotion-afflicted speech (e.g., shouting, screaming, etc.), or limited sample duration (amongst others), the unfavourable quality of the speech material will likely render it unsuitable for applying the methodologies and analytical techniques which laboratory-based research have focussed on. For forensic research to truly be as beneficial as possible for FVC practice, it should be founded on speech data which is characteristic of the speech material which forensic analysts have to work with. Unfortunately, however, gaining access to real-life speech data which is representative of that found in forensic casework is problematic as such data is simply not available to researchers due to ethical constraints (other than where the lawful interception of speech data is permitted, and access is subsequently granted for use within research). One way in which researchers have looked to alleviate this issue is through the development of forensically relevant speech databases (e.g., DyViS (Nolan et al., 2009); WYRED

(Gold et al., 2018)). These databases have been specifically designed to elicit speech which is more similar to that which is found in forensic recordings, such as a suspect being interviewed whilst in custody, a suspect in telephone conversation with an accomplice, and a suspect leaving an incriminating voicemail message. Making use of the spontaneous (or semi-spontaneous) speech material afforded by databases such as these is one way in which forensic phonetics research can be more targeted towards being applicable to FVC practice. Overall, whilst it may be a good starting point to use controlled data for forensic research, it is important that such research should subsequently be extended to more realistic data. In relation to the present thesis and its focus on speech rhythm, previous research which has examined speech rhythm features in relation to their speaker discriminatory potential has largely focussed on controlled speech data (e.g., Dellwo & Koreman, 2008; He and Dellwo, 2016; Leemann et al., 2014). With such studies indicating that the analysis of speech rhythm features could be of use in speaker discrimination tasks, determining the applicability of measuring speech rhythm parameters (and using rhythm metrics) when forensically realistic data is involved is an obvious next step – a step in which the present work sets out to take.

The final area of research-practice disparity highlighted above is that there are some speech features analysed within FVC that are significantly under researched in comparison to others. As has been demonstrated above, vowels have been subjected to a vast quantity of forensic research. On the other hand, there are some features that have been overlooked within the literature, despite carrying potential for contributing to FVC. It is therefore self-evident that directing focus towards such features will be beneficial for FVC practice.

The discussion provided in this section serves to illustrate that regardless of the amount of research carried out on a specific feature, there will always be limitations and potential pitfalls which arise. This discussion also goes to show that where an acoustic approach to analysis may reach an impasse, an auditory approach may provide a preferable option. With this in mind, it is apparent that the development of methodologies which are focussed on the perceptual analysis of speech would be of benefit to FVC practices. It is acknowledged, however, that strict adherence to all of

these idealisms could be somewhat prohibitive in relation to the advancement of forensic speech research as a whole. All forensic phonetics research which furthers our understanding of the speaker discriminatory potential and robustness of various speech features is, of course, valuable and welcomed. Nevertheless, it would be pleasing to see more research emerging within the forensic field which has direct applicability to the tasks carried out within FVC casework.

1.2. Research contributions

The research presented in the present work looks to contribute to FVC practice by specifically focussing on the aforementioned factors which were highlighted as being beneficial for FVC casework. As such, this thesis is committed to focus on:

(1) A comparatively under researched speech feature within the forensic domain: *speech rhythm*. Whilst vowel formants and other speech features have been subjected to a vast quantity of forensic research, speech rhythm has been overlooked within the literature, despite being proposed as a feature that could contribute to FVC (e.g., in Gold and French's (2011) survey, 73% of the experts stated that they examine speech rhythm with varying regularity). Within the auditory-phonetic and acoustic approach to FVC, there is currently no structured analysis framework practitioners can use to effectively account for speakers' speech rhythm patterns. When an analyst suspects a speaker's speech rhythm is relevant to an analysis, it is usually only described at an impressionistic level. That is, in the absence of any formalised practice/framework to analyse speech rhythm, any relevant observations will likely be documented in the form of a short descriptive summary which will then be incorporated into their final report. Using both production and perception experiments, the present research explores whether there are acoustic and auditory cues that could capture speech rhythm and subsequently be used to discriminate between speakers in forensic casework. The findings from these experiments will inform the development of a perceptual framework for speech rhythm analysis, providing the forensic practitioner with a structured analytical framework to assist in making their auditory (impressionistic) judgements (see (4) below).

(2) *Spontaneous speech* from a *forensic phonetics database*. The majority of previous forensically-motivated speech rhythm research has made use of controlled, read speech. Adopting and extending a variety of approaches used in previous research and transferring these to spontaneous speech will be of obvious interest to the forensic cause. Using mock police interview speech data from the WYRED corpus (Gold et al., 2018), two production experiments are carried out. The first analyses spontaneous speech utterance data and the second extends this analysis to a group of, so-called, "frequently occurring speech units" which are hypothesised as potentially bolstering the effectiveness of quantifying spontaneous speech patterns (see Chapter 2, Section 2.6). These experiments focus on a range of speech parameters – intensity, f_0 and duration – the three parameters most commonly attributed to speech rhythm. These parameters are all measured both individually and in combination to provide a comprehensive evaluation of their comparative usefulness in assessing variation between speakers and their ability to distinguish between speakers.

The perception experiments make use of incriminating voicemail data from the WYRED corpus. From these data, delexicalised speech samples, which foreground the rhythmic attributes of speech, are created and presented to listeners in a series of discrimination tasks. These experiments will therefore shed light on the extent to which speech rhythm is useful for the purpose of speaker discrimination.

(3) Incorporating the perceptual evaluations of *forensic experts* into the research findings. For the main perception experiment, *expert listeners* are consulted, and their assessments of speech rhythm patterns play a key role in evaluating the extent to which speech rhythm patterns can be used to discriminate between speakers. *Forensic practitioners*, *forensic phonetics researchers*, *forensic phonetics research students* as well as other *experienced phoneticians* are consulted to give a comprehensive overview as to how these experienced groups of listeners assess speech rhythm patterns within a speaker discrimination context. Of the forensic experts consulted, many have firsthand experience in FVC casework, with their contributions further strengthening the link between research and practice for the current work.

(4) *Method development for use within FVC practice.* The present thesis uses the results obtained from the aforementioned experiments to propose a perceptual rhythm

framework that analysts can use to more effectively account for speakers' speech rhythm patterns within FVC. In particular, the evaluations of *forensic experts* are incorporated into the assessment framework design.

1.3. Research aims

In order to achieve the goals presented above, this thesis is broken down into the following research objectives:

1. Measure spontaneous speech rhythm patterns using a range of parameters and measurement techniques and assess the extent to which these patterns are speaker-specific.

2. Measure the rhythmic properties of selected frequently occurring speech units and assess the extent to which these could be useful for discriminating between speakers.

3. Determine the extent to which listeners are able to distinguish between speakers based solely on attributes of speech rhythm.

4. Develop a perceptual framework for the analysis of speech rhythm which can be applied within the context of forensic voice comparison tasks.

1.4. Thesis outline

As a means of accomplishing the research aims outlined above, this thesis is composed as follows:

• In Chapter 2, an overview is presented of the existing literature relating to speech rhythm research. Particular focus is cast upon forensically-motivated rhythm research and research which has accounted for the three most relevant rhythm parameters: intensity, f_0 and duration (i.e., the components most commonly attributed to contributing to speech rhythm). A review of the literature relating to the frequently occurring speech units which are the focus of Chapter 4 is also provided.

• In Chapter 3, the spontaneous speech rhythm patterns of a group of homogeneous speakers are analysed in terms of their intensity, f_0 and durational characteristics. A number of different measurements and quantification metrics are used to assess the rhythmic variation exhibited between speakers. Statistical analysis of the data is performed to examine the speaker-specificity of the rhythm patterns as well as to determine which parameters and measures carry the most speaker discriminatory potential.

• In Chapter 4, the rhythmic characteristics of four frequently occurring speech units are analysed in terms of their intensity, f_0 and durational characteristics. These speech units are normalised against the spontaneous speech data presented in Chapter 3 in order to capture the rhythmic characteristics of these specific units relative to the spontaneous speech patterns. Statistical analysis is carried out to test which speech units and which parameters possess the most speaker discriminatory power (n.b., within the present thesis the terms "frequently occurring speech units" and "speech units" are used interchangeably and refer specifically to the four monosyllabic units being analysed unless otherwise stated).

• In Chapter 5, perception experiments are carried out to determine to what extent listeners are able to discriminate between speakers based on just speech rhythm characteristics. Speech samples were subjected to delexicalisation which foregrounded the rhythmic characteristics. These delexicalised samples were presented to both expert and non-expert listeners in online perception experiments. The experiments consisted of three sections. In Section One and section Two, participants were required to make a binary decision as to which delexicalised samples contained the same speaker as the original (non-delexicalised) samples whilst, for Section Two, also providing qualitative feedback. In Section Three, listeners had to rate the similarity of pairs of delexicalised speech samples on a nine-point Likert scale from very similar (1) to very different (9).

• In Chapter 6, off the back of the results obtained from the previous chapters, in particular the qualitative feedback from the perception experiments, a framework for

the assessment of speech rhythm within the forensic context is proposed. Current forensic frameworks are presented in the first instance to illustrate the processes involved in developing and testing a framework in preparation for forensic application.

- In Chapter 7, a summary of the thesis is provided, and further discussion is provided in relation to the opportunities for future research. The final conclusions of the thesis are then drawn.

CHAPTER 2

Literature Review

2.1. Introduction

This thesis is founded upon both production and perception experiments of speech rhythm with the goal of having implications and applications within the forensic domain. This chapter therefore provides an overview of the literature surrounding speech rhythm. Following this, attention is turned to the forensic implications of the present work and provides an overview of the existing forensically-motivated speech rhythm research. This will highlight the void which this thesis intends to occupy in investigating speech rhythm production and perception for forensic applications.

2.2. What is speech rhythm?

Speech rhythm is one of the most difficult elements of speech to both describe and quantify (Lloyd James, no date, p.11). This observation can be ascribed to the idea that speech rhythm is typically regarded as a manifestation of several distinct speech parameters that overlap and interact with one another in complex ways. Nonetheless, speech rhythm research is still a prevalent field of study, with efforts towards arriving at agreeable definitions of speech rhythm and means of capturing speech rhythm patterns being far from exhausted. In the context of rhythm typology, work continues to further understand the rhythmic properties and patterns found across different linguistic varieties (Arvaniti, 2009; Dauer, 1983; Fuchs, 2016; Krivokapić, 2013; Liu & Takeda, 2021; Mok, 2009; Nespor, 1990). Moreover, establishing reliable ways of

measuring speech rhythm and having concrete definitions of different types of speech rhythm will have a number of practical uses.

For example, from a forensic point of view, having a robust framework with which speech rhythm can be quantified would allow the forensic analyst to make reliable comparisons of the speech rhythm between unknown speakers and suspects who feature in forensic recordings. Indeed, this forensic application is the focus of the present study in which individual speaker variation in speech rhythm is examined. However, describing and quantifying speech rhythm remains a perpetual challenge for all those seeking to do so, and this can be attributed to the fact that it is often understood as being a manifestation of a number of different speech parameters overlaying and interacting with one another (e.g., see Handel, 1993). The parameters usually associated with contributing towards perceived speech rhythm include:

- **Speech tempo:** the durations of speech units (e.g., syllables) which can vary from short to long.
- **Pitch:** the fundamental frequency (f_0) of the voice which can vary between low and high.
- **Intonation:** the variation in pitch across stretches of speech such as rising or falling pitch.
- **Loudness:** the intensity or vocal effort of the voice which can vary between loud and soft.
- **Stress:** the prominence of speech units (e.g., syllables), with this prominence commonly being attributed to variations in:

    - Duration (typically longer duration) and/or
    - Pitch or a pitch movement within a syllable and/or
    - Loudness (typically increased intensity).
    - Full vowel vs centralised vowel (i.e., schwa).

The final of these parameters, ‘stress’, which is generally accepted as describing a speech unit (e.g., a stressed syllable) which perceptually ‘stands out’ in comparison to neighbouring units, can be seen to carry its own complexities given that it could be a

combination of acoustic cues which determine a unit's prominence. Furthermore, these acoustic cues will also likely be contributing to the prominence to varying degrees, with certain features carrying more weight over others, and with this weighting also potentially varying even over a singular utterance. Stress is yet further complicated in that it is also conceptualised as either corresponding to *lexical stress* or *prosodic stress* (e.g., see Beckman & Pierrehumbert, 1986), with the former relating to an abstract word-level phonological property in some languages (e.g., English), and the latter relating to the intentional emphasis placed on a word (e.g., to highlight its importance).

In consideration of all the above speech components, and the potentially complex interrelations between them, it is hardly surprising that there is not one conclusive answer to the question: *'what is speech rhythm?'*. Nevertheless, definitions as to what constitutes speech rhythm have been put forward by linguistic scholars, such as the following:

> *Rhythm:* An application of the general sense of this term in phonology, to refer to the perceived regularity of prominent units in speech. These regularities may be stated in terms of patterns of stressed v. unstressed syllables, syllable length (long v. short) or pitch (high v. low), or some combination of these variables.
>
> (Crystal, 1985: 266-67)

Other definitions generally follow along the same lines of the above (e.g., Laver, 1994, p. 527; Trask, 2006, p. 311), but what most, if not all, definitions tend to point towards is the notion of there being some form of perceived *regularity* associated with speech rhythm, whether this be in terms of duration, pitch or intensity. So, it would appear that if speech *is* in some way rhythmic, that we should be able to substantiate these claims of regularity, and indeed that is what a great deal of research spanning over the last 80 years has attempted to do.

2.3. Review of speech rhythm research

As the purpose of this thesis is to determine whether speech rhythm patterns can be measured and described for the purpose of speaker discrimination, it will of course be necessary to highlight previous speech rhythm research which has also had a similar focus. As will be seen, such forensically-motivated research is far from abundant within the literature, nevertheless, the final sections of this chapter will focus on this body of work. Before this though, the initial following subsections will provide a brief chronological review of speech rhythm research from over the last 80 years. This initial review of the literature will provide a useful backdrop as to how speech rhythm research and its associated methodologies evolved over the decades, ultimately resulting in the various methods being used in the present day and informing the methods used in the present thesis.

2.3.1. Rhythm typology and the quest for isochrony

Much of the early research into speech rhythm, starting around the 1940s, focussed on trying to establish regularities in relation to one particular parameter: *duration*. Such was this focus on duration (that is, the relative duration of speech units such as syllables), that the concept of speech being *isochronous* – being governed in some way by durational regularities – became widely accepted, subsequently establishing a new area of study known nowadays as *rhythm typology* (e.g., Abercrombie, 1967; Lloyd James, 1940; Pike, 1945).

The field of rhythm typology is founded on the notion that there are different types of speech rhythm and that these different types are associated with specific languages. The earliest descriptions of these different types of rhythm were made by Lloyd James (1940) who differentiated between ‘machine-gun’ and ‘Morse-code’ speech rhythm varieties, with these terms later being displaced by Pike’s (1945) proposition of the terms ‘syllable-timed’ and ‘stressed-timed’ to describe these two contrasting rhythmic types. At this early juncture in speech rhythm research, these two types of speech rhythm were not explicitly associated with any particular language or languages, that is, rhythm typology was yet to be formally applied.

It was Abercrombie, some 22 years later in 1967, who embraced Pike's terminology of 'syllable-timed' and 'stressed-timed' speech rhythms and further asserted that all of the world's language could be classified under one or the other of the two terms. Further still, he asserted that both of these rhythm types were underpinned by isochronous intervals. Giving examples of which languages were classified under the two different rhythmic classes, Abercrombie cited French, Telugu and Yoruba as being syllable-timed and English, Russian and Arabic as being stress-timed, further stating that a language either belongs to 'one or the other type of rhythm but not both since the two types are incompatible' (Abercrombie 1971; reported in Adams, 1979, p. 52).

It was the work of Abercrombie around this time which is acclaimed as being the inception of the rhythm typology movement. With Abercrombie focussed solely on the distinction between syllable-timed and stressed-timed rhythmic types, others such as Bloch (1950) and Trubetzkoy (1958) had earlier highlighted how Japanese rhythm also displayed isochronous tendencies in relation to the repetition of morae (a minimal unit of metrical time equivalent to a short syllable (Hoequist, 1983)), however, any such 'mora-timed' distinction was not formally acknowledged under Abercrombie's rhythm typology at this time. It was only later in the 1980s and 1990s that rhythm typology researchers proposed the addition of mora-timed languages. In summary, and for the sake of completeness, including the later proclaimed mora-timed distinction, rhythm typology dictates that languages fall under three distinct rhythmic classifications:

1) *syllable-timed languages* in which the duration of every syllable is equal (e.g., Italian).
2) *stressed-timed languages* in which the interval between two stressed syllables is equal (e.g., English).
3) *mora-timed languages* in which the duration between every mora is equal (e.g., Japanese).

However, even during the early stages of the rhythm typology movement, that is throughout the 1960s and 1970s, research was being conducted which offered no support for Abercrombie's claims relating to the presence of isochrony in these rhythmic groupings (e.g., Bolinger, 1965; O'Connor, 1965; Shen & Peterson, 1962). For example, Bolinger's (1965) study which analysed characteristics of both duration and pitch within spontaneous speech, found no evidence of isochrony, with Bolinger suggesting the reason for this being the interrelationship between syllable length and pitch accent, and the fact that the latter is highly variable within spontaneous speech. With a greater number of studies still failing to corroborate the claims for isochrony, the idea that isochrony could be a perceptual phenomenon rather than actually evidenced within the speech signal, that is, subjective rather than objective, became a focal point for investigation for a number of researchers (e.g., Allen, 1972; Donovan & Darwin, 1979; Fowler, 1979; Morton et al., 1976). Although the methodologies employed and the specific objectives of such studies varied, they were all united in the overarching finding that perceived timing was not a manifestation of any actual acoustic regularity within the speech signal.

Overall, during this period of time from the 1940s up to the 1970s, speech rhythm research was primarily concerned with the field of rhythm typology and the concept of isochrony being present within the different rhythmic classifications. However, towards the latter stages of this period, research was already starting to contradict the claims for isochrony, and this is a trend which was set to continue over the coming decades.

2.3.2. Abandoning the isochrony quest and categorical rhythm classes

As speech rhythm research progressed through the 1980s and 1990s, so did the weight of evidence against there being any actual objective isochrony evidenced within speech rhythm patterning. This was evident for all three of the proposed rhythmic classes, with there being studies which dismissed isochrony for syllable-timed languages (e.g., Dauer, 1983, Pointon, 1980; Roach, 1982, Wenk & Wioland, 1982), stress-timed languages (e.g., Dauer, 1983; Faure et al., 1980; Jassem et al., 1984; Roach, 1982) and mora-timed languages (e.g., Hoequist, 1983a, 1983b). Alongside

the demise of any sort of rhythmic isochrony being evidenced, the same fate was set to befall the rhythm typologists' assertions that the world's languages could be neatly classified into one of the three rhythmic classes.

Throughout the 1980s, a body of research emerged which dispelled the idea of neat rhythmic classifications being evident. Where some of these studies showed that both stress- and syllable-timing could exist within a single language (e.g., Miller, 1984), others introduced alternative labels for classifying specific languages such as 'stress language' and 'boundary language' (Vaissière, 1991, p. 118), as well as 'leader-timed' and 'trailer-timed' (Wenk & Wioland, 1982). Abandoning any strict rhythm classification labels altogether, Dauer (1983, 1987) instead suggested that cross-linguistic differences pertaining to differing degrees of durational variability are best depicted as being on a continuum by which languages are able to exhibit both syllable-timed and stress-timed characteristics. Dauer's proposition of a continuum being more appropriate was reinforced by a later study conducted by Auer (1993) whose survey of 34 languages resulted in the proposition that the prosodic rules which languages possess relate to either the syllable or the word, and therefore 'syllable languages' and 'word languages' should be seen as being at each end of a prosodic continuum.

The shift away from syllable-timed and stress-timed classifications meant that some researchers throughout the 1980s and 1990s also looked to consider other spectral properties such as within-syllable energy distributions (e.g., Harsin, 1997, Howell 1984, Pompino-Marschall 1989, Scott 1994) when postulating their speech rhythm theories. Despite some of these perceptual studies indicating that other speech parameters might be equally as important in the conceptualisation of speech rhythm as temporal attributes, duration still remained the most widely studied rhythmic parameter throughout the 1980s and 1990s, with this set to continue into the 21st century in light of the development of a number of rhythm metrics (e.g., Low, 1994, 1998; Ramus et al., 1999).

Overall, throughout the 1980s and 1990s, speech rhythm research dispelled the idea of there being physical, measurable isochrony within the speech signal, with focus being shifted towards determining the influential factors relating to perceived isochrony. Research during this period also proposed a shift away from the strict

dichotomy of languages being either syllable-timed or stress-timed, with studies instead suggesting that the rhythmic properties of the world's languages should be conceptualised as being on a continuum.

2.3.3. Rhythm metrics

As briefly alluded to above, as speech rhythm research progressed into the 21st century, there was a marked upsurge in speech rhythm production experiments which sought to make use of the rhythm metrics developed in the previous decade by Ramus et al. (1999) and Low (1994, 1998). As such, the vast majority of these studies were concerned with the temporal and durational characteristics of speech and how these quantification methods could be utilised for classifying the rhythmic properties of different languages. Whilst some of these rhythm metrics were specifically designed for measuring temporal attributes such as Ramus et al.'s (1999) %V (the percentage over which speech is vocalic), ΔV (the standard deviation of the vocalic segments across an utterance), and ΔC (the standard deviation of the consonantal segments of an utterance), others such as Low's (1998) PVI (Pairwise Variability Indices which measure the difference for a given parameter between immediately consecutive intervals and average these differences over an utterance) could also be applied to other speech parameters such as intensity. Indeed, in her study of prosodic prominence in Singapore English, Low (1998) demonstrated that PVIs could be implemented to quantify speech rhythm in terms of measures of duration and intensity, with both parameters performing equally well.

However, despite the applicability of some metrics which could facilitate the quantification of speech rhythm on multiple levels, much of the experimental research of the early 2000s was still centred on measuring durational characteristics. Some of these studies seemingly sought to reignite the rhythm typology flame as they investigated rhythmic differences across numerous different languages. For example, Grabe and Low (2002) used PVIs to measure consonantal and vocalic properties for 18 languages with their results showing that languages that were previously designated as being either syllable-timed or stress timed were, in general, separated as such, however there were a number of languages which lay between these classifications. This led Grabe and Low to conclude, in agreement with Dauer (1983)

and Auer (1993), that the rhythmic properties of these languages were best conceptualised as falling on a continuum between syllable-timed and stress timed. Ramus (2002), however, claimed that the results obtained from utilising Low's (1994, 1998) metrics (i.e., PVI) and Ramus et al.'s metrics (i.e., %V, ΔV, ΔC) allowed for languages to be more strictly categorised as either syllable-timed or stress-timed.

As this durational focus continued, so did the development of new temporal-based rhythm metrics which sought to explore the rhythmic properties of the world's languages (e.g., Bertinetto & Bertini, 2008; Dellwo, 2006; Deterding, 2001; Duarte et al., 2001; Gibbon & Gut, 2001). The conclusions drawn from studies such as these continued to attempt to categorise the languages studied into different rhythm-based groups, whether this be under the conceptualisation of a rhythm continuum or a more dichotomous approach.

Given that studies such as these often produced results which were not explicitly compatible with one another, it was not long before research started to emerge which cast aspersions on the reliability of rhythm metrics and their methodological groundings. One such study, conducted by Arvaniti (2009), critiques the use of these rhythm metrics, for the classification of languages on a number of levels. In her investigation of six languages, Arvaniti found that the metrics utilised (ΔC, %V, PVIs and Varco) were highly sensitive to both elicitation method as well as syllable complexity resulting in inconsistent rhythm classifications and cross-linguistic differences (based on scores from the metrics) which were statistically non-significant. In light of her results, Arvaniti suggests that cross-linguistic differences captured by these metrics are not robust and warns that making cross-linguistic comparisons and rhythmic classifications based on these metrics is not reliable. (However, cf., Prieto et al. (2012) who, using a similar methodology to Arvaniti, found that some of these metrics (nPVI-V, ΔV, and VarcoV) are good for discriminating between the languages which they studied (English, Spanish and Catalan.))

Joining the critique of these rhythm metrics, Barry et al. (2009) showed that using PVIs in the same way as Grabe and Low (2002) produced different results depending on speaking style, again highlighting the sensitivity of these metrics to different

methodological procedures and their apparent unstableness for making cross-linguistic comparisons.

Alongside the issues surrounding the methodological implementation and the reliability of these metrics, researchers also drew attention to the fact that the utilisation of these metrics was only focussing on one parameter: duration. This sole durational focus was subsequently deemed problematic by a number of studies.

For example, in their speech rhythm perception study, Barry et al. (2009), demonstrated that parameters other than duration can be equally as important in influencing listeners' perception of rhythmicity being evidenced. They tested the importance of duration, f_0, intensity and vowel quality by manipulating these features to ascertain the relative contribution of the four parameters to the impression of rhythmicity amongst their subjects. Results showed that, although duration did rank in top spot amongst the four parameters (albeit to different extents across the different listeners), f_0 was also a highly relevant factor in the perception of rhythmicity (followed by intensity and then vowel quality).

Arvaniti (2009) agrees that this durational focus is misplaced and points towards 'the host of other factors' which should be taken into account when conceptualising speech rhythm. She utilises prior research within the psychological domain (e.g., Woodrow, 1951; Fraisse, 1963, 1982) to illustrate the problematic nature of the concept of syllable-timing and a sole durational focus. Making close reference to Dauer (1983), Arvaniti highlights the importance of *stress* as being the crux from which different languages could potentially be defined and classified, and that this *stress* (or *prominence*) is derived from more than just temporal characteristics.

Further studies (e.g., Shattuck-Hufnagel & Turk, 2013; Kohler, 2009) also emerged which supported the notion that these rhythm metrics were not wholly reliable when attempting to categorise a given language and again suggested that a sole durational focus is not sufficient and that other speech parameters such as f_0 and intensity need to be factored into speech rhythm research.

Despite these criticisms of rhythm metrics emerging within the linguistic literature, this did not dissuade all metric-based rhythm research, with new metrics being

introduced and new research which sought to quantify rhythmic variability at various different levels such as in relation to the durational differences of vocalic and consonantal intervals (e.g., Bradshaw, Hughes, & Chodroff, 2020; Dellwo, 2006, 2008; Grabe & Low, 2002), the duration of voiced and unvoiced intervals (e.g., Dellwo et al., 2007), and the durations between peaks in syllabic intensity (e.g., Leemann et al., 2014; He & Dellwo 2016). In light of the continuing development of new rhythm metrics, it is important to have an understanding as to the limitations of using such metrics. In the first instance it is necessary to ascertain that a given metric is indeed capturing rhythmic characteristics. That is, ensuring that the metric is robust in its design and will not be susceptible to capturing any unwanted characteristics which are not attributed to speech rhythm (e.g., vowel voicing vs. vowel devoicing, longer vowels vs. shorter vowels). Additionally, from a forensic standpoint, it is also necessary to be mindful if using such rhythm metrics within speaker discrimination tasks – that is, being mindful that although a given metric may present as being a good speaker discriminant, this does not necessarily mean that it is a good measure of speech rhythm. Once again, this necessitates ensuring that the metric is robust in its design and is capturing the proposed rhythmic characteristics whilst omitting any unwanted non-rhythmic differences between speakers.

Given that the present thesis looks to examine speech rhythm in relation to spontaneous speech data, where comparisons between speakers' rhythm patterns will be based on content-mismatched utterances, it is necessary to provide discussion pertaining to previous research that has investigated the effects of utterance content on rhythm measures. The three studies discussed below all make use of rhythm metrics and are focussed solely on durational characteristics, hence their inclusion at this juncture, before attention is shifted to other rhythm parameters.

Wiget et al. (2010) investigated how robust a number of durational rhythm metrics were to variation between speakers, sentence materials, and measurers. The study, which was comprised of six speakers of Standard Southern British English, each reading five sentences, sought to asses the impact of these sources of variation on various metrics. The results showed that, of the three factors assessed, it was the differences between sentences which resulted in the most rhythmic variability.

Furthermore, the results highlighted that the variability demonstrated as a result of sentence are greater than the between language variability. This finding gives support to the supposition that individuals may generate a distinctive rhythm through the deliberate choice of lexical elements and/or morphosyntactic configurations, which can produce specific rhythmic characteristics in their spontaneous speech. This supposition is given further merit when taking into consideration the results obtained by Prieto et al. (2012) discussed below.

Prieto et al. (2012) investigated how variations in syllable structure influence speech rhythm metrics across three languages that are recognized as belonging to various rhythmic classifications (English, Spanish and Catalan). The results from the experiments showed that rhythm metrics reveal differences that remain apparent even when syllable structure is controlled within the experimental materials. This is especially true in the contrast between English and Spanish/Catalan, suggesting that there are essential differences in durational patterns that cannot be exclusively linked to phonotactic factors. The experimental setup of this study also highlighted the role of syllable structure within languages. Specifically, sentences composed mainly of phonotactically simple syllables exhibit distinct rhythmic variations when compared to those with more complex syllable structures. It is therefore plausible that a particular choice of vocabulary or morphosyntactic forms, which predominantly includes either simple or complex phonotactic features, could consequently shape the measurable rhythmic features of speech. A further observation from the study was that the stressing of prosodic heads or pre-final syllables produces systematic differences in the measurements of speech rhythm. Given that speech rhythm, along with intonation and stress, is commonly categorised under the term prosody, this finding lends support to the suggestion that other prosodic features, including intonation and stress, could impact the duration-related aspects of speech rhythm.

Another study which sought to examine the impact of utterance content on rhythmic variability is that of Dellwo et al. (2015) who investigated the effects of within-speaker variability of linguistic structural characteristics on a range of durational rhythm metrics. The study consisted of 16 speakers who were recorded whilst engaged in spontaneous speech during interviews. Transcripts of 16 selected sentences from these

interviews were then created, and the speakers were instructed to read them. Each speaker read their own previously produced spontaneous sentences as well as the transcripts from the other speakers, resulting in a total of 256 sentences across the 16 individuals. The results regarding the influence of sentence structural features on rhythm scores suggest that sentences differ in the complexity of their consonantal and vocalic intervals, and this variability affects rhythmic measurements to a certain extent. Dellwo and colleagues also report that when they examined the rhythm scores of sentences produced by the speakers in comparison to those formulated by different speakers, they found no indication that the variations in phonotactic complexity could account for the differences in variability amongst speakers. Dellwo et al. use these findings to suggest that differences in speech rhythm between speakers (within their experiment) cannot be linked to speakers' individual preferences for lexical and morphosyntactic choices. In addition, they discount the notion that speaker-specific speech rhythm is contingent upon distinctive prosodic elements, given their results which show that the prosodic variability introduced by varying speaking styles did not influence the differences between speakers. Overall, they propose that their results support the hypothesis that individual differences in articulatory movements are the primary contributors to rhythmic variability observed between speakers.

Although the three studies reviewed above are focussed solely on durational properties of speech, and more specifically have a focus on the variability of consonant and vocalic intervals, their inclusion here has demonstrated the effects which utterance content can have on rhythmic variability. In relation to the present thesis, in particular Chapter 2 in which spontaneous speech utterances are assessed to determine if their rhythmic characteristics exhibit any speaker-specificity, the results of these studies may present as concerning. However, as suggested by Dellwo et al. (2015), the differences observed between speakers are more plausibly attributed to the individual control mechanisms governing their articulators. Section 2.5 of the present chapter provides further discussion relating to forensically-motivated speech rhythm research for which this articulatory rationale for between-speaker rhythmic differences is also claimed.

Although such experimental research has undoubtedly furthered our understanding of speech rhythm in relation to differences in timing, what about the other contributing factors such as pitch and intensity? As already alluded to above, previous studies which have looked to incorporate these factors have been for the most part perceptual, with acoustic features (e.g., duration, pitch and intensity) being manipulated and listeners' judgements being indicative as to which feature was most important in determining prominence (e.g., Bertinetto, 1980; Kohler, 2009; Llisterri et al., 2003; Sautermeister & Eklund, 1997). Results from studies such as these revealed that all of these acoustic parameters have the potential to carry the most weight in determining prominence, and therefore our perception of differences in speech rhythm may be dependent upon more than temporal information alone.

Experimental research which sought to investigate these acoustic measures alongside one another, however, was far less abundant, and the prevalence of duration-based studies seemingly suggested that researchers had all but forgotten about the findings of these perceptual experiments – that is until relatively recently. These more recent studies which have looked at unifying these speech parameters in terms of speech rhythm are discussed in Section 2.4.

In summary of the speech rhythm research discussed above which has focussed on rhythm metrics, none of the metrics developed or utilised were able to provide evidence for isochrony being evidenced in the speech signal, nor were they able to successfully classify languages into neat rhythm categories based on measures of timing alone. There have, however, been some studies (e.g., Tilsen & Arvaniti, 2013) which have found some (limited) success in determining cross-linguistic differences using metrics when other acoustic parameters have been the focus.

As shall be seen, a small number of these rhythm metrics are utilised in Chapter 3 of this thesis in order to determine their capacity for discriminating between individual speakers (as opposed to between groups of speakers for cross-linguistic purposes). However, given that prior cross-linguistic research has shown that these metrics in general harness little promise with regards to reliably categorising specific languages, these metrics are not utilised extensively, with only a few being selected for analytical purposes. Nevertheless, the present section has been important in demonstrating how

speech rhythm research was advanced and broadened by the development and testing of these metrics and in highlighting the complexity of capturing, measuring and describing speech rhythm.

Finally, despite the overall finding that the application of these metrics still results in there being no conclusive evidence for physical isochrony and languages belonging to different rhythm classes, it is important to note that within the literature such rhythm classes will still be made reference to, as will those prototypical languages which are most commonly associated with belonging to a specific rhythmic class.

2.3.4. Is there rhythm in speech?

Given all that has been discussed above, it seems appropriate to raise the question as to whether speech is actually rhythmic. One important consideration when formulating a response to this question is the idea of *regularity* being inherent when conceptualising speech rhythm. As has been shown from the numerous studies referenced above, *regularity*, whether this be in relation to temporal attributes or other acoustic features, is something which has never been evidenced in experimental speech rhythm research. The lack of evidence for periodicity within the speech signal has led a number of researchers to make the claim that rhythm within speech is therefore best conceptualised as being a perceptual phenomenon rather than something that can be evidenced physically (e.g., Arvaniti, 2009; Dauer, 1983; Roach 1982; Shattuck-Hufnagel & Turk, 2013; White & Mattys, 2007). That is, listeners may make claim to being able to perceive rhythmic variability and/or different rhythmic categories, however this does not mean that there will necessarily be any evidential rhythmic patterning (e.g., isochronous phenomena) when the acoustic signal is analysed.

Before describing the nature of the research which has suggested that speech rhythm is best thought of as a perceptual construct, it is necessary to consider the impact of such claims in relation to the tenability of analysing speech rhythm from a forensic point of view. If it transpires that speech rhythm can be accounted for in a more robust way by perceptual means as opposed to measurable characteristics within the speech

signal, it could be supposed that speech rhythm as a feature would not be of use for forensic voice comparison purposes. That is, having a feature which is analysed solely at a perceptual level may be discounted as carrying little weight in comparison to other 'more measurable' features such as voice pitch and vowel formant frequencies. However, when taking into account the degraded nature of the speech data which forensic practitioners are often faced with, it is often the case that an expert's perceptual judgements will be more robust than obtaining vowel formant measurements from a degraded speech signal, for example. Furthermore, there are some features which are routinely analysed within FVC casework at solely a perceptual level, with voice quality being the standout example here. Within FVC casework, voice quality is analysed through auditory means as opposed to taking acoustic measurements, with this auditory analysis being aided by a perceptual assessment framework – the Vocal Profile Analysis (VPA) framework (Laver, 1980). It stands then that discounting the analysis of speech rhythm within FVC casework on the basis of it being potentially a feature which is best assessed through perceptual means would be unfounded (see Chapter 5 and Chapter 6 which focus on the perceptual assessment of speech rhythm from a forensic perspective).

Returning now to the empirical work which has suggested that speech rhythm is best thought of as being a perceptual phenomenon, Lehiste's (1970) study demonstrated that listeners are able to perceive durational differences (and thus perceived regularity) which are above a specific threshold. Additionally, in a later study, Lehiste (1977) showed that listeners are more likely to perceive isochronous events when presented with non-speech stimuli than they are when presented with authentic speech. Lehiste uses these findings to suggest that listeners are therefore not as sensitive to differences in duration when speech is concerned and thus could be more likely to make claims of isochrony being present when, in actuality, it is not.

Another study which bares relevance here is that of Arvaniti and Ross (2010) who investigated whether listeners could classify low-pass filtered utterances of six different languages into rhythm classes. In their experimental design, they examined how listeners (who were speakers of different languages) rated each utterance's rhythm in comparison to a series of non-speech trochees (a metrical foot containing

one stressed syllable followed by one unstressed syllable). Their results showed that none of the languages under investigation were deemed to be similar to the non-speech trochees by any group of listeners. Although one explanation for this result could be due to the relative simplicity of the trochee pattern in comparison to the rhythmic patterns of the languages, the finding that English was judged as being the least 'trochee-like' rhythm is somewhat surprising given that English is predominantly acclaimed as being the quintessential stress-timed language (and therefore should have in fact been rated as the most 'trochee-like'). An additional finding which serves to support the notion that speech rhythm and its supposed regularity is likely to be a perceptual phenomenon is that of the three types of stimuli presented to the listeners, it was the "uncontrolled" stimuli that were rated as more 'trochee-like' than the rest (i.e., the stress-timed stimuli and the syllable-timed stimuli). The suggested explanation put forward by the authors here is that the uncontrolled stimuli were likely more natural and that they could have therefore been read more fluently – thus resembling what one would perceive of a more natural rhythmic pattern despite the stimuli in fact having the least regulated (and more variable and complex) syllabic makeup.

Overall, studies such as the ones mentioned above serve to promote the idea that speech being rhythmic is something which may in some circumstances be perceived despite there being no physical evidence of any form of periodicity within the acoustic signal. It would seem, therefore, that the answer to the question '*is there rhythm in speech?'* should be a quite simple: '*no*'. However, before landing on such a definitive conclusion, it is worth briefly taking a step back and considering *rhythm* from a more general standpoint.

For example, when we consider rhythm from a musical perspective, *regularity* is something which is easily perceived. If we take a drum beat for example, the beat will be confined to a given time signature in which certain hits (e.g., of the snare drum) will occur on certain beats (e.g., the second and fourth beats of a given measure) resulting in a regular rhythm being perceived by the listener. Although this regularity is something which is often thought of as being a fundamental and integral aspect of musical rhythm, in actuality, strict periodicity (i.e., regularity/isochrony) is something

which is rarely present in the physical output of the musical signal. Similar to what has been shown with regards to there being no evidence for periodicity within speech rhythm, Large and Jones (1999) illustrate the lack of periodicity within musical rhythm by showing how the time intervals between the onsets of key presses for a given section of musical notation lack any kind of regularity or periodicity. Nevertheless, a listener will still perceive the presence of regularity within a given piece of music even if there is no physical evidence of isochrony in the external rhythm (i.e., if there are differences in timing between the onsets of key presses/drum hits/string plucks, etc.). The lack of strict isochrony within musical rhythm further highlights the overdependence on timing/duration when rhythm, even in its more general sense, is conceptualised.

As such, this once again raises the question as to the role which all of the other features generally associated with rhythm have to play. If one was asked as to what these other features might include with regards to the concept of musical rhythm, one might expect to receive answers such as pitch, loudness (or intensity) and stress (or prominence). As has been shown in the previous subsections of this chapter, these are all features which play a part in our conceptualisation of speech rhythm. Therefore, in consideration of this, making the claim that speech is rhythmic may not seem all too outlandish. That is, all of the components which play a part in our perception of musical rhythm (duration/timing, pitch, loudness, prominence, etc.) are also the same components which feed into our conceptualisation of speech rhythm.

However, if we return to the drum analogy introduced above, if we were to play this drumbeat with a faster tempo or with greater force (e.g., increasing the loudness of the snare drum hits) this would have little effect on the pitch produced (that is providing that these drum hits did not exceed 50 beats per second). Such duration and pitch regularities do not transfer over to speech. Given that rhythm is defined by the presence of regularity, it raises the question as to how the term ever actually came to being applied to speech to begin with. Indeed, arguments have been put forward that speech may be in its inherent makeup actually *antirhythmic* - a term suggested by Nolan and Jeon (2014) to describe, in the case of a language such as English, ‘the blatant disregard for proper sequential alternation in favour of syntagmatic

irregularity' (Nolan & Jeon, 2014, p.7). Despite entertaining the idea that speech might be more antirhythmic than rhythmic, Nolan and Jeon do not advocate abandoning the pursuit to understand the relationship between speech and rhythm, but rather suggest that investigating this relationship in less arbitrary ways could be more useful.

2.3.5. Summary

In summary, rhythm is not self-evident when observing the complex acoustic speech signal and various interrelated speech parameters seemingly combine to contribute towards what we may perceive as rhythm in speech. If taking the stance that speech is at least in some way rhythmic, then what exactly is the patterning that we perceive and, more importantly for the purposes of this thesis, how can we capture, measure and describe these patterns? Could it be that there are certain parameters and certain fragments within a given utterance that evidence rhythmic patterning to a greater extent than others? Where the experiments which form the foundation of this thesis look to provide answers to these questions over the subsequent chapters, the remainder of the present chapter considers prior speech rhythm research which has accounted for acoustic parameters other than duration (namely, pitch and intensity) as well as the small body of research which has been motivated by its potential forensic implications and applications.

2.4. Beyond duration: speech rhythm research with other acoustic parameters

As the previous sections of this chapter have shown thus far, the vast majority of speech rhythm research has been exclusively focussed on measuring and describing durational differences and variation between different languages. This sole durational focus has subsequently been shown to be problematic given that it is known that our perception of speech rhythm takes into account various other acoustic parameters also.

In accordance with this, there have been a number of studies which have looked to account for and assess these features with regards to their role in the conceptualisation of speech rhythm. Although some of studies have been briefly mentioned above (e.g., Berry et al., 2009; Kohler, 2009; Shattuck-Hufnagel & Turk 2013), the following subsections address this body of research in greater depth given the multidimensional approach to speech rhythm which the current thesis embodies.

2.4.1. Intensity-focussed speech rhythm research

Studies which have focussed on the variability of intensity within the speech signal have sought to determine the extent to which intensity characteristics can be used to capture speech rhythm patterns. The rationale for assessing speech rhythm through intensity variability stems from the observed correlation between mouth aperture size and signal intensity. Specifically, an increase in the area of the mouth opening is associated with elevated intensity levels, and conversely, a smaller mouth opening correlates with reduced intensity. The dynamic opening and closing gestures, which constitute the articulatory basis for speech rhythm, perpetually modify the shape of the vocal tract. This modification influences the filter characteristics that act upon the source signal, thereby altering its spectral properties and intensity levels. As a result, the cycles of opening and closing can be approximated by variations in signal intensity.

Research which has focussed on intensity in relation to its role in capturing speech rhythm patterns has served different purposes such as:

- investigating rhythmic differences within a single language (e.g., Low, 1998)
- investigating differences between different varieties of the same language (e.g., Fuchs, 2014)
- investigating rhythmic differences between different dialects (e.g., Ferragne, 2008; Cichocki et al., 2014)
- investigating differences between children and adults (He, 2018)
- investigating differences between first and second language English (He, 2012).

Where some of these studies have focussed solely on intensity and its role in defining speech rhythm, others have also considered intensity alongside other rhythmic parameters such as duration. In order to provide an insight into some of the findings from this intensity-focussed rhythm research, as well as the methodologies employed, three of these studies are described in more detail here.

Fuchs (2015) sought to investigate speech rhythm differences between two varieties of English – British English and Indian English – by accounting for intensity and duration variability. He developed a novel metric which combined two existing metrics which served to simultaneously account for intensity and duration variability amongst the speakers. Using the newly designed metric alongside the two existing metrics, Fuchs demonstrated that Indian English is less variable in terms of intensity and duration both as separate entities and also as a simultaneous commodity. These results therefore contribute to the understanding that speech rhythm is realised in different dimensions, through different acoustic and perceptual correlates, and that a multidimensional model of speech rhythm that accounts for more than just duration is recommended.

Chichoki et al. (2014) looked to assess cross-dialectal differences in speech rhythm from the perspective of intensity and duration using a number of rhythm metrics. They analysed utterances of read speech from 140 speakers from three different dialects of French spoken in New Brunswick, Canada. Using discriminant analyses, their results showed that both intensity- and duration-based rhythm metrics played a part in distinguishing between the three dialects. They conducted three classification experiments which accounted for (1) duration-based metrics, (2) intensity-based metrics, and (3) the combination of both intensity-based and duration-based metrics. Classification results for all three experiments were above chance level (33.3%). Overall, intensity-based metrics performed slightly better than duration-based metrics (45.7% vs. 41.4%), with the combination of both types of metrics yielding the strongest result (47.1%). In interpreting their findings, they suggest that intensity, given its better performance than duration, is an acoustic indicator of prominence for the three dialects under study. As such, they go on to advocate for a speech rhythm

model which is multidimensional in its makeup, accommodating different prosodic features.

He (2012) carried out an intensity-focussed study which examined whether three specially developed metrics would be able to differentiate between L1 speakers and L2 speakers. Using the metrics developed, He analysed the variability in intensity patterns within elicited (read) speech across three speaker groups: L1 English, L1 Mandarin and L2 English (Mandarin speakers). Results showed that all three of the intensity metrics had reasonable success at distinguishing between L1 English and L1 Mandarin, with L1 English exhibiting significantly higher degrees of intensity variability than L1 Mandarin. This result supported He's hypothesis which drew upon the notion that English, being a "stress-timed" language, may exhibit greater intensity variability across the course of an utterance owing to stressed syllables having higher amplitude levels in comparison to unstressed syllables. Mandarin, on the other hand, a language often classified as being "syllable-timed", may exhibit comparatively more levelled intensities across the course of a given utterance. The most important finding from He's study, however, was attributed to the results obtained for L2 English. There was no significant difference between L2 English and L1 Mandarin, indicating that, when speaking L2 English, these Mandarin speakers exhibited more (native) Mandarin-like intensity patterns (i.e., less intensity variability). In highlighting the importance of this finding, He compares these results to previous studies (He, 2010, 2011 (same dataset, speakers, etc.)) which found that, for durational measures, L2 English had similar metrics scores to native (L1 English) scores. He points out the disparity between the results and suggests that for L2 English learning amongst Mandarin speakers, durational characteristics such as vowel reductions and syllable structures may be more easily learnt than characteristics pertaining to intensity. Therefore, although L2 English is similar to L1 English in terms of duration metrics scores, this is not sufficient for supporting L2 speakers acquiring native-like speech rhythm patterns. He concludes by proposing that future speech rhythm research should take measures of intensity into consideration as accounting for duration alone is not sufficient.

2.4.2. f_0-focussed speech rhythm research

Studies which have examined f_0 in terms of speech rhythm and its role alongside other prosodic parameters such as duration and intensity have been, for the most part, geared towards examining cross-linguistic rhythmic differences. Where some of this research has looked into f_0 variability alongside durational variability (Cumming, 2011; Niebuhr, 2009; Niebuhr and Winkler 2017; Polyanskaya et al., 2020), other studies have investigated f_0 variability alongside intensity variability (e.g., Alku et al., 2002; Jessen, 2005; Köster, 2002; Lee & Todd, 2004; Plant & Younger, 2000; Traunmüller & Eriksson, 2000). The following three studies are summarised to provide an example as to what some of the methodological approaches and results look like for this f_0-focussed research. These specific studies were selected as they cover both production and perceptual findings as well as assessing f_0 through various measurements and alongside other parameters.

Polyanskaya et al. (2020) investigated cross-linguistic rhythmic differences between Italian and English focussing on measurements of f_0 as well as durational measurements. They accounted for this cross-linguistic variation by quantifying a number of different parameters, namely, the regularity of tonal alternations in time; the magnitude of f_0 excursions; the number of tonal target points per intonational unit; and the similarity of f_0 rising and falling contours within intonational units. They analysed semi-spontaneous speech from 20 female speakers (10 from each language) and found that Italian possessed a stronger tonal rhythm than English as they had hypothesised. Italian demonstrated a higher regularity in the distribution of f_0 minima turning points, larger f_0 excursions, and more frequent tonal targets. In explaining their results, Polyanskaya et al. point out that a listener's native language determines the significance of f_0 and durational ratios in the perception of speech rhythm and where some languages pay much more attention to durational ratios than to tonal cues, the contrary will be true for other languages. They draw upon the findings of Cumming's (2011) study in which it was shown that (Swiss) German listeners pay much more attention to durational ratios than to tonal cues, whilst French listeners pay equal attention to durational and tonal cues. Polyanskaya and colleagues relate this to their study on the concept of Italian and French being rhythmically similar and English

and German being rhythmically similar, and, as such, a greater degree of pitch variation would be expected in Italian rather than English. They conclude by highlighting that speakers of different languages use different phonetic means and strategies to construct speech rhythm patterns, with these varying in terms of the weight placed on specific acoustic cues (e.g., f_0, duration, intensity, etc.). As shown from their results, f_0 is one parameter which should be considered in speech rhythm research.

Niebuhr (2009) sought to determine the role of f_0 within speech rhythm from the perspective of its role in signalling prominence. He investigated the performance of 32 German native speakers with regards to their ability to perceive and subsequently reproduce a number of speech stimuli which contained target sections in which the f_0 of certain syllables had been manipulated. He found that that the perceived position of the prominent syllable in the target section was affected by the prominence pattern and the resulting rhythm of the context section of the stimuli. Consequently, this determined that the perceived position of the prominent syllable in the target section was shifted so that the local prominence pattern matched the context pattern, creating an overall consistent speech rhythm. Niebuhr uses his findings to highlight the notion that speech rhythm is a perceptual phenomenon which is brought about through changes in acoustic parameters such as f_0, duration, intensity, and sound quality. He goes on to stress the multidimensional nature of speech rhythm and how attempting to account for speech rhythm through measurements obtained from the acoustic signal is a somewhat futile exercise. Instead, Niebuhr concludes by advocating that future speech rhythm research should be more focussed on understanding what speech rhythm actually is, and how it is constructed from a perceptual perspective.

Cumming (2011) conducted a speech rhythm study in which she examined whether f_0 and duration are interdependent cues for the perceived rhythmicality of sentences, and whether or not this depends on the native language of listeners. The experimental design assessed the judgements of two groups of listeners, one being native speakers of Swiss German and one native speakers of Swiss French, in relation to which stimulus sentences had the most natural sounding rhythm. The two language varieties were selected on the basis that they sound rhythmically different from one another

owing to them differing in terms of prosodic properties involving f_0 and duration. The stimuli within the experiment were manipulated in terms of their f_0 and duration, with this being implemented on a given specific syllable in order to determine whether a deviant duration results in a less natural sounding rhythm than a deviant f_0 movement, or vice versa. Cumming found that duration and f_0 are interdependent cues for perceived rhythmicality, and that the relative significance of a non-deviant duration and a non-deviant f_0 excursion in the rhythmicality judgements of listeners depends on their native language. For Swiss German, duration contributed more than f_0 with regards to signalling rhythmicality, whereas for Swiss French f_0 and duration were weighted more evenly with the different durational properties of the two languages being proposed as the reasoning being this finding. In summarising her findings, Cumming encourages future research to investigate speech rhythm as a perceptual phenomenon rather than trying to measure speech rhythm through production tasks. She proposes that future experiments should not have a sole durational focus and should incorporate the analysis of f_0 given the apparent interdependence of these two features, and, further still, should include other parameters such as intensity and vowel quality which likely also are cues to rhythmic prominence.

2.4.3. Summary

To sum up, the previous two subsections served to illustrate the development of speech rhythm research as it progressed beyond the sole durational focus which had previously dominated. The inclusion of additional parameters within speech rhythm research, whether that be in relation to f_0 or intensity, have bolstered the claim that a multidimensional approach is needed when conceptualising speech rhythm or carrying out rhythm-related research. Indeed, as the findings from some of the above studies indicate, it may be that some parameters bear more importance rhythmically than others, and that this may be dependent on the language being studied. Moreover, the different acoustic parameters involved in our perception of speech rhythm are likely to have a level of interdependence upon one another, that is, duration, f_0 and intensity will be to a greater or lesser extent interrelated, with these complex interrelations manifesting as perceivable attributes of rhythmic patterning (e.g., as

prominence). Although the body of research discussed above has almost exclusively made use of laboratory data, the findings provide reasonable cause for the present thesis to examine how well a multidimensional approach to capturing speech rhythm transfers over to spontaneous, content-mismatched speech which is predominantly the type of material encountered in forensic casework.

2.5. Previous forensically-motivated research on speech rhythm

Given the forensic focus of this thesis, the following subsections provide a summary of the small body of speech rhythm research which has had a forensic focus. Similar to the generalised speech rhythm research which has been discussed up until this point, forensically-motivated speech rhythm studies have often been focussed on an individual rhythmic parameter (as opposed to taking a multidimensional approach). As such, the research summary provided below is presented in subsections which deal with duration-focussed, intensity-focussed and f_0-focussed studies respectively. Following on from this summary of production-based research, section 2.5.4 introduces the even smaller body of research focussed on speech rhythm perception and its forensic implications.

Before the aforementioned research summary is discussed, it is first important to outline the reasons as to why speech rhythm could be useful as a feature for discriminating between speakers. The rationale for this supposition is, on the one hand, owing to the unique anatomical characteristics associated with a speaker's vocal apparatus and speech organs, and, on the other hand, accounting for the individual ways in which speakers operate their articulatory mechanisms. It is the interplay of these two factors which ultimately results in the emergence of speaker idiosyncrasies within the speech signal.

For example, in relation to intensity, one of the three main parameters most commonly associated with contributing toward speech rhythm (alongside pitch and duration), earlier research has established a direct relationship between the size of the mouth aperture and the intensity of the speech signal. Specifically, a larger mouth opening is

associated with greater intensity, while a smaller opening corresponds to lesser intensity. In addition, studies focusing on subglottic and pulmonic air pressure - both of which are intrinsically linked to speech intensity - have uncovered significant inter-speaker variability, even within strictly controlled syllable contexts. These individual characteristics result in speaker-specific variations in pulmonic and subglottic pressure, which are reflected in the speech signal. Similarly, in relation to the parameters of pitch and duration, both can also be expected to exhibit speaker discriminatory potential as a result of idiosyncratic anatomical and articulatory factors as well as individual conversational behaviours. For example, where some speakers might mark syllabic prominence through variations in pitch (e.g., an increase in pitch), others may do so through durational means (e.g., prolongations of syllables). Aside from syllabic prominence, it might be that some speakers' speech is characterised by specific speech units such as filled pauses or verbal fillers (e.g., *yeah*, *well,* etc.), with such units being marked by specific prosodic behaviours which could result in distinctive speech rhythm patterns emerging.

Investigations into the speaker-specificity of speech rhythm patterns through the analysis of intensity, duration and pitch have primarily been conducted in controlled speech environments, such as through read speech. It is reasonable to assume that individual differences may be more evident in spontaneous speech, influenced by distinctive connected speech processes. By analysing spontaneous speech data, the present thesis aims to provide an initial insight into the efficacy of analysing speech rhythm for forensic purposes through these measures, whilst also testing the effectiveness of rhythm metrics to speech data applicable to forensic contexts. The production experiments carried out in the present work (see Chapter 3 and Chapter 4) will therefore also provide an initial insight as to how the results from acoustic measurements can be generalised for application within FVC casework when limiting factors affecting the speech material (e.g., poor audio quality, limited duration of speech material, etc.) are present. Such limiting factors are, unfortunately, commonplace within FVC casework, and it may very well transpire that endeavouring to capture and compare speech rhythm patterns through acoustic means would rarely be afforded (see discussion below for how analysing speech rhythm through perceptual means could alleviate this problem). Nevertheless, testing how acoustic

measures of speech rhythm perform on forensically-relevant, content-mismatched speech data is the logical next step from previous work which has focussed on controlled laboratory data and will provide a basis for the tenability of using such measures within FVC casework.

The present work also explores the forensic potential of analysing speech rhythm through perceptual means. Given that the speech data which is found in FVC casework is often of suboptimal quality, measuring speech rhythm through acoustic means may not be possible. This is particularly true with regards to obtaining reliable measurements in relation to intensity and pitch characteristics as these are likely to be compromised if the speech data is severely degraded. Although accounting for speech rhythm patterns through durational characteristics (realised in the temporal domain as opposed to the spectral domain like intensity and pitch) may still be possible under some signal degradation conditions, it is more likely to be the case that assessing speech rhythm from a perceptual perspective is the more feasible approach. For example, within a FVC case, the questioned speech material and/or the known speech sample might be of suboptimal quality meaning that obtaining acoustic measurements relating to speech rhythm characteristics is not possible. Depending on the severity of the degradation, it might also be the case that obtaining other measurements for other speech features such as vowel formant frequencies and pitch measurements are also compromised. However, it is possible that a speaker's speech rhythm patterns may still be accessible on a perceptual level, even if the lexical content of the speech is distorted and perhaps unintelligible. If a forensic practitioner was faced with transcribing the speech in a case in which the audio was of degraded quality and in which there were multiple speakers, it could be that the perceptual assessment of speech rhythm could assist with ascribing a given utterance to a given speaker. That is, even if the lexical content of what is being said is unclear, and other speech features (e.g., voice quality) are affected by the nature of the degradation, speech rhythm properties may still be discernible and be of use to the forensic analyst. To give a further example, a questioned sample might have been recorded at a problematic distance, or perhaps the recording device was situated in a different room to the questioned speaker(s). In such circumstances, it is also likely that some speech features will not be accessible to the forensic analyst. It is plausible, however, that

speech attributes pertaining to the speakers' rhythmic patterns are still perceivable auditorily, such as fluctuations in loudness and pitch. Durational information (e.g., the rate of speech) as well as features such as pausing behaviour could also still be discernible. These perceptual characteristics all contribute to an individual's perceived speech rhythm behaviour, therefore, if still accessible, would permit a practitioner to use speech rhythm to assist with speaker discrimination. Taking this into consideration, conducting research into the perceptual assessment of speech rhythm from a forensic perspective is something which stands to be of benefit to FVC casework and is an area in which the present thesis looks to bolster through perception experiments (Chapter 5) and the subsequent proposal of an auditory framework for the assessment of speech rhythm within the forensic context (Chapter 6).

2.5.1. Duration

Perhaps unsurprisingly, it is studies pertaining to durational measures which are the most readily available, with such studies for the most part focussing on between-speaker variability in durational information (e.g., Dellwo & Koreman, 2008; Dellwo et al., 2012; Leemann et al., 2014; Zhang, et al., 2019). Taking Leemann et al.'s (2014) study as an example, they examined the speaker individuality of temporal features and used a wide range of rhythm metrics to investigate how robust they were to channel variability (high-quality vs. mobile-transmitted speech) and speaking style variability (read speech vs. spontaneous speech). For all metrics, they found high levels of between-speaker variability and low levels of within-speaker variability across both speaking styles. Of the ten metrics included in the study, two, namely the percentage over which speech was vocalic (%V) and the percentage over which speech was voiced (%VO), significantly outperformed the others in explaining between-speaker variability as well as proving to be robust across speaking styles, with just the one (%VO) being robust to channel variability. Such results seemingly suggest that these rhythm metrics could have potential within the forensic domain, and at the very least can aid in explaining some perceptually salient rhythmic differences across languages (e.g., Ramus et al., 1999). (However, as previously discussed, cf. Arvaniti (2012) who finds these metrics to be wholly unreliable when making rhythmic classifications and

linguistic comparisons.) In looking to provide explanations for their findings, Leemann and colleagues suggest that, on the one hand, speaker-specific rhythmic patterns might originate from the anatomical traits of the speaker, which are influenced by the neurological motor patterns that function within the speaker's brain. They also offer the explanation that speaker-specificity in speech rhythm features could stem from their unique idiolectal patterns of articulation, advocating that further research is needed to establish a more rigid basis for speculation. As mentioned previously, these studies have neglected the fact that speech rhythm is perceived and realised not only in terms of its temporal characteristics but also its loudness and pitch differences. It therefore stands that a thorough examination of speech rhythm should take into account all of these prosodic elements. Studies which have looked to unify these three parameters into a multidimensional model of speech rhythm have not been geared towards forensic voice comparison but rather automatic speaker recognition (e.g., Adami et al., 2003; Bartkova et al., 2002). Nevertheless, there are a handful of studies which have investigated both f_0 and intensity in relation to their roles within speech rhythm from a forensic perspective, whether this be as separate entities, or whilst also considering their interactions with duration.

2.5.2. Intensity

Studies which have focussed on the variability of intensity within the speech signal have served different purposes such as investigating rhythmic differences within a single language (Low, 1998), investigating rhythmic differences between different dialects (Ferragne & Pellegrino 2008; Cichocki, et al., 2014), investigating differences between children and adults (He, 2018) and differences between first and second language English (He, 2012).

Only relatively recently have studies with forensic motivations focussed on how intensity varies across individual speakers. Before elaborating on the nature of such research, it is worth addressing in the first instance whether using intensity-based methods could be suitable for forensic purposes given that the data found in forensic casework is likely to be spontaneous speech which is often suboptimal in quality. It is widely recognised that the analysis of intensity in spontaneous speech presents

significant challenges due to its high sensitivity to extraneous noise. For instance, the presence of background noise can lead to an unconscious increase in vocal effort by the speaker. Additionally, even minor movements, such as a slight head turn or a hand gesture near the mouth, can cause a noticeable decrease in the measured intensity. Variations in the distance between the speaker and the microphone, as well as differences in the type of microphone utilised for recording, can also substantially influence the intensity measurements. Despite these challenges, it still stands that intensity has been identified as an important factor in signalling prominence in spontaneous speech. Given the potential significance of prominence in evaluating speech rhythm, failing to explore the forensic potential of intensity-based methods would be somewhat defeatist.

It is conceivable that if the recording conditions for both a known sample and a questioned sample in a specific FVC case are found to be relatively stable, the examination of speakers' intensity patterns could contribute towards capturing distinctive rhythmic behaviours. Alternatively, measuring intensity over shorter time spans, such as individual speech units (such as those analysed in Chapter 4 of the present thesis), could diminish the likelihood of interference from the aforementioned problematic factors, thereby enabling the acquisition of more reliable and robust measurements that could potentially aid in assessing the significance of intensity in forensic casework.

Turning now to the small body of existing research which has considered intensity in terms of speaker individuality, He and Dellwo (2016) tested a number of rhythm metrics (those typically used for quantifying temporal information (e.g., normalised variation coefficients, normalised pairwise variability indices, means, standard deviations, etc.) to investigate between-syllable intensity variability (intensity means and peaks), whilst also looking at durational variability of vocalic and consonantal intervals as well as syllable-sized duration variability. In analysing both intensity measures and temporal measures, and making comparisons between the two, their results showed intensity measures to contain more speaker-specific information than durational measures, highlighting the importance of intensity features in between-speaker rhythmic differences.

Adopting slightly different methods, He and Dellwo (2017) and Zhang et al. (2021) examined intensity dynamics in terms of the speed of increases and decreases of amplitude between syllable peaks (the point at which intensity reaches its maximum value within a syllable) and inter-peak troughs (the point at which intensity reaches its minimum value between adjacent syllable peak intensities), thus unifying and capturing both intensity and temporal attributes simultaneously. They divided dynamics into positive (speed of increases in intensity from amplitude troughs to subsequent peaks) and negative (speed of decreases in intensity from amplitude peaks to subsequent troughs) subcategories and applied quantification metrics (means, standard deviations and pairwise variability indices) to evaluate the variability of these dynamics across utterances, finding that negative dynamics contained more speaker-specific information (around 70% of between-speaker variation was explained by measures of negative dynamics).

Adopting those methods employed by He and Dellwo (2017), Machado (2021) conducted a cross-linguistic study of between-speaker variability in intensity dynamics in L1 and L2 spontaneous speech. Although results showed that there was between-speaker variability in both of the languages studied, results indicated that for both languages positive and negative dynamics seemed almost equally able to explain inter-speaker variability (positive dynamics = 48%; negative dynamics = 52%), with this being attributed to the nature of the data and the greater degree of gestural overlap between the start and end of syllables in spontaneous speech. Further linear discriminant analyses of the intensity dynamics revealed low speaker classification rates in both of the languages, and although negative measures were the better classifiers for both languages (L1 = 4.8%; L2 = 4.4%), these classification rates were still only marginally above chance level (1.9%).

These studies all serve to highlight the role of intensity in between-speaker rhythmic variability, but what are the explanations for between-speaker variability in relation to intensity? Reasonings are twofold: firstly, one must consider the anatomical idiosyncrasies relating to a speaker's speech organs and vocal tract, and secondly the individualities in the way speakers operate their articulators. It is a combination of both these factors which inevitably lead to speaker idiosyncrasies manifesting within

the speech signal (Dellwo et al., 2007). Previous research has shown that the size of the mouth aperture relates directly to intensity in the speech signal (e.g., Chandrasekaran et al., 2009; He & Dellwo, 2017): the bigger the mouth opening the greater the intensity and vice versa. Similarly, studies on subglottic air pressure (Plant & Younger 2010) and pulmonic air pressure (Wilson & Leeper, 1992), both of which are intrinsically related to intensity in the speech signal, have shown a great deal of between-speaker variation even within tightly-controlled syllable contexts. These individualities give rise to speaker-specific pulmonic and sub-glottal pressure fluctuations which are in turn evidenced within the speech signal.

Studies which have focussed on such phenomena, and which have commented on the between-speaker variation evidenced, have predominantly, if not wholly, been obtained through controlled speech conditions (e.g., read speech). It stands to reason that individual differences in such phenomena may be even more prominent within spontaneous speech, for example as a result of idiosyncratic connected speech processes. In utilising spontaneous speech data, the present thesis will be able to address such speculations, whilst also testing the efficacy of rhythm metrics and the suitability of intensity-based rhythm research to speech data more relevant to forensic casework. Where previous forensically-motivated research has focussed on read speech (Leemann et al., 2014; He & Dellwo, 2016), meaning direct comparisons can be made between speakers' rhythmic attributes given the matched lexical contents of utterances, situations such as these are extremely rare within forensic casework.

Given that a speakers' perceived rhythm will be dependent largely on the content of what is being said, the application of acoustic measurements in relation to rhythm is perpetually difficult within casework and therefore descriptions within forensic reports are largely based on perceptual assessments alone. It will therefore be of interest to the forensic analyst to see how rhythm measurements based on intensity and duration transfer over to spontaneous, content-mismatched data such as that analysed here in this thesis.

Whilst previous research has shown that the analysis of speech rhythm through durational measures may be useful within forensic casework, given that temporal attributes are largely unaffected by degradation of the speech signal (e.g., Leemann et

al., 2014) or by voice disguise (Leemann & Kolly, 2015), there is comparatively less known about the potential forensic application of intensity measures. Research by Kolly and Dellwo (2014) does highlight the potential forensic relevance of intensity measures in observing that intensity patterns may not easily be manipulated by speakers (e.g., as a disguise strategy) due to lack of possible auditory feedback. However, it remains that further research into the tenability of intensity for forensic applications is much needed.

2.5.3. f_0

Unlike intensity, for which speaker individuality research is scarce within the forensic literature, studies into the speaker-specificity of fundamental frequency are much more prevalent, with this parameter being generally regarded as an important feature for analysis within forensic voice comparison casework (e.g., Braun, 1995; Hudson et al., 2007; Kinoshita et al., 2009; Leemann at al., 2014; Lindh & Eriksson, 2007; Nolan, 1983). Much of this forensically-motivated research has been focussed on issues pertinent to forensic voice comparison casework such as the extent of between-speaker variation within homogenous groups (e.g., Hudson et al., 2007; Lindh, 2006; Künzel, 1989; Skarnitzl & Vaňková, 2017) and factors which can influence within-speaker variation (e.g., physiological factors (e.g., age); psychological factors (e.g., emotional state); technical factors (e.g., effects of mobile phone transmitted speech); see Braun (1995) for a thorough overview). Within FVC, f_0 is typically quantified by providing measurements of a speaker's mean f_0, with this being seen as pointing towards a speaker's anatomy and physiology of the vocal folds, with another option being to measure a speaker's standard deviation for f_0, with this measurement relating more to behavioural choices adopted by a speaker whereby they might be placed on a scale with regards to whether their speaking manner is monotonous or melodic in nature (e.g., Hollien, 1990; Jessen, Köster & Gfroerer, 2005; Rose, 2002).

Studies which have examined f_0 in terms of speech rhythm and its role alongside other prosodic parameters such as duration and intensity have, for the most part, not been forensically focussed (f_0 and duration research has included: Cumming, 2011; Fuchs, 2014; Niebuhr & Winkler, 2017; f_0 and intensity research has included: Alku et al.,

2002; Köster, 2002; Lee & Todd, 2004; Plant & Younger, 2000; Traunmüller & Eriksson, 2000).

One forensic investigation with relevance to the present work is that of Jessen et al.'s (2005) large-scale study which examined the relationship between f_0 and intensity in terms of the influence vocal effort has on average f_0 and the variability of f_0. They found that an increase in vocal effort from neutral to loud speech resulted in increases in mean f_0 for all 100 of the speakers analysed in both spontaneous and read speech. Further analysis revealed that, even after differences in amplitude level were accounted for, the size of this effect differed between speakers, and that for 91 of the 100 speakers f_0 variability (the standard deviations of f_0) was higher in loud speech as opposed to neutral speech. These results marry well with the prior literature which has focussed on this relationship, and, although not specifically orientated towards describing between-speaker and within-speaker differences in relation to speech rhythm, provide good evidence for further investigating the relationship between f_0 and intensity, whilst also assessing whether one has more speaker-distinguishing potential than the other.

2.5.4. Perception studies

As the present thesis is comprised of both speech rhythm production experiments and speech rhythm perception experiments, there should be some background information on existing perception-based speech rhythm research which has looked to serve forensic applications. There is only a handful of studies which fall into this domain, with three of these being summarised below. These three studies were selected on the basis that they use a variety of methodologies and focus on different features associated with the perception of speech rhythm, with these methodologies and features being relevant to the present thesis (e.g., speech which has been degraded through different types of signal manipulation (delexicalised), and features such as pausing behaviour and speaking rate).

Kolly and Dellwo (2013) investigated the importance of different temporal and rhythmic prosodic characteristics for the recognition of French- and English-accented

German. In their experimental design, they used Swiss German listeners to judge stimuli which preserved only time domain characteristics and different degrees of rudimentary information from the frequency domain (i.e., stimuli devoid of linguistic content and voice quality characteristics). They created the stimuli using a variety of (delexicalisation) techniques in order to assess which cues within the signal were the most importance with regards to facilitating correct accent recognition from the listeners. They found that listeners could recognise French- and English-accented German above chance even when their access to segmental and spectral cues was strongly reduced. It was shown that different types of temporal cues led to different recognition scores, with segment durations found to be the most salient temporal cue for accent recognition. It was also determined that stimuli which contained fewer segmental and spectral cues led to lower accent recognition scores. In discussing the forensic implications of their findings, Kolly and Dellwo highlight that a good deal of the speech material which forensic practitioners work with is often degraded with the frequency domain information available being reduced. They point out the importance of foreign accent recognition as a means of narrowing down a group of suspects in cases where an expert must establish the geographical origin and identity of an individual based solely on their voice (i.e., speaker profiling). They explain that some individuals may use L2 speech as a form of voice disguise and therefore having an improved understanding as to what acoustic cues could be relevant for recognition of the individual's L1 is a desirable commodity within forensic voice comparison and speaker profiling cases.

Kolly et al. (2015) conducted a cross-linguistic study in which they assessed speech rhythm patterns from the perspective of speakers' pausing behaviour. 16 speakers of Zürich German were prompted to read 16 Zürich German sentences, 16 English sentences and 16 French sentences, which were subsequently analysed in terms of the number of pauses present within a sentence and the sum of the durations of all the pauses within a given sentence. Results showed that the fewest and the shortest pauses were produced in the speakers' native Zürich German speech and the most and the longest pauses were produced in their French speech, with pausing behaviour in English placed in between the two. Kolly et al. explain this finding by highlighting that speaking a second language is cognitively more demanding than speaking a first

language and draw upon previous research which has shown that increased cognitive load has an effect on both the number of pauses produced (Riazantseva, 2001) and the duration of the pauses (Grosjean, 1980). Moreover, the results obtained in relation to speakers' individual pausing behaviour showed promising forensic potential. Both pausing measures exhibited significant between-speaker variability on the one hand and little within-speaker variability on the other – both of which are desirable properties within the forensic domain. Kolly et al. point out that speech samples which feature in forensic casework are often of degraded quality which has an impact on the analysis of features realised in the spectral domain. This is in contrast to temporal features, such as pausing behaviour, which will be largely unaffected by such degraded conditions, thus making them a potentially promising feature for forensic voice comparison. Although this study is, in essence, a production study, its inclusion here (as opposed to alongside other production experiments reviewed) was deemed more appropriate as it helps situate the research alongside the related speech rhythm research of Kolly and colleagues. Furthermore, the results from this study are directly correlated with the findings obtained from the study described below.

In a follow-up study to the one outlined above, Kolly (2016) looked to determine the extent to which speakers' strength of foreign accent was speaker-specific across different non-native languages. The stimuli presented to listeners in this study were not subjected to any kind of manipulation (cf. Kolly & Dellwo (2013) above). However, its inclusion within the current review relating to speech rhythm is merited in light of the acoustic correlates which have been shown to be markers of perceived accent strength. As Kolly points out, features such as segment durations (Tajima et al., 1997; Holm, 2008; Quen´e & van Delft, 2010; Winters & O'Brien, 2013), pausing behaviour (Trofimovich & Baker, 2006), and speaking rate (Dellwo, 2010) increase the perception of foreign accent. Given that these three features (duration, pausing behaviour, and speaking rate) are all features which have been focussed on within speech rhythm research (both production and perception experiments), it is reasonable to suggest that listeners would be considering these features when assessing the speakers, and therefore, albeit potentially subconsciously, be considering speakers' speech rhythm. Using the same speech data as the aforementioned study (16 Zürich German speakers reading 10 sentences in English and 10 sentences in French), 16

native French listeners and 16 native English listeners were tasked with rating the read sentences for their native language.

Using a quasi-continuous scale comprised of 100 intervals, with each end of the scale labelled as ‘rather weak’ and ‘rather strong’, listeners were asked to rate the intensity of the speakers’ foreign accent. Results showed that speakers were perceived to have a stronger accent in their French non-native speech than in their English non-native speech, with this finding being related to the fact that Zürich German speakers are likely more proficient in English than in French. This finding seems to corroborate with the previous finding that these speakers produce more pauses and pauses of greater length when speaking their non-native French as opposed to non-native English. Another result that echoed the findings of the previous study was the significant effect of speaker on accent strength as rated by native listeners of French and English. That is, there is significant between-speaker variation evidenced along with little within-speaker variation – conditions which are highly desirable within forensic voice comparison. Kolly suggests that accent strength being speaker-specific could be a result of not only external factors such as age of acquisition, but also because of cognitive and social-psychological factors. In highlighting the forensic implications of the results obtained, Kolly notes that the speaker-specificity of accent strength could also be leveraged for forensic cases in which a speaker uses different non-native languages in different contexts, possibly in the presence of earwitnesses who may recall the strength of the speakers' accent. Although Kolly’s study here is not explicitly discussed in terms of speech rhythm, it is probable that the perception of foreign-accented speech will involve listeners making their assessments based on speech rhythm features, as described above.

2.5.5. Summary

As has been shown above, the pool of forensically-motivated speech rhythm research is a comparatively shallow one. Nevertheless, such studies have covered a range of different acoustic parameters such as duration, intensity and f_0 and have shown these, to a greater or lesser extent, to carry some potential within the forensic sphere. The results of these studies therefore lend support to the examination of speech rhythm

using these parameters, such is the focus of the production experiments of the present thesis. The previous subsection provided a summary of some of the forensically-motivated speech rhythm perception studies. Research in this very specific area is scarce, even more so than production-based studies. Nevertheless, this research has demonstrated that listeners are able to discriminate between different types of foreign accented speech when presented with speech which has been degraded through different types of signal manipulation. These findings lend support to the nature of the perception experiments in the present thesis in which listeners are tasked with making speaker identification assessments based on delexicalised speech material which conveys only rhythmic characteristics.

The following section returns focus to the nature of the production experiments of the present thesis. In particular, it summarises the rationale for extending the examination of spontaneous speech rhythm patterns to specific frequently occurring speech units. Discussion is then provided relating to previous forensic phonetics research which has investigated these units.

2.6. Frequently occurring speech units

Within spontaneous speech there is likely always to be some items which occur more frequently than others such as verbal pauses of hesitation (e.g., *er* and *erm*) as well as other verbal pauses (*you know*, *yeah*, *like*, *well*, etc.). Given that these speech units might serve different functions within a given utterance (e.g., to mark the start or end of a speech turn or to signal that a speech turn is not complete), it might be that they frequently occur in similar positions within a given speaker's utterances (e.g., Braun et al., 2023; Gósy, 2023). Some speakers might use items such as these subconsciously and consistently, such that their presence within the speech of an individual becomes a characteristic feature. The frequency of such items and the notion that they may punctuate the speech of an individual to serve common purposes (e.g., starting a turn) might mean that they play a role in an individual's speech patterns. Furthermore, these speech units might also often be marked by different prosodic inflections (depending

on their function) to a greater extent than the rest of an utterance (e.g., Benus et al., 2007; Grivičić & Nilep, 2004; Trouvain & Truong, 2012; Truong & Heylen, 2010)

As has been shown in the review of the forensically-motivated speech rhythm research presented previously, the vast majority of such research has been carried out on controlled, usually read, speech. Even where spontaneous speech has been examined, these studies tend to explicitly exclude material which contain items such as filled pauses and verbal fillers (e.g., Leemann, et al., 2014). The preference for using controlled speech material for forensic speech rhythm research is perhaps not surprising as this methodological setup will have speakers all producing the same speech material, allowing for within- and between-speaker variation to be assessed with greater ease. It is likely that using spontaneous speech material for determining the speaker-specificity of speech rhythm patterns will prove more problematic (i.e., yield less discriminatory potential) as this will involve comparisons being made across utterances that are different with respect to their phonetic content, level of stress and whole-utterance factors, for example. As all of these factors will contribute to the variables used to capture speech rhythm patterns, these measures are likely to be too sensitive to the variation that spontaneous, content-mismatched speech contains.

The present thesis looks to offer a potential solution to this problem by factoring frequently occurring speech units into the analysis of spontaneous speech rhythm. Given all of the factors associated with these speech units mentioned above (i.e., their prevalence, their discourse position patterning, their prosodic patterning, etc.), this thesis suggests that these units can serve as being potentially fruitful with regards to acting as 'anchors' or 'control units' from which idiosyncratic speech rhythm patterns could be determined. As such, four frequently speech units (*er*, *erm*, *yeah* and *no*) will be analysed in terms of their rhythmic characteristics and using a novel methodological procedure which will facilitate comparisons to be made between the spontaneous speech utterances analysed in the previous chapter. This thesis therefore hypothesises that directing focus towards these speech units could be a way in which we might start to measure, at least to some degree, spontaneous speech patterns.

The following subsections provide some background as to the forensic research which has been carried out with regards to the frequently occurring speech units studied in this thesis.

2.6.1. Filled pauses: *er* and *erm*

Disfluencies are a normal and natural part of communication and serve an important function in the planning, production and comprehension of speech, with a great deal of research being carried out along these lines (e.g., Blankenship & Kay, 1964; Brennan & Schober, 2001; Corley et al., 2007; Fraundorf & Watson, 2011; Goldman-Eisler, 1961; MacGregor et al., 2010; Shriberg, 1994, 1996, 2001). However, there is no evidence to suggest that all speakers plan, produce and comprehend speech in the same way, and therefore it is logical to assume that disfluency behaviour has the potential to vary from speaker to speaker. A speaker's disfluency behaviour may be influenced by a number of factors, such as the topic of conversation, a speaker's cognitive-processing ability, psycho- or socio-linguistic demands, along with possible psychological and prosodic explanations (e.g., Brennan & Schober 2001; Corley et al. 2007; Fraundorf & Watson 2011; Goldman-Eisler 1968). Accordingly, there is a good deal of non-forensic research in which individual variation in disfluency features has been observed. Such studies have found individual differences in the use of:

- **Silent pauses:** (e.g., Bortfeld et al., 2001; Butterworth, 1980; Goldman-Eisler, 1968; MacGregor, Corley & Donaldson, 2010).
- **Filled pauses:** (e.g., Fraundorf & Watson, 2011; Shriberg & Lickley, 1993; Rose & Watanabe, 2019).
- **Repetitions and revisions:** (e.g., Eklund, 2004; Fox Tree, 1995; Shriberg, 1995, 2001).
- **Prolongations:** (e.g., Betz, Eklund & Wagner, 2017; Betz & Wagner, 2016; Eklund, 2004).

The analysis of speakers' disfluency patterns is appealing to forensic casework due to the notion that disfluencies are a normal and natural part of speech and thus such phenomena will be difficult to deliberately and consistently disguise as speakers (and listeners) are generally unaware that they are occurring (Finlayson & Corley, 2012). Recent research into the speaker-specificity of speakers' disfluency behaviour has looked to account for the frequencies and types of disfluencies used by individual speakers, with McDougall and Duckworth (2017) devising a forensic framework, TOFFA, by which a range of disfluencies can be quantified. The application of this framework has gone on to show how speakers show consistency in their disfluency behaviour across different speaking styles (e.g., Carroll, 2019c; McDougall & Duckworth, 2018; however, cf., Harrington et al., 2021 who report the opposite trend) and the benefits of utilising this more objective taxonomy within real forensic casework (see McDougall et al., 2019).

The TOFFA framework is potentially a useful tool for the forensic analyst when its application is appropriate and feasible, nevertheless, in providing only the relative frequencies in which different types of disfluencies occur within speech samples, this approach fails to capture acoustic characteristics which may also contain useful speaker-specific information. Research which has sought to investigate the acoustic make-up of speech disfluencies has, perhaps unsurprisingly, been focussed largely on the filled pauses *er* and *erm*, with this being partly due to their assumed prevalence in relation to other disfluency phenomena, and also the comparative ease in obtaining reliable and replicable acoustic measurements. As such, filled pauses have been studied in relation to durational characteristics (e.g., Kaushik et al., 2010; Shriberg, 2001; Stouten & Martens, 2003), variability in formants F_1-F_3 (e.g., Audhkhasi et al., 2009; Brander, 2014; Foulkes et al., 2004; Hughes, et al., 2016; Kaushik et al., 2010), and variability in f_0 (e.g., Brander, 2014; Clark & Fox Tree, 2002; Shriberg & Lickley 1993; Tschäpe et al., 2005; Verkhodanova & Shapranov, 2016). In generalising the findings from these studies, filled pauses tend to be longer in duration and exhibit a lesser degree of F_1-F_3 and f_0 variability in comparison to other syllables, meaning that they are broadly realised as long, stable syllables of a low pitch.

Of these studies, one which looked to serve forensic purposes is Foulkes et al.'s (2004) investigation of the filled pauses *er* and *erm* in relation to their F_1-F_3 midpoint measurements. They examined 1,695 filled pauses within the spontaneous speech of 32 individuals from Newcastle upon Tyne, northern England, and did so in comparison to the lexical vowels /ɪ, ɛ, a, ə/ to determine whether filled pauses or lexical vowels possessed greater discriminatory value. They analysed their data using linear discriminant analysis and found that both filled pauses had classification rates close to or better than the best-performing lexical vowels studied, thus highlighting the discriminatory potential of these filled pauses (n.b., *er* and *erm* were treated as separate variables given that the nasal portion of *erm* was predicated to affect formant patterning; only the vocalic portion of *erm* were analysed; *er* performed better than *erm* at discriminating between speakers).

A later study, conducted by Hughes et al. (2016), also focussed on the variability of F_1-F_3 measurements, but as well as analysing the midpoints of the vocalic portions of *er* and *erm*, they also investigated the dynamic measurements of the formant trajectories (i.e., quadratic curves fitted to 9 measurement points over the full vowel), whilst also accounting for the duration of the vocalic potions of both filled pauses and the nasal portion of *erm*. The study made use of spontaneous speech data from a group of 60 male speakers of standard southern British English and results were presented using the likelihood ratio framework in line with what is argued to be the most appropriate for forensic casework (see Rose & Morrison, 2009). Their results showed that it was the dynamic measurements of all three F_1-F_3 formant trajectories across the vocalic portion of *erm*, combined with the durational measurements of both the vocalic and nasal portions which resulted in the best performance at distinguishing between speakers. Furthermore, they established that in general *erm* outperformed *er*, especially with regards to measurements of formant trajectories with this being attributed to the increased degree of formant movement in the former due to the transition from the vocalic to the nasal portion. For *er*, it was also noted that taking a dynamic approach did not improve system performance, but rather taking static midpoint measurements of the three formants was more useful. Finally, for both *er* and *erm*, combining durational information, whether with regards to the vocalic or the

vocalic and nasal portions of the filled pauses, always culminated in greater speaker-distinguishing potential.

Another study with forensic implications is Tschäpe et al.'s (2005) investigation of f_0 patterns within *er* and *erm.* They analysed the speech of 72 male German speakers performing a picture description task within two different speaking conditions (neutral speech vs. Lombard speech (e.g., speech which has an increase in vocal effort due to background noise or some other factor such as poor telephone transmission). They found that, for both speaking conditions, filled pauses exhibited a lesser degree of f_0 variability than that of speakers' intonational phrases, demonstrating a high degree of between-speaker variation and a low degree of within-speaker variation, both of which are desirable to the forensic analyst. Furthermore, the finding that there was low within-speaker f_0 variability for the filled pauses within the Lombard speech bolsters the view that acoustic analysis of filled pauses could be useful within FVC tasks which feature telephone transmitted speech given that a speaker's f_0 may be affected by the Lombard effect in such cases.

2.6.2. Monosyllabic responses: *yeah* and *no*

There has been very little forensic research which has focussed on the speech units *yeah* and *no*, with a search of the literature returning just two relevant studies pertaining to the former, and a complete void with regards to the latter.

In terms of *yeah*, Gibb-Reid et al. (2022) assessed the vowel formant dynamics of this speech unit in terms of F_1 and F_2 trajectories and found that the formant trajectories varied based on the function of the word as well as its positioning with respect to pauses. Additionally, their results showed an indication that *yeah* possessed distinctive formant trajectories across speakers as well as exhibiting low within-speaker variability. They use these findings to tentatively suggest that word-specific variation is worthy of further investigation with respect to the application to forensic voice comparison tasks.

Braun et al. (2023) conducted a study on the speaker-specificity of various speech disfluencies, particularly focusing on a group of "verbal fillers" where *yeah* (or rather

its German equivalent *ja*) and *und* were found to be the most frequent and relevant examples. The researchers analysed the frequency and placement of these verbal fillers within an utterance, as well as their f_0 relative to the surrounding context. They emphasised that these verbal fillers have been relatively overlooked compared to other disfluency phenomena like filled pauses. The study advocates for further research on these verbal fillers, as the findings suggest that they contribute to individual disfluency patterns and enhance the discriminatory potential of analysing disfluency behaviour.

There is, however, a body of research which has examined *yeah* (and to a lesser extent *no*) in relation to their polyfunctionality and associated discourse position patterns. As this is a factor in why these speech units have been selected for analysis in this thesis, these are briefly summarised here. In the research literature (mostly pertaining to discourse/conversation analysis), *yeah* has been shown to have many functions within dialogue such as acting as a backchannel (an indicator that the speaker is being listened to and may carry on with the conversation), an assessment item (evaluating something that was said previously), or a marker of speaker incipiency (an indicator of the speaker taking the conversational floor; see e.g., Drummond & Hopper, 1993; Gardner, 1998; Jefferson, 1984).

A small body of research which has investigated the prosodic makeup of *yeah* in relation to its conversational function has shown that these different functions carry with them different prosodic inflections in relation to intensity, f_0 and duration (e.g., Benus et al., 2007; Grivičić & Nilep, 2004; Trouvain & Truong, 2012; Truong & Heylen, 2010). Although there are a handful of studies which have looked into the potential multifunctionality of *no* (e.g., Jefferson, 2002; Lee-Goldman, 2011), this work has not looked into the associated acoustics of this specific speech unit. Despite this research not having any direct forensic relevance, nor being (for the most part) focussed on the acoustics of these speech units, its mention here is merited in support of the hypothesis that these units could be useful ‘anchor units’ or ‘control units’ from which spontaneous rhythm patterns can start to be measured.

2.7. Chapter summary

The present chapter has provided a review of the speech rhythm literature and highlighted how research within the field progressed over time as a result of methodological developments and shifts in focus. Following a review of the more generalised speech rhythm research, a more focussed approach was taken which looked more closely at research which has been dedicated to the forensic implications and applications of such research. From this forensic perspective, the literature was reviewed in relation to the three parameters most commonly attributed to speech rhythm (intensity, f_0 and duration) as well as for both production- and perception-based studies. Finally, a review of the literature pertaining to four, so-called, "frequently occurring speech units" was also provided owing to the proposition put forward that these speech units could be key in facilitating an approach by which spontaneous speech rhythm patterns might be quantified more robustly.

CHAPTER 3

Speech Rhythm in Spontaneous Speech

3.1. Introduction

This chapter details the findings relating to the speaker discriminatory potential of speech rhythm measurements across spontaneous speech utterances. As discussed in Chapter 2, Section 2.5, previous forensic research which has sought to capture speech rhythm patterns for the purpose of speaker discrimination has taken account of intensity, f_0 and duration measurements and, as such, the analysis conducted in the present chapter also focuses on these three parameters. Where the vast majority of this previous work has made use of controlled, usually read speech, the analysis which follows seeks to determine how well these measures transfer over to spontaneous, content-mismatched speech which is predominantly the type of material encountered in forensic casework. At present, if an individual's speech rhythm is a feature which a forensic expert wishes to analyse in a given FVC case, any such patterns will only be described at an impressionistic level. The experiments which follow therefore look to assess whether there are acoustic cues that could capture speech rhythm in spontaneous speech and subsequently be used to discriminate between speakers in forensic casework.

In consideration of the forensically-motivated rhythm research which has investigated the three parameters under study in the present chapter (See Chapter 2, Section 2.5), it is hypothesised that certain parameters may perform better with regards to

distinguishing between speakers than others. Specifically, given that intensity has been shown to demonstrate more speaker discriminatory potential than duration (see e.g., He and Dellwo, 2016), measures of intensity are proposed as potentially being more useful speaker-discriminators than measures of duration. The degree to which one parameter might perform 'better' than another, however, is somewhat more difficult to predict. Given the spontaneous, content-mismatched nature of the data under analysis, one can predict with a greater deal of confidence that it is unlikely that any one parameter will exhibit the capacity to categorically discriminate between all of the speakers under investigation. Nevertheless, it remains that the efficacy of these measures should be tested on spontaneous speech data as a natural next step forward from previous studies which have made use of controlled, usually read, speech data. That is, as prior research has demonstrated that these measures harness (to greater or lesser extents) speaker discriminatory power when applied to controlled, laboratory quality speech data, at present there has yet to be any thorough testing as to their potential for distinguishing between speakers using forensically-relevant speech data.

Given that previous research has often made use of rhythm metrics as a means of quantifying the measurements of different rhythm parameters, the experiments conducted in the present chapter will also look to test whether the quantification of measurements through these means leads to improved speaker discrimination rates (in comparison to a contour-approach (see below for elaboration on the contour-approach vs. variability-approach). As the application of these metrics has been shown to be of use within the few studies that have made use of them for testing between-speaker rhythmic variability, it is hypothesised here that the application of one (or more) of these quantification metrics may lead to higher speaker discrimination rates.

The final hypothesis posed for the present chapter concerns the application of dynamic measurements to the speech data (as opposed to static by-syllable measurements). Given that the dynamic measures applied in this chapter relate to combining measures of intensity and duration together across the spontaneous utterances (see Section 3.2.4.4.), it is hypothesised that dynamic measurements may well yield higher speaker discrimination rates in comparison to static measurements. It is further predicted that the application of quantification metrics to the dynamic measurements may also result

in improved discriminatory potential in comparison to the 'contour-approach' method (see below for elaboration on the contour-approach vs. variability-approach).

The structure of this chapter is as follows. The first section outlines the materials used and the methodological procedures followed including details of the speakers, the data extraction and editing procedures, the measurements taken, and the statistical analyses conducted. Following this, the results are then presented. Firstly, results from the static syllabic measures of intensity, f_0 and duration are provided before the overall findings relating to these measures are summarised. The results of the dynamic intensity measures are then presented and summarised with the chapter being concluded with an overall discussion and summary of the results from both static and dynamic measures as a whole.

3.2. Methodology

The following subsections provide detail relating to the materials used in the present chapter, the methodological procedures followed, and the statistical analyses carried out. Firstly, in section 3.2.1, details of the corpus from which recordings were obtained are provided, including details of the speakers, the type and amount of speech elicited, and the recording methods used. Following this, section 3.2.2 describes the nature of the utterances which were extracted from the interview data and the criteria used to ensure comparable data between speakers. Section 3.2.3 provides details on how the data were prepared and how syllables were segmented. For each of the syllables in the dataset, three acoustic parameters were analysed: intensity, f_0 and duration. Section 3.2.4 defines each acoustic parameter in turn and explains how each parameter was measured. Following this, in section 3.2.5, the normalisation procedures implemented are explained, with section 3.2.6 detailing the statistical methods used to analyse the data.

3.2.1. The WYRED corpus

The West Yorkshire Regional English Database (WYRED; Gold et al., 2018) is the largest forensically-relevant collection of Northern British English speech, consisting of 180 male speakers of West Yorkshire English between the ages of 18-30, divided evenly across three boroughs: Bradford, Kirklees, and Wakefield. Speakers produced four samples of spontaneous speech, three of which were under simulated forensic conditions, including a mock police interview (Task 1), a telephone conversation with an "accomplice" (Task 2) and a voicemail message relating to the fictitious crime (Task 4). Task 3 is a non-crime related discussion between a speaker and another participant (or in some cases a friend). The data analysed in the present chapter were obtained from 20 speakers from the borough of Bradford who were undertaking the mock police interview task in which speakers were being questioned by a research assistant imitating a police officer for approximately 20 minutes. Speakers wore a Sennheiser HSP 4 omnidirectional headband microphone situated approximately 2 cm from their mouth and recordings were made on a Marantz PMD661 MKII Handheld Solid State Recorder in PCM WAV format (44.1kHz, 16 bit). The 20 speakers were selected on the basis that they produced the required quantity of speech data needed for analysis (see Section 3.2.2).

3.2.2. Utterance length

For each speaker, 18 utterances of nine syllables were extracted from the mock police interview data. 9-syllable utterances were judged as appropriate in line with previous speech rhythm research (e.g., Dellwo et al., 2012; He & Dellwo, 2017). Originally, a target of 20 utterances per speaker was set, however, 18 utterances per speaker was the highest amount possible in order to keep a balanced dataset following the removal of problematic utterances (see Section 3.2.2.2). The audio data were analysed within Praat (Boersma & Weenink, 2020) and all utterances of nine syllables were isolated and extracted from the mock interviews. All utterances were declarative responses to questions being asked by the research assistant. The isolated utterances had to form meaningful units and be free from filled pauses and unfilled pauses (>200ms), however, given the challenge in obtaining these requirements, there were no further

formal criteria for the identification of utterances as complete units. To provide an example, in response to the "police officer" asking a question such as, "What did you and your brother do last night?", a response such as, "Watched T.V. I think for the most part" would be acceptable, whereas a response such as "We watched T.V. for a bit and then-" would not be acceptable as the utterance is incomplete (regardless as to whether the speaker ended their turn at this point). In total, 360 utterances were extracted (18 utterances × 20 speakers), meaning the dataset consisted of 3240 syllables (9 syllables × 18 utterances × 20 speakers).

3.2.3. Data preparation

Utterances were first transcribed orthographically and then force-aligned and segmented using the WEBMAUS Basic online interface (Kisler et al., 2017). This segmentation provided a visual guide to the initial placement of syllable boundaries; however, all syllabification was adjusted manually based upon phonetic criteria (acoustic cues drawn from the waveform and spectrogram along with auditory judgement). Syllables were judged as being the most suitable unit from which to obtain measurements in line with previous speech rhythm research (e.g., He & Dellwo, 2016; Leemann et al., 2014). Figure 3.1 shows the waveform, spectrogram and TextGrid of a 9-syllable utterance uttered by speaker WY171.

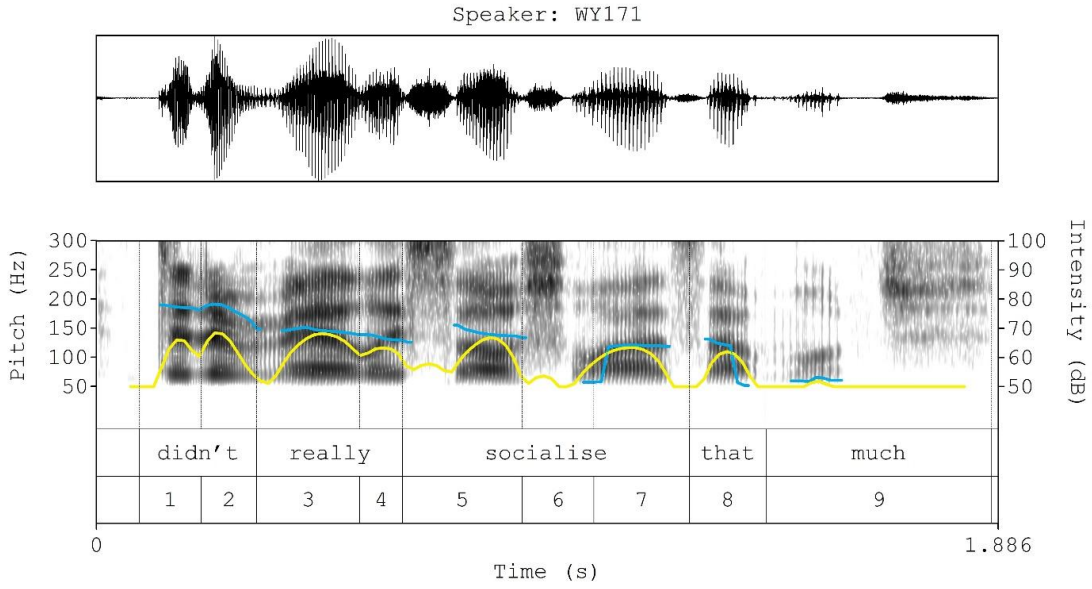


Figure 3.1. Waveform, spectrogram and TextGrid of a 9-syllable utterance uttered by speaker WY171. Tier 1 shows the orthographic transcription and Tier 2 shows the syllable tier from which the static intensity and f_0 measurements (mean, peak and trough) and syllable durations were derived.

3.2.4. Acoustic parameters

Rhythmic patterning across utterances was accounted for by taking static measurements of syllabic intensity and f_0 measurements as well as measuring syllabic duration. Dynamic measurements of intensity were also calculated. The following subsections detail how each of these parameters was measured.

3.2.4.1. Intensity

Intensity measurements were conducted within Praat through the use of a script which was written by the present author. This Praat script used the 'get intensity' function of the Praat software in order to obtain the following measurements:

(1) Mean intensity of the syllable: the mean (in dB) of the intensity values of the frames within the specified time domain (averaging method = "dB").

(2) Maximum (peak) intensity of the syllable: the maximum value within the specified time domain, expressed in dB (interpolation method = "cubic").

(3) Minimum (trough) intensity of the syllable: the maximum value within the specified time domain, expressed in dB of the syllable (interpolation method = "cubic").

The algorithm Praat uses to calculate intensity is as follows:

The values in the sound are first squared, then convolved with a Gaussian analysis window (Kaiser-20; sidelobes below -190 dB). The effective duration of this analysis window is 3.2 / *pitchFloor*, which will guarantee that a periodic signal is analysed as having a pitch-synchronous intensity ripple not greater than 0.00001 dB (Boersma and Weenink (2020).

Praat's standard settings were used.

The algorithm Praat uses to calculate mean intensity (using the dB method) is as follows:

$$1/(t_2 - t_1) \int_{t1}^{t2} x(t)\, dt$$

Where:

(t_1, t_2) constitute the time range (Boersma and Weenink (2020).

3.2.4.2. f_0

Measurements of f_0 were obtained using VoiceSauce (Shue, 2011), an application, implemented in MATLAB, which provides automated voice measurements over time from audio recordings. This application permitted measurements of f_0 using four different algorithms. Each of these algorithms, along with measuring f_0 using Praat, was tested on the data to determine which provided the most consistent and robust measurements. The measurements obtained from each of these different methods were manually checked alongside the corresponding audio files and spectrograms

(visualised within Praat) in order to determine the extent to which incorrect measurements were being calculated (e.g., unrealistic values being attributed to voiceless or creaky segments). Overall, the 'STRAIGHT' algorithm (Kawahara et al., 1998) proving most reliable. This method was deemed the most reliable as it produced the fewest erroneous values and fewest instances in which no f_0 reading could be obtained. See Kawahara et al. (1998) for details pertaining to how f_0 is calculated using this method.

The following f_0 measurements were calculated:

(1) Mean f_0 of the syllable: the mean (in Hz) of the f_0 values of the frames within the specified time domain.

(2) Maximum (peak) f_0 of the syllable: the maximum value within the specified time domain, expressed in Hz.

(3) Minimum (trough) f_0 of the syllable: the minimum value within the specified time domain, expressed in Hz.

VoiceSauce permitted manual adjustment of some settings, with the following being applied:

(1) The frame duration was set at 10ms (i.e., 100 pitch values were computed per second (Praat standard setting)).

(2) The pitch floor was set at 75 Hz (Praat standard setting)

(3) The pitch ceiling value was set at 300 Hz (Praat recommended setting for male voices (Boersma & Weenink, 2020)).

In addition to applying the settings described above, a sample of the data were inspected prior to any measurements being taken to ensure that this homogenous group of speakers' average pitch was within the upper and lower limits imposed and

that these settings did not result in any categorical or consistent f_0 tracking errors (e.g., octave jumping).

Nevertheless, as alluded to above, on occasion, the automatic extraction of f_0 produced erroneous values due to factors such as creak and voiceless segments causing tracking errors. In order to remove errors of this type, the raw data were inspected and unrealistic values such as values with improbable shifts from one syllable to the next were manually removed. In order to preserve as many tokens as possible for analysis (rather than the more reductive approach of removing utterances entirely), missing values were replaced with the mean of the two adjacent syllable values. Where missing values occurred at the in the initial or final syllable, the entire utterance was removed. This process removed a total of 36 utterances from the analysis across the 20 speakers meaning that the highest number of 9-syllable utterances that could be obtained for every speaker was 18 (the original target being 20 utterances).

3.2.4.3. Duration

Absolute durations of each syllable within an utterance were obtained using a Praat script written by the present author, by calculating the duration of the interval between the marked onset and offset points of each syllable.

3.2.4.4. Intensity dynamics

Intensity dynamics were calculated within Praat and follow the procedure outlined in He and Dellwo (2017). The following procedure (and associated wording) is taken for the most part verbatim from He and Dellwo (2017, p. 141) and is as follows:

(1) Peak points (*timeP* in Figure 3.2) were placed at the maximum intensity between syllable boundaries and trough points (*timeT* in Figure 3.2) were placed at the minimum intensity between adjacent peak points.

(2) The intensity values at each peak and trough points (*intP* and *intT* in Figure. 3.2) were obtained from the intensity curve at each *timeP* and *timeT* using cubic interpolation.

(3) Peak and trough points (*timeP* and *timeT*) and their associated intensity values (*intP* and intT) were obtained from each utterance.

(4) Positive dynamics (posDyn) were defined as $v_I[+] \stackrel{\text{def}}{=} (intP - intT) / (timeP - timeT)$, where *intP* and *intT* refer to the intensity values at peak and trough points represented by timeP and timeT.

(5) Similarly, negative dynamics (negDyn) were defined as $v_I[-] \stackrel{\text{def}}{=} |intT - intP| / (timeT - timeP)$.

(6) Absolute values were taken given that the magnitude was the point of focus.

(7) Thus, the speed of intensity increases, and decreases were measured.

(8) Geometrically, $v_I[+]$ and $v_I[-]$ can be demonstrated as the secant lines intT→intP and intP→intT in Figure 3.2 and the steepness of these lines were measured.

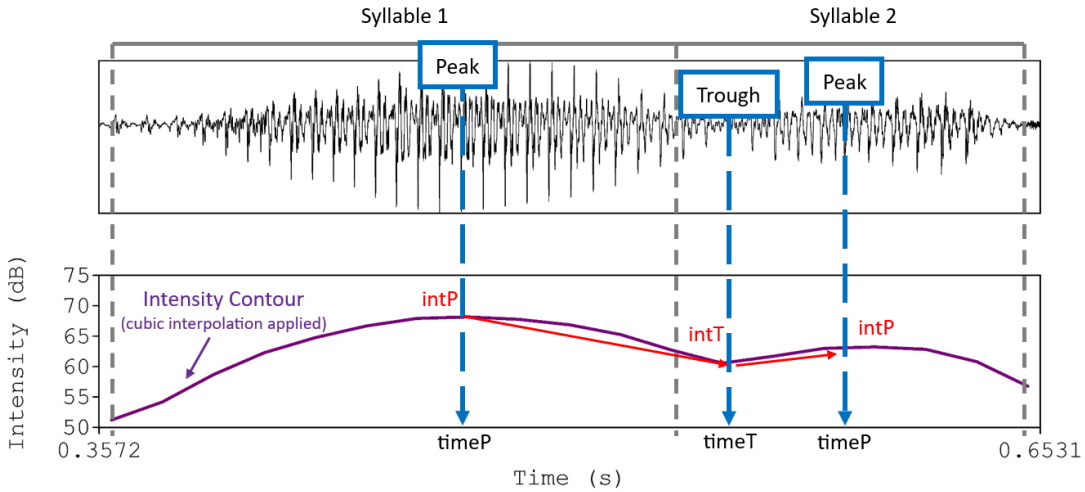


Figure 3.2. Illustration of calculating positive and negative intensity dynamics from a speech signal. The intensity contour (lower plot) was calculated from the speech waveform (upper plot). The amplitude envelope (superimposed over the waveform in the upper plot) was used to facilitate locating the peak and trough points (timeP and timeT). The peak and trough intensity values (intP and intT) were obtained from the intensity contour at timeP and timeT using the cubic interpolation. Intensity dynamics were calculated as how fast the intensity level dropped from a peak to its adjacent trough (intP→intT, i.e., negative dynamics), or increased from a trough to its adjacent peak (intT→ intP, i.e., positive dynamics).

Given that each utterance was nine syllables in length, there were 8 positive and 8 negative dynamic measurements calculated per utterance. Negative and positive dynamics were analysed as separate entities (following He & Dellwo (2017)) in which an utterance would be represented by 8 values (i.e., either the 8 negative or 8 positive dynamics of that utterance), and also considered them together as would be the natural sequence of the increases and decreases throughout an utterance (i.e., 16 dynamics (values) in total per utterance). Similar to the static analysis, the dynamic measurements were analysed through a contour-approach in which the raw dynamic values obtained were subjected to z-score normalisation (see Section 3.2.5 below), and also through using two quantification metrics:

(1) Standard deviations:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Where:

- s = Sample standard deviation symbol
- $\bar{x}$ = Arithmetic mean of the observations
- n = total number of observations

(2) Variation coefficients (varcos):

$$varco = \frac{stdev}{mean} \times 100$$

These two metrics were selected in line with previous speech rhythm research which has focussed on intensity dynamics (e.g., He & Dellwo, 2017).

In total there were 11 different types/combinations of dynamics subjected to linear discriminant analysis:

(1) pos – the positive dynamics of an utterance (8 values, z-score normalised)

(2) neg – the negative dynamics of an utterance (8 values, z-score normalised)

(3) both – the positive and negative dynamics of an utterance (16 values, z-score normalised)

(4) pos_stdev – the standard deviation of the 8 raw positive dynamic values of an utterance (1 value per utterance, with the 18 values in total per speaker being z-score normalised subsequently)

(5) neg_stdev – the standard deviation of the 8 raw negative dynamic values of an utterance (1 value per utterance, with the 18 values in total per speaker being z-score normalised subsequently)

(6) both_stdev – the standard deviation of the 16 raw positive and negative dynamic values of an utterance (1 value per utterance, with the 18 values in total per speaker being z-score normalised subsequently)

(7) pos + neg_stdev – the combination of the z-scores for (4) and (5) (36 values in total per speaker)

(8) pos_varco – the variation coefficient of the 8 raw positive dynamic values of an utterance (1 value per utterance, with the 18 values in total per speaker being z-score normalised subsequently)

(9) neg_varco – the variation coefficient of the 8 raw negative dynamic values of an utterance (1 value per utterance, with the 18 values in total per speaker being z-score normalised subsequently)

(10) both_varco – the variation coefficient of the 16 raw positive and negative dynamic values of an utterance (1 value per utterance, with the 18 values in total per speaker being z-score normalised subsequently)

(11) pos + neg_varco – the combination of the z-scores for (8) and (9) (36 values in total per speaker).

3.2.5. Normalisation

All raw static measures (means, peaks, troughs and syllable duration) and dynamic measures were subjected to z-score normalisation (by-speaker)in order to control for effects such as imprecisions in the distance between mouth and microphone, articulation rate and the likelihood that some speakers will be inherently louder or quieter than others. This normalisation method was deemed appropriate in order to isolate the features in focus for this study.

For a particular measure, the z-score of a particular syllable, or of a particular intensity dynamic, *k*, was calculated as:

$$z_k = \frac{(y_k - \bar{y}_k)}{\sigma_k}$$

Where:

y_k = the raw value of the syllable;

$\bar{y}_k$ = the mean of the nine raw values across the utterance;

σ_k = the standard deviation of the nine raw values across the utterance.

In line with previous research which has looked to assess the variability of measurements through quantification metrics (e.g., He and Dellwo (2016)), one of these metrics – varco (variation coefficient) – was applied to the static measurement data (see Section 3.2.4.4 above for the metrics used on the dynamic intensity data) to determine how such quantification compares to considering a contour as a whole. This particular quantification metric was selected on the basis that previous research (e.g., He & Dellwo, 2016, Wiget et al., 2010) has shown it to capture more between-speaker variation than other metrics such as Pairwise Variability Indices (PVIs).

For the static measurements of intensity and f_0, the following was calculated:

- The normalised variation coefficient of mean (varcoM), peak (varcoP) and trough (varcoT) syllable intensity / f_0 for each utterance.

Similarly, the following measurement was calculated for syllable durations:

- The rate-normalised variation coefficient of syllable durations (varcoSyll).

3.2.6. Statistical analysis

All of the following statistical analyses were carried out using R (R Core Team, 2019). In order to statistically test the extent of speaker-specificity exhibited by the spontaneous utterances, discriminant analysis was used for the 20 speakers under examination. The R package MASS (Venables & Ripley, 2002) was used to carry out the linear discriminant analyses – a multivariate technique used to assess whether a set of predictors can be combined to predict membership to a specific group (see Tabachnick & Fidell, 2014, ch.9). For example, for the static measurements of the spontaneous utterances, predictors are the nine sequential values attributed to a given utterance (using the contour-approach), whereas for the variability-approach, predictors are the singular value calculated for a given utterance (i.e., the variation coefficient). A 'group' is an individual speaker as represented by the collection of 18 utterances analysed for the study. The discriminant analysis procedure constructs discriminant functions which can be used to allocate each 9-syllable intensity/f_0/duration contour/individual value in the data to one of the speakers and determines a 'classification rate' according to the accuracy of the allocation. The 'leave-one out' method was used such that each intensity/f_0/duration contour/individual value was allocated to a speaker using discriminant functions calculated from all contours/values except the contour/value itself. A classification rate was then calculated to reflect the accuracy of this allocation process.

It is worth noting here that studies which deal with speech material for forensic purposes, specifically those whose analysis is intended for application within forensic casework, should use a likelihood ratio approach (see, for e.g., Morrison, 2014; Robertson & Vignaux, 1995; Rose & Morrison, 2009). However, the likelihood ratio approach is one which functions most effectively with larger groups of speakers (e.g.,

Hughes, 2014; Ishihara & Kinoshita, 2010), with Hughes (2017) finding that stable LR output was only achieved with more than 20 speakers. Therefore, the implementation of linear discriminant analysis in the present work was deemed appropriate in offering an initial statistical exploration, which, if merited, could be built on to include likelihood ratio analysis in future research of a larger scale.

In order to model the intensity, f_0 and duration contours for the 9-syllable spontaneous utterances, Generalised Additive Mixed Models (GAMMs; Wood, 2017) were used, which allow for the modelling of sequential, non-linear effects over time (e.g., see Sóskuthy, 2017). GAMMs were fitted in order to observe between-speaker and within-speaker variation for intensity and f_0 measures (mean, peak and trough) as well as durational measures for the spontaneous utterances. Additional models were also fitted in order to test the significance of the interactions between intensity, f_0 and duration measures. All GAMMs models were fitted using the R package mgcv (Wood, 2017).

For the GAMMs analysis, the data set consists of intensity (mean, peak and trough), f_0 (mean, peak and trough), and duration trajectories measured at each of the nine syllables across a given utterance. The data set also contains the following variables:

- syllable: see above
- utteranceID: a grouping variable by trajectory
- speaker: a grouping variable by speaker

In relation to the 'syllable' variable, it is recognised that this variable could be categorised as an ordered categorical variable which for which using ordinal GAMMs (as opposed to conventional GAMMs) could be deemed more appropriate. However, for the present study, no assumption was made that there necessarily be a close ordinal relationship between syllable numbers, with this being deemed appropriate for dealing with spontaneous speech.

As an example, the following R notation corresponds to the model fitted for peak intensity (int_peak) for which the result can be seen in Table 3.2:

```
int_peak ~ speaker + s(syllable, bs = "cr", k = 8) +
s(syllable, by = speaker, bs = "cr", k = 8) +
s(syllable, utteranceID, bs = "fs", m = 1, k = 8)
```

In order to assess the significance of interactions between measures, so-called 'tensor product interactions' (indicated by `ti` in the notation below) were used following the procedure given by Sóskuthy (2017). As an example, the following R notation corresponds to the model fitted for peak intensity and its interaction with duration for which the result can be observed in Table 3.2:

```
int_peak ~ speaker + s(syllable, bs = "cr", k = 8) +
s(syllable, by = speaker, bs = "cr", k = 8) +
s(duration, bs = "fs", k = 8) +
ti(syllable, duration, k = 8) +
s(syllable, utteranceID, bs = "fs", m = 1, k = 8)
```

Following the procedures outlined by Sóskuthy (2017), the issue of autocorrelation in trajectories was addressed using an autoregressive error model. For all GAMMs models fitted, where *p*-values are reported, $\alpha = 0.05$ in line with other dynamic speech analysis research which has utilised GAMMs (e.g., Sóskuthy, 2017, 2021).

3.3. Results

The analyses carried out over the following section serve the purpose of attempting to measure speech rhythm in spontaneous speech. Accordingly, measurements of intensity, f_0 and duration, the three parameters most commonly associated with speech rhythm, are subjected to statistical analyses in order to investigate variation exhibited between speakers. In particular, given the forensic motivations of this work, the following analyses focus on individual speaker variation and whether speakers can be distinguished from one another through measurements of intensity, f_0 and duration. These analyses will subsequently reveal whether measurements of a certain parameter are more useful than that of another, or whether it is the combinations and interrelations of these parameters which signal speaker individuality.

The degree to which one parameter might perform 'better' than another, however, is somewhat more difficult to predict. Given the spontaneous, content-mismatched nature of the data under analysis, one can predict with a greater deal of confidence that it is unlikely that any one parameter will exhibit the capacity to categorically discriminate between all of the speakers under investigation

In addition, measurements will be taken in a number of different forms (e.g., means, peaks, troughs, dynamic measurements, etc.) in order to determine which present as being the most useful in capturing variation between speakers. As indicated at the beginning of the present chapter, it is predicted that it is unlikely that any one single parameter will exhibit the capacity to categorically discriminate between all of the speakers. However, as previous research has shown measures of intensity to discriminate between speakers with greater proficiency than durational measures, it is speculated that measures of intensity may show the greater discriminatory potential.

The 9-syllable spontaneous speech utterances were analysed to determine whether any speaker-specific patterning could be accounted for in relation to largely uncontrolled, content-mismatched data. Static measurements of syllabic intensity, f_0 and duration are analysed first in Section 3.3.1 to determine which harnesses the most speaker discriminatory potential, before the parameters of intensity and duration are unified in a dynamic approach in Section 3.3.2 to assess whether intensity dynamics are more useful than static measurements.

3.3.1. Static syllable measures

Figure 3.3 shows the classification rates yielded from the linear discriminant analysis results for each of the rhythm measures calculated, ranked from highest to lowest.

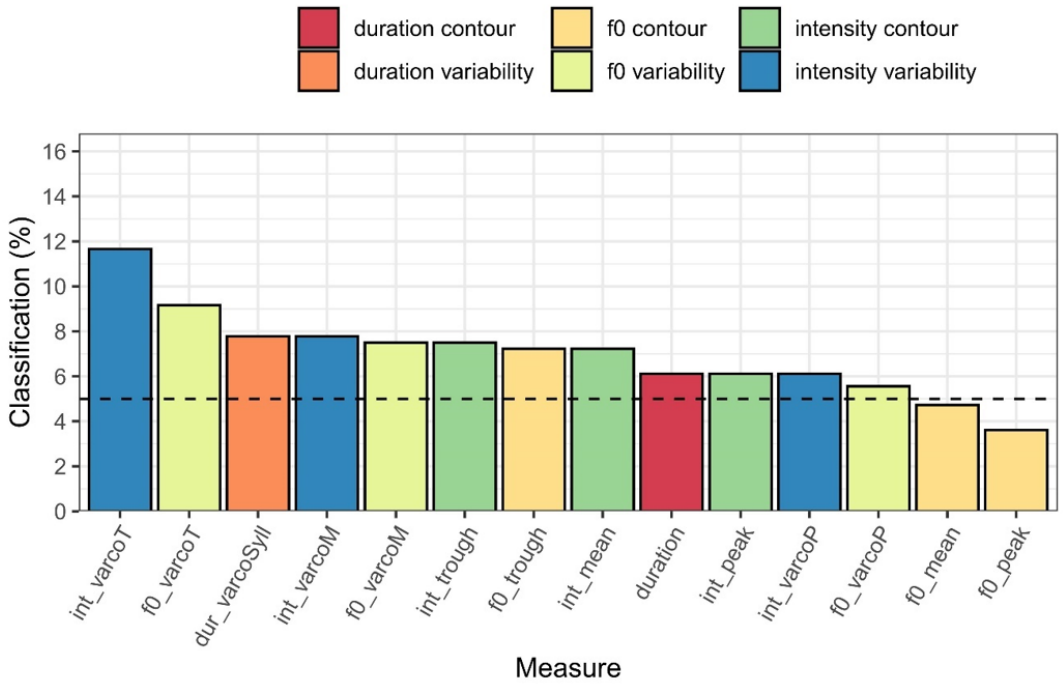


Figure 3.3. Discriminant analysis classification rates for each of the rhythm measures calculated (ranked from highest to lowest). Chance level is 5% (indicated by dotted line) as there are 20 speakers. N.b., chance level is calculated as the maximum probability (1) divided by the number of speakers (20). Therefore, 1/20 = 0.05, which, when expressed as a percentage, gives 5%. Where chance level is expressed in further plots throughout the thesis, calculations follow the same method.

The results indicate that no one single measure has the ability to distinguish all twenty speakers from one another. Some measures are shown to perform better than others, and it is the measures quantified by the single-value variation coefficient which present as most promising (top five results are varcos). There is, however, no clear pattern with regards to which rhythm parameter (intensity, f_0, duration) is the best performing as the top three results are spread over the three parameters. One trend that is apparent is that it appears to be trough measures, those associated with minimum syllabic measurements, which distinguish between speakers most competently. This is the case with regards to both the variability-approach (i.e., int_varcoT and f_0_varcoT) and the contour-approach (i.e., int_trough and f_0_trough). In general, peak measures appear to be the least encouraging although the results for many of the measures are evidently very much alike. Figure 3.4 provides an illustration of the LDA results for the peak intensity measures within the contour-approach. It can be observed

that there are no clear patterns or individual speaker clusters apparent, with speakers' values distributed sporadically across the plot in addition to there being areas which all 20 speakers' values occupy (highlighted in the upper left of the plot in the top panel).

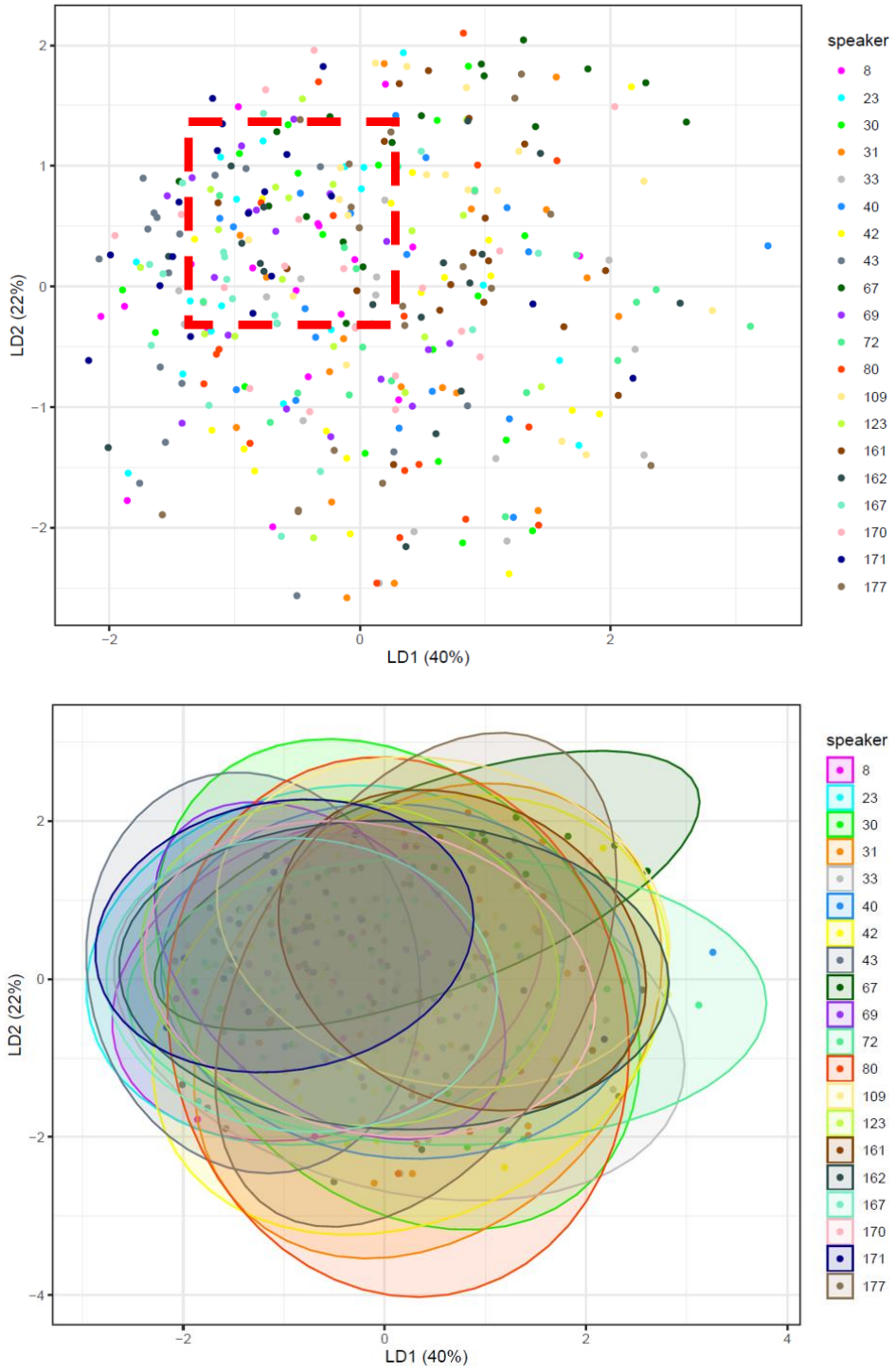


Figure 3.4. Visualisation of the LDA results for the peak intensity measures within the contour-approach. Highlighted in the upper left of the plot in the top panel is an area occupied by all 20 speakers. Bottom panel plot includes ellipses to further emphasise the sporadic distribution of speakers' data values. Overall classification rate = 6.1% (chance = 5% as there are 20 speakers).

A univariate ANOVA testing speaker by measure was calculated for those rhythm measures attributed to the variability-approach. ANOVAs were selected instead of LMEs as the latter cannot be calculated if the number of observations per speaker (18) is equal or less than the number of speakers (20). Table 3.1 displays the summary of these results. The contour-approach measurements were not tested in this fashion given the in-depth analysis afforded to through the application of GAMMs.

Table 3.1. Summary of statistics for the tested rhythm measures from the variability-approach.

Measure Type	Measure	Test	Factor tested	Result
Duration	dur_varcoSyll	One-way ANOVA	Speaker	F = 2.22, p = 0.145
f_0	f_0_varcoM	One-way ANOVA	Speaker	F = 3.03, p = 0.082
f_0	f_0_varcoP	One-way ANOVA	Speaker	F = 1.75, p = 0.397
f_0	f_0_varcoT	One-way ANOVA	Speaker	F = *23*.58, p < .0001
Intensity	int_varcoM	One-way ANOVA	Speaker	F = 0.64, p = 0.506
Intensity	int_varcoP	One-way ANOVA	Speaker	F = 1.65, p = 0.168
intensity	int_varcoT	One-way ANOVA	Speaker	F = 37.97, p < .0001

Only two (f_0_varcoT and int_varcoT) of the seven measures showed significant effects of speaker. In line with the results yielded from the LDA analyses above, it is those measures associated with syllabic trough measurements which appear the most promising. To investigate this trend further, and for a visual comparison across these metrics, Figure 3.5 shows the boxplots for the 20 speakers in relation to their f_0_varcoT (panel a) and f_0_varcoP measures across their 18 utterances.

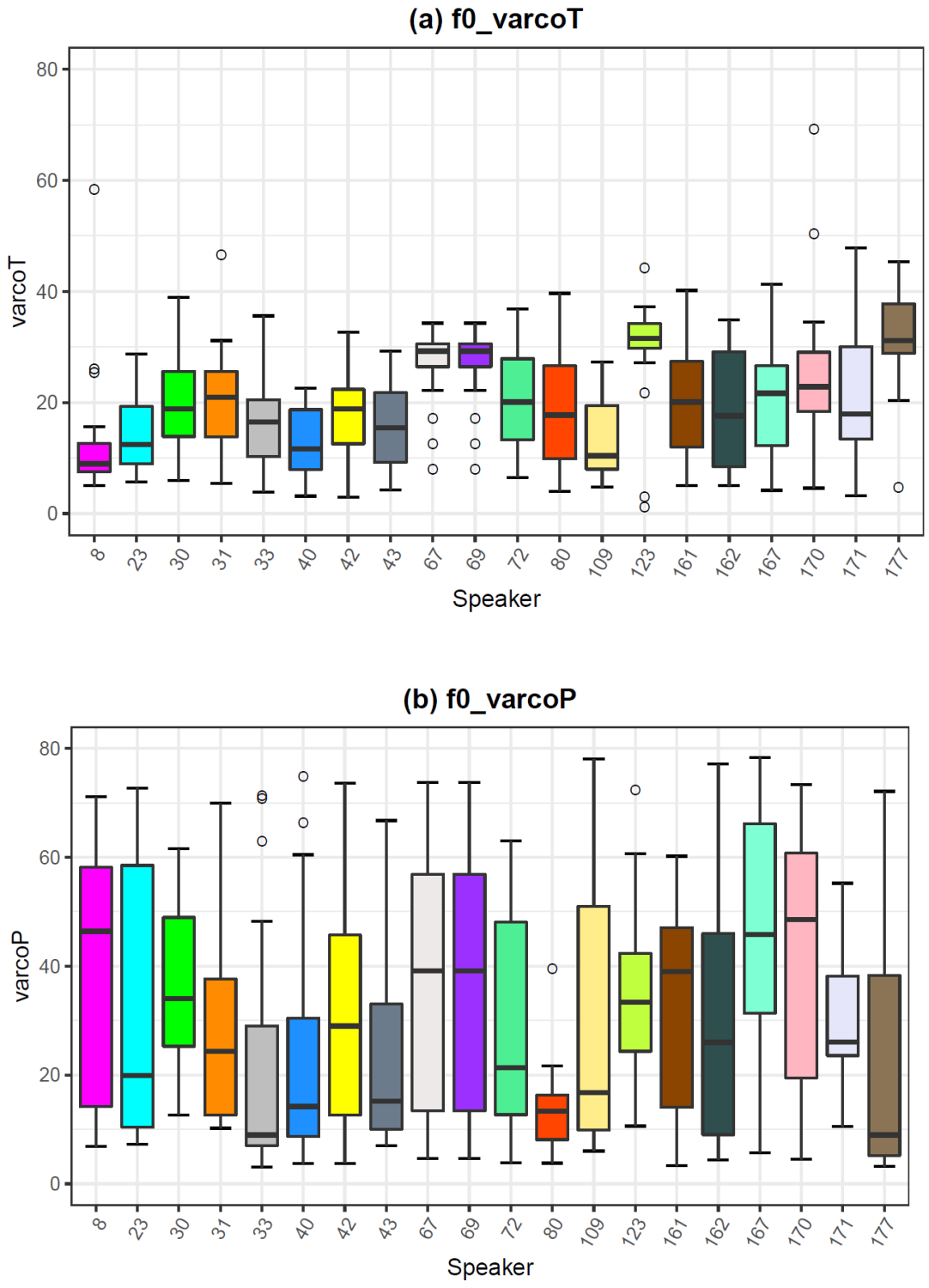


Figure 3.5. Boxplots of the 20 speakers f_0_varcoT (panel a) and f_0_varcoP (panel b) measures for each of their 18 utterances.

In comparing f_0_varcoT with f_0_varcoP, on the whole there is greater between-speaker variation in relation to speakers' trough f_0 measures than their peak f_0 measures (e.g., compare speaker 008 with speaker 177). There is also less within-speaker variation for the trough measures in comparison to the peak measures (e.g., f_0_varcoT boxes for speakers 008, 067, 069, 123). For the forensic analyst, greater between-speaker variation and little within-speaker variation is desirable meaning that these trough measures are much more favourable – at least in comparison to their peak counterparts. It is noted though that for some speakers (e.g., speakers 080, 167, 171) a good deal of within-speaker variation is also present for f_0 trough measurements and therefore this is something which would need to be carefully considered if taking such measures forward into the forensic domain.

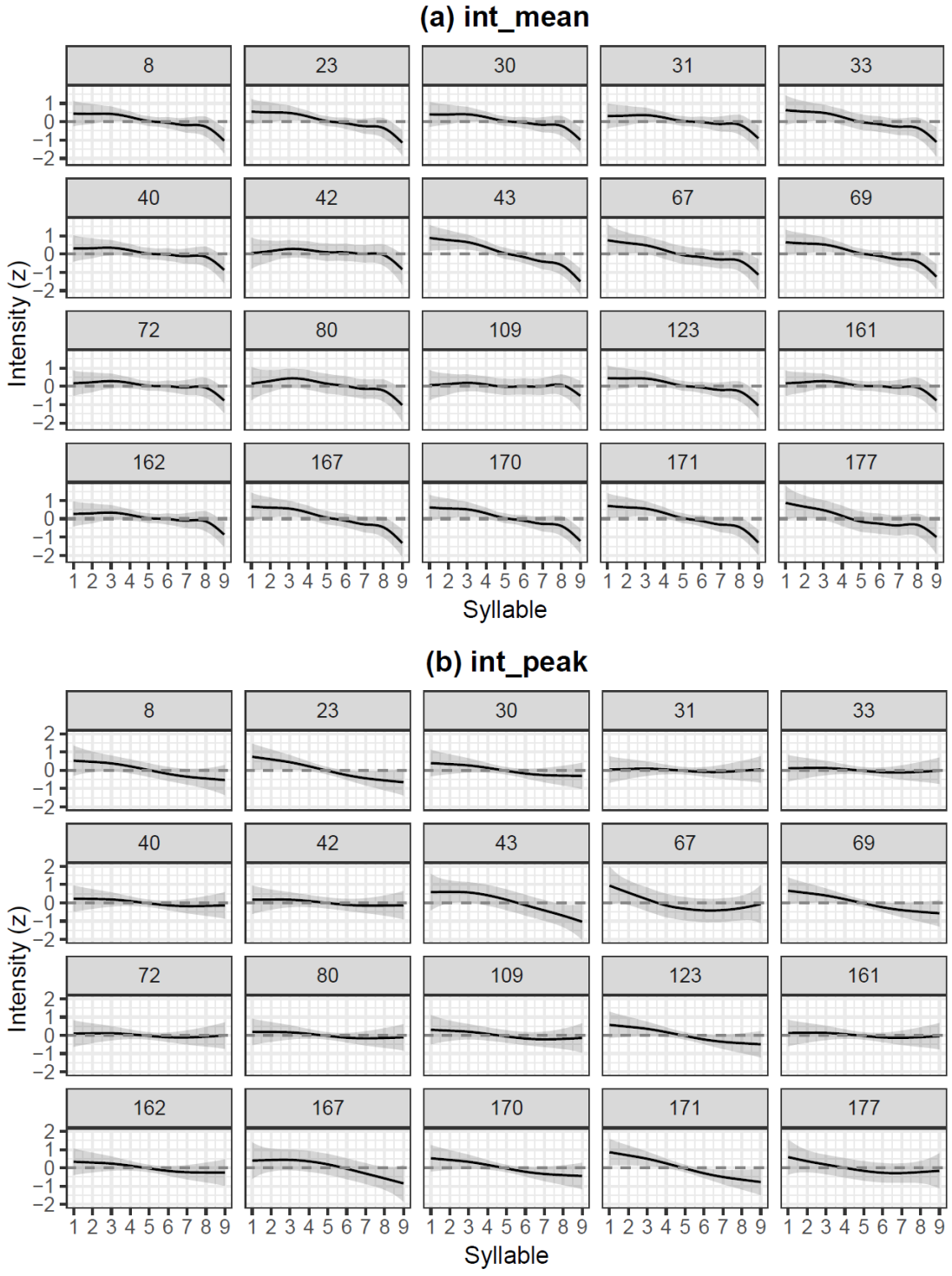
Table 3.2 displays the results for each of the GAMMs fitted in relation to the deviance explained by each model (%) and the interaction between the three rhythm parameters. See Section 3.2.6 for further details (e.g., model notation) relating to the GAMMs results presented below.

Table 3.2. The percentage of deviance explained by each of the rhythm measures (2nd column) along with the significance of the interactions between the parameters. Significant interactions are highlighted in bold, and the deviance explained by the interaction models is given in brackets.

		Interaction Measure		
Variable Measure	**Deviance explained**	**Intensity**	**f_0**	**Duration**
int_mean	37.4%		*p* = 0.0586 (37.8%)	***p* = 0.0001** (39.8%)
int_peak	29.3%		*p* = 0.3253 (30.5%)	***p* = 0.0028** (55.5%)
int_trough	40.1%		*p* = 0.1450 (40%)	***p* = 0.0139** (47.2%)
f_0_mean	36.6%	***p* = 0.0365** (38%)		*p* = 0.3581 (38.5%)
f_0_peak	47.6%	***p* = 0.0002** (48.8%)		*p* = 0.3155 (48.4%)
f_0_trough	63%	*p* = 0.2270 (62.5%)		*p* = 0.1845 (63.1%)
duration	39.1%			

It can be observed that of the models without interactions that it is the durational model which accounts for the most variation (30.1%), followed by mean intensity (23.2%) and trough intensity (20.5%). Figure 3.6 shows the by-speaker contours for each of the intensity measures (panels (a) – (c)) along with the duration contour (panel (d)) across the 9-syllable utterances. In offering visualisations of individual speakers'

intensity and duration contours, this will allow for it to be observed as to whether there are any speakers which exhibit any differentiating rhythmic behaviour for these parameters and whether there are any observable trends or correlations between the two parameters.



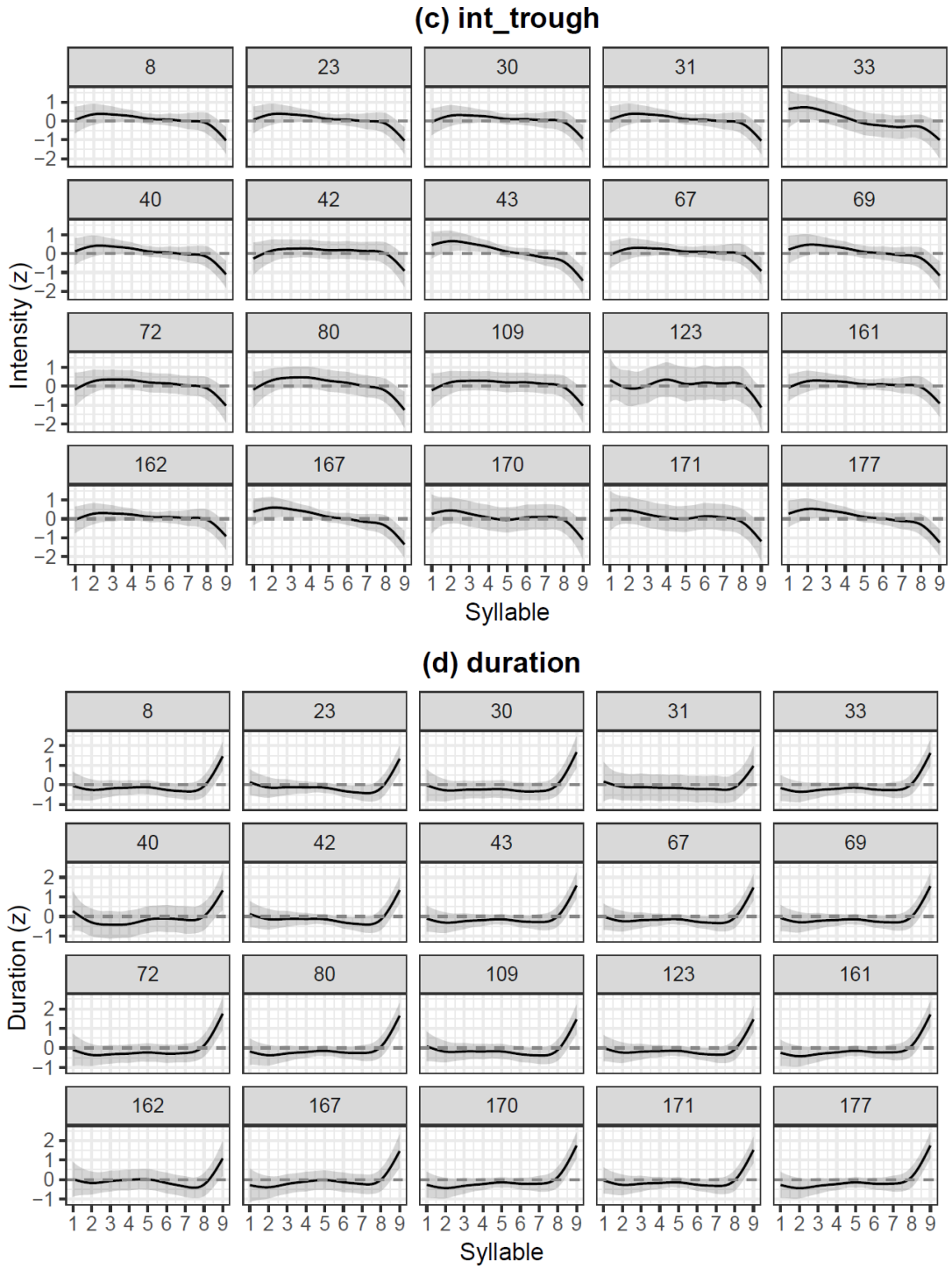


Figure 3.6. GAMM plots of by-speaker syllable-varying intensity contours (panels (a) – (c)) where higher z-scores correspond to greater intensity and the durational contour (panel (d)) where higher z-scores correspond to longer duration.

For duration, there is a marked upward rise of the contour from syllable eight to syllable nine for all speakers, with this trend seemingly supporting the well-

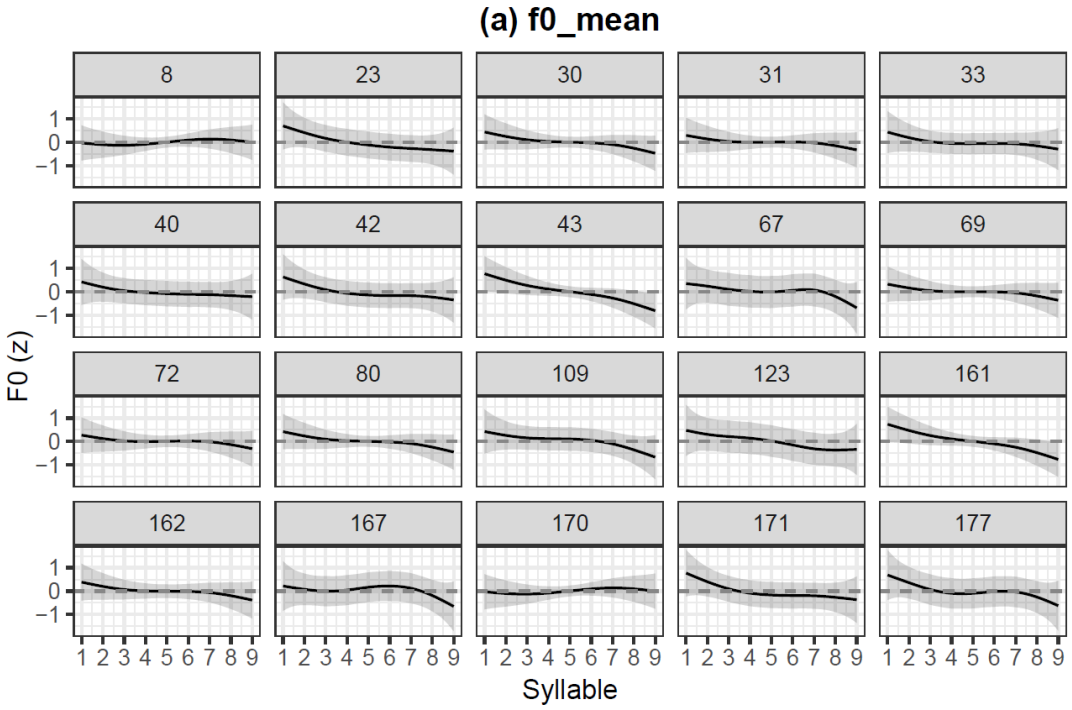
established phenomenon of phrase-final lengthening (e.g., Klatt, 1976; Shattuck-Hufnagel & Turk, 1998; Wightman et al., 1992). For all three intensity measures (i.e., panels (a) – (c)), it is observable that there is a general trend of a decrease in syllabic intensity for all speakers across the 9-syllable contours with this decrease being more marked for some speakers (e.g., speaker 043). Contra to the trajectory for duration, there is also a pattern (most noticeably for trough and mean intensity) of a distinct decrease in intensity between the penultimate and final syllable in these contours. This visual correlation between final-syllable lengthening and a final-syllable decrease in intensity may be the main contributing factor in there being significant interactions between the three intensity measures and duration. To investigate this interaction further in relation to the individual speakers, Table 3.3 displays the intensity and duration GAMMs fitted in relation to the significance of the smooth terms for each speaker (*p*-values), and the interaction between the three intensity measures and duration (*p*-values).

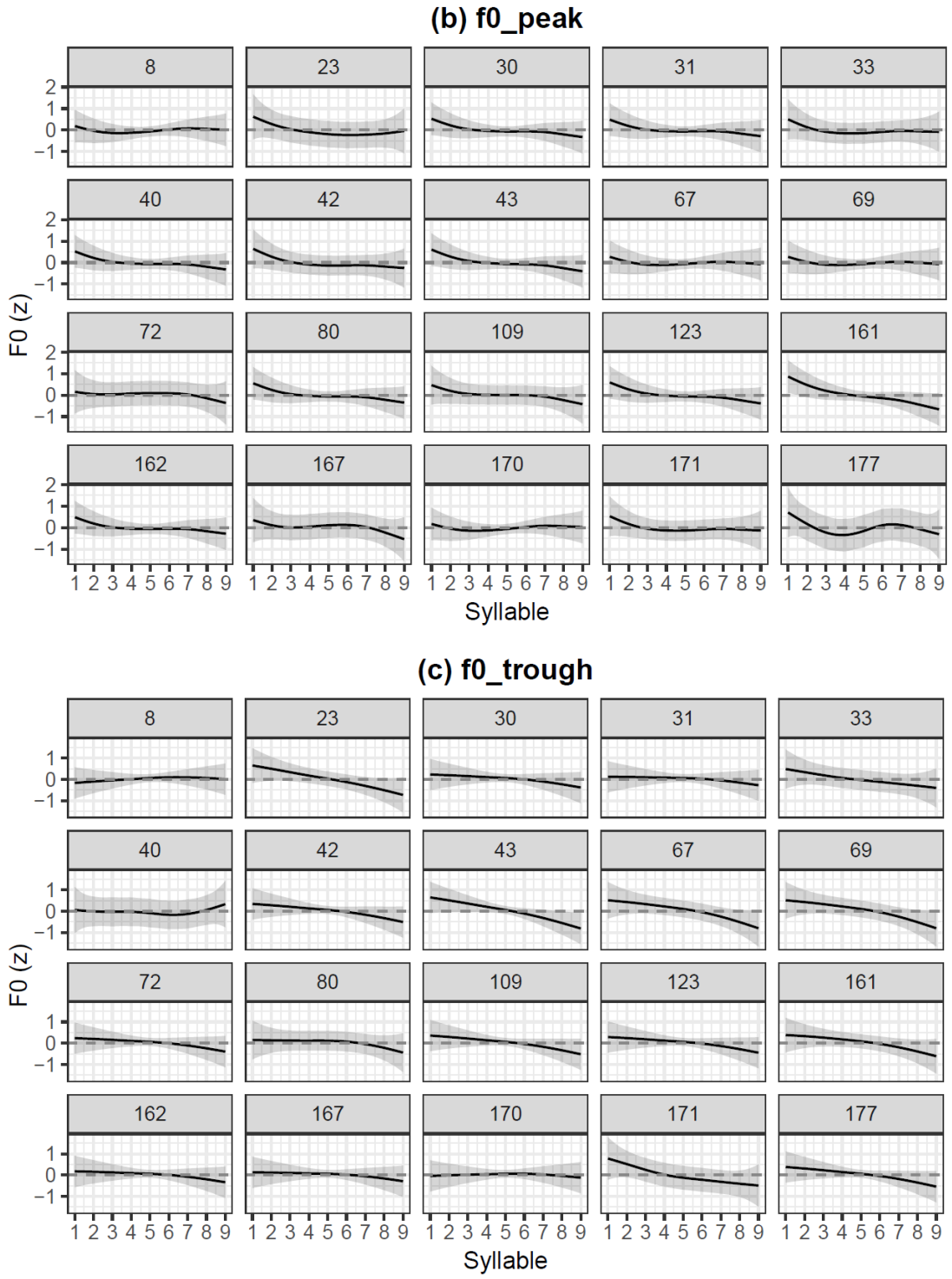
Table 3.3. *p*-values for the approximate significance of smooth terms for each speaker across each of the fitted intensity and duration GAMMs. Yellow highlighted cells indicate the five speakers who had significant smooth terms across four or more of the fitted GAMMs. Grey highlighted rows indicate the two speakers who yielded no significant smooth terms across any of the fitted GAMMs.

	Measure						
	Peak	**Trough**	**Mean**	**Duration**	**Peak & Duration**	**Trough & Duration**	**Mean & Duration**
Speaker ID							
008	0.2969	0.6529	0.2831	0.2334	**<0.0001**	0.2889	0.4952
023	**0.0312**	0.7056	0.0860	0.6078	**<0.0001**	0.9484	0.1226
030	0.7424	0.2910	0.4140	0.0751	**0.0082**	0.2278	0.6091
031	0.1241	0.7283	0.7607	0.0751	0.9986	0.9976	0.9991
033	0.2437	**0.0108**	0.1175	**0.0185**	0.2463	**0.0054**	0.1644
040	0.5877	0.9755	0.8789	**0.0359**	0.1406	0.8622	0.7352
042	0.5833	0.1016	0.3993	0.5612	0.3683	0.0810	0.2203
043	**0.0026**	**0.0380**	**0.0006**	**0.0394**	**<0.0001**	**0.0358**	**<0.0001**
067	**0.0046**	0.2523	**0.0452**	0.1497	**<0.0001**	0.2452	0.0417
069	0.0830	0.7268	**0.0233**	0.0626	**<0.0001**	0.5370	**0.0280**
072	0.2189	0.3410	0.6018	**0.0200**	0.2849	0.1396	0.4115
080	0.4696	0.1178	0.3590	**0.0133**	0.1113	0.0529	0.3959
109	0.7474	0.1861	0.2002	0.4055	0.2365	0.2550	0.0615
123	0.2004	0.0505	0.2546	0.1742	**0.0003**	0.2352	0.3546
161	0.2939	0.2292	0.5957	**0.0037**	0.2064	0.0723	0.3230
162	0.9983	0.2428	0.9992	0.09888	0.0966	0.3928	0.6495
167	**0.0167**	0.1160	**0.0256**	**0.0306**	**<0.0001**	0.1638	**0.0281**
170	0.3016	0.3063	0.0971	**0.0025**	**0.0001**	0.3202	0.1288
171	**0.0055**	0.1224	**0.0081**	0.0998	**<0.0001**	0.4952	**0.0074**
177	0.2861	0.3431	**0.0101**	**0.0022**	**0.0007**	0.4952	**0.0094**

Table 3.2 revealed that the strongest interaction was between peak intensity and duration, and this is evidenced in Table 3.3 in that 11 of the 20 speakers are indicative of this significance. Interestingly, the model which accounted for the least variation was the peak intensity model (29.3%), suggesting that for some speakers, especially speakers 008, 069 and 123 who exhibited no significance in either peak intensity or duration when considered separately, the interaction between duration and peak intensity is especially meaningful.

Figure 3.7 shows the by-speaker contours for each of the f_0 measures (panels (a) – (c)) along with the duration contour (panel (d)) across the 9-syllable utterances.




(b) f0_peak
8
23
30
31
33
40
42
43
67
69
72
80
109
123
161
162
167
170
171
177
F0 (z)
Syllable
(c) f0_trough
8
23
30
31
33
40
42
43
67
69
72
80
109
123
161
162
167
170
171
177
F0 (z)
Syllable

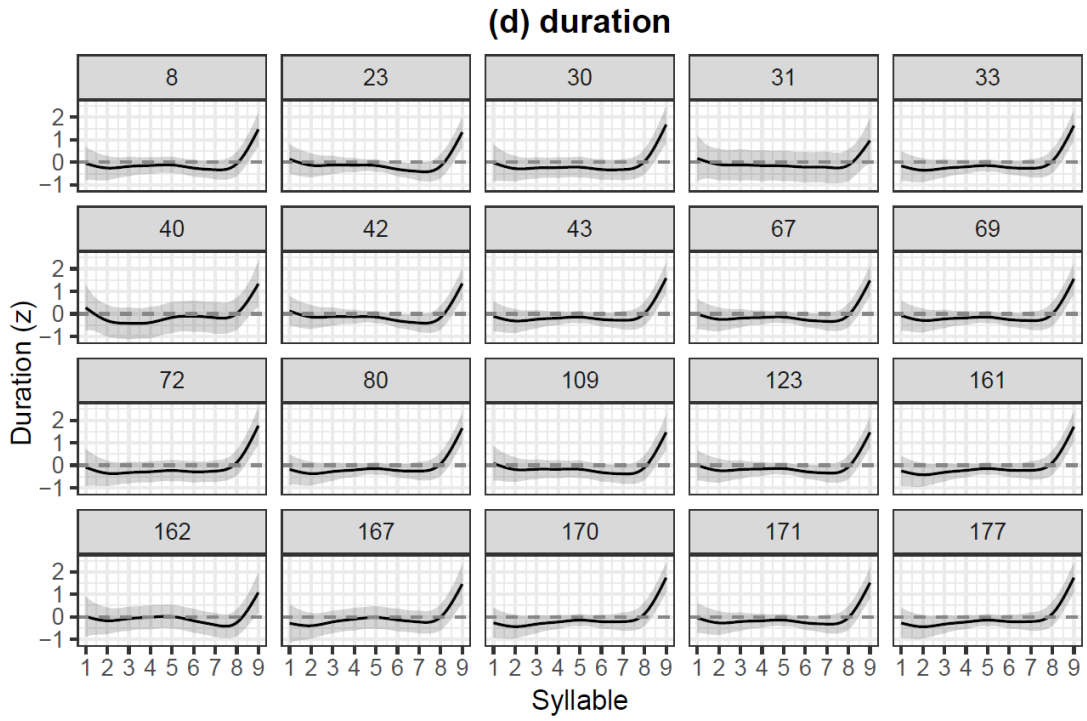


Figure 3.7. GAMM plots of by-speaker syllable-varying f_0 contours (panels (a) – (c)) where higher z-scores correspond to higher f_0 and the durational contour (panel (d)) where higher z-scores correspond to longer duration.

In examining the f_0 contours (panels (a) – (c)), it is observable that there is somewhat less variation in speakers' trajectories when compared to the intensity contours (n.b., smaller scale y-axis). This suggests that speakers exhibit less variation in their pitch across the utterances and more variation in terms of changes in loudness (generally decreasing over an utterance). Although slight, for a number of speakers there is also a general decrease in f_0 (i.e., slight downward sloping trajectories) over the utterances which could suggest a relationship between f_0 and intensity, in that as intensity decreases over an utterance, so does f_0, albeit to a lesser extent. Table 3.2 showed that the only significant interactions in this regard are between f_0 mean / f_0 peak and intensity, with there being no significant interactions between f_0 measures and duration across the utterances. Where for intensity (mean and trough) there is a marked fall in the trajectory between syllable 8 and syllable 9 for all speakers, this is not replicated for f_0, which, given the comparative lack of significant interactions for f_0, lends further support to the importance of the penultimate and final syllables in determining the significant interaction between intensity and duration.

3.3.2. Dynamic intensity measurements

Figure 3.8 shows the classification rates yielded from the linear discriminant analysis results for the dynamic intensity contours across the 9-syllable utterances.

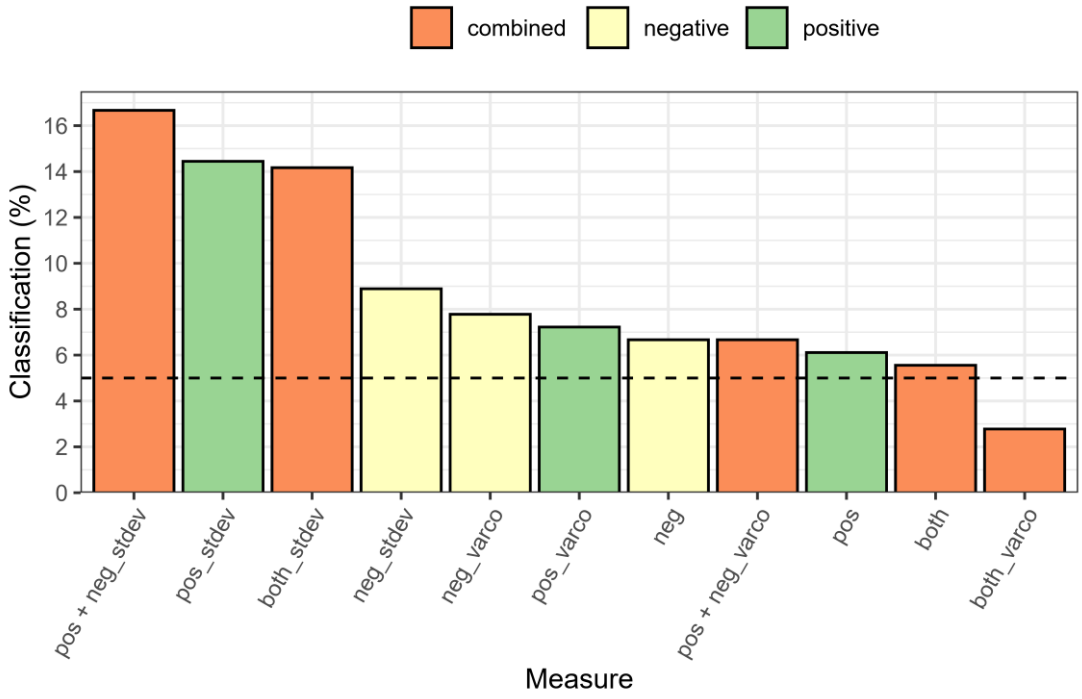


Figure 3.8. Discriminant analysis classification rates for each of the dynamic intensity measurements (ranked from highest to lowest). Chance level is 5% (indicated by dotted line) as there are 20 speakers.

Similar to the results for the static measurements, there is a general tendency for the application of single-value quantification metrics to perform better at distinguishing between speakers, specifically the normalised standard deviations across utterances. The best performing measure (pos + neg_stdev) is derived from combining the standard deviations of the positive dynamics (pos_stdev) with the standard deviations of the negative dynamics (neg_stdev). In comparing this measure to both_stdev (which is the third best performing), its potential to distinguish between speakers to a slightly greater degree lends support to the argument of it being beneficial to treat positive and negative dynamics as separate entities. This is further highlighted by those measures from the contour-approach (i.e., pos, neg and both), in that both the

positive dynamics (pos) and the negative dynamics (neg) perform better when treated separately than when integrated together (both). In addition, there is no clear pattern as to whether it is positive dynamics or negative dynamics which discriminate between speakers most efficiently. For the contour-approach negative dynamics (neg) perform slightly better than the positive dynamics (pos), whereas for the single-value variability-approach the results are mixed with pos_stdev performing better than neg_stdev but conversely neg_varco performing better than pos_varco (further discussion in Section 3.4.2).

3.4. Discussion

This chapter reported evidence that speakers show variation in intensity, f_0 and durational features and that such variation results in some speakers being more distinguishable than others. Overall, it manifested that measures of intensity generally perform better at distinguishing between speakers than measures of f_0 and duration, although it was shown that accounting for all three parameters together produced the most effective results. In consideration of these findings, it would appear that a multidimensional approach to measuring speech rhythm carries the most potential for studying individual speaker variation and for distinguishing between speakers. The following subsections provide discussion relating to the results reported above.

3.4.1. Static syllable measures

In comparing the contour-approach and the variability-approach, results from the LDA showed that the variability-approach (i.e., the single-value variation coefficient) was generally the most efficient at distinguishing between speakers, particularly with regards to the quantification of syllabic trough intensity and trough f_0, as well as syllabic duration. This is perhaps not surprising given that the classifications for the variability-approach are based upon 360 data values (18 utterances × 20 speakers), whereas the intensity-contour classifications are based on 3240 data values (18 utterances × 9 syllables × 20 speakers). This disparity in relation to the quantity of data being handled by the LDA is something which should be taken into consideration

when making such comparisons. Additionally, the homogeneity of the dataset should also be acknowledged in that the results obtained from both the contour-approach and the variability-approach are derived from the same raw data values, and these raw data values were obtained from speakers from the same specific accent group (i.e., Bradford English).

Of particular interest was the observation that the best performing measure across all the measures was the normalised variation coefficient int_varcoT (the variation coefficient attributed to trough (minimum) intensity values) and likewise the best performing measure from the contour-approach was that of the normalised trough-intensity contours. In their intensity-based study, He and Dellwo (2016) did not measure syllabic trough intensity variability, and just accounted for peak intensity (stdevP, varcoP, rPVIp, and nPVIp) and mean intensity (stdevM, varcoM, rPVIm, and nPVIm). They found that the normalised variation coefficients varcoP and varcoM performed slightly better than their raw counterparts stdevP and stdevM, with this result lending to the inclusion and selection of this specific quantification metric in the present study. Where they found that peak measures (varcoP) performed better than the mean measures (varcoM), the opposite effect was found in the present work, although both were surpassed by intensity trough measures. One possible explanation for the difference in results obtained between the present experiment and that of He and Dellwo (2016) could be attributed to the nature of the data in that where the former used spontaneous, content-mismatched speech, the latter used content-controlled, read speech. Within content-controlled, read speech peak intensity measurements might be expected to produce better results as speakers will likely be more controlled and regular in their articulations. For spontaneous speech conditions, syllabic emphasis is likely to be less regulated and therefore intensity can be expected to show greater within-speaker variation resulting in poorer speaker discriminatory potential.

The contour-approach was afforded further inspection through the application of GAMMs in which a significant interaction between intensity measures (mean, peak and trough) and duration was established. Visualisations of the intensity, f_0 and durational contours revealed some perhaps predictable patterns such as speakers exhibiting a general decrease in intensity across their utterances, with this being more

pronounced around the penultimate and final syllable, as well as phrase-final lengthening (again, markedly around the final two syllables). Both of these patterns have been found to be indicators within spontaneous speech of a given speaker finishing their turn and thus allowing another speaker to take the floor. Given that the decrease in syllabic intensity and the increase syllabic duration seem to cooccur in the present data, it could be that the significance of the interaction between these two parameters holds more weight in a phrase-final context. As for f_0, these measures (mean, peak and trough) yielded no significant interactions between duration, with the plotted contours showing this parameter to be much less variable, both within and between speakers. Despite the common assumption that f_0 plays a major role in signalling syllabic prominence, previous studies which have focussed on spontaneous British English have found that the most significant cue for emphasis is intensity, either followed by, or accompanied with, duration (e.g., Herment, 2012; Kochanski et al., 2005). Given the finding that measurements of syllabic intensity, along with its interaction with duration, carry the most speaker-specific information in the present study, it could be the case that it is the stressed/prominent syllables within speakers' utterances that are most useful for distinguishing between individuals, although further investigation would be needed to substantiate this notion. GAMMs analysis revealed that the best-performing models (i.e., which explained the most variation) were the trough intensity + duration model and the peak intensity + duration model, and therefore focussing on the relationship between these measures (rather than syllabic mean intensity on f_0 measures) may be the most fruitful in determining idiosyncratic rhythmic behaviour.

Overall, the results obtained from the static syllabic measurements showed that using these methods and parameters to capture spontaneous speech rhythm patterns yields little with regards to discriminatory power. The reasons that these speech rhythm measurements (and metrics) do not transfer over well to the spontaneous speech condition are perhaps obvious. They effectively involve making syllable-to-syllable comparisons across utterances (i.e., the first syllable's relative duration measurement of utterance X from speaker 1 is compared against the first syllable's relative duration of utterance X from speaker 2). While this is a good setup for read speech, it does not translate so well to the spontaneous speech condition. The approach involves making

comparisons across syllables that are different with respect to their phonetic content, level of stress, whole-utterance factors, etc.; all of which will contribute to the variables we are aiming to use to capture speech rhythm. In essence, these rhythm measures are too sensitive to the variation that spontaneous, content-mismatched speech contains.

3.4.2. Dynamic intensity measures

For the dynamic intensity measurements, it was again the application of single-value metrics which resulted in the best speaker-distinguishing performances, with standard deviation quantifications proving to be the most effective. Although, similar to the results obtained for the static measures, the variation approach results are based upon fewer data values than the results for the contour-approach. For the contour-approach, negative dynamics (the speed of decreases in intensity from syllable peaks to between-peak troughs) performed marginally better than positive dynamics (the speed of increases from these troughs to syllable peak intensity). This finding seemingly reinforces the results of He and Dellwo (2017) who also found negative intensity dynamics to be more speaker-specific than their positive counterparts. In offering an explanation for this finding, they highlight the articulatory rationale, suggesting that the positive and negative dynamics studied are related, respectively, to the opening and closing gestures of the mouth, and that these gestures serve different purposes. They draw upon motor plant theory and the notion of controllable and intrinsic articulatory properties (Perrier, 2012) to argue that the controllable properties play a greater role in opening gestures (in order to reach articulatory targets (e.g., Birkholz et al., 2011; Ghez & Krakauer, 2000)), whereas the intrinsic properties play a greater role in closing gestures, in which control of the articulators is reduced resulting in movements conditioned to a greater extent by idiosyncratic biophysical properties. It could be then that, for the present study, this particular result of negative dynamics performing better than positive dynamics (within the contour-approach) is potentially also a result of such idiosyncratic articulation.

However, He and Dellwo (2017) found that negative dynamics explained around 70% of between-speaker variation (thus, positive dynamics 30%), where the present study

has the two sets of dynamics much more closely approximated. Where, for the single-value metrics, negative measures also outperform positive measures in terms of the variation coefficients, the opposite effect is found in terms of the standard deviations, which, as already mentioned, was the best-performing dynamic metric overall. He and Dellwo's (2017) result which highlights the potential usefulness of negative dynamics over positive dynamics in capturing speaker-specific information is one which is reinforced in a later study by Zhang et al. (2021) who report similar figures for their study on Thai speech. Both of these studies, however, used controlled, read speech and therefore it may well be the case that transferring this method of analysing intensity dynamics to spontaneous speech results in this apparent difference between positive and negative dynamics becoming less obvious. Indeed, in transferring this approach to measuring intensity dynamics over to spontaneous speech, Machado (2021) found this difference between positive and negative dynamics in relation to the between-speaker variation they explain to be marginal, with negative dynamics explaining 52% of variation and positive dynamics 48% h – a finding much more in tune with the present experiment. One possible explanation for this could be the greater degree of gestural overlap between the start and end of syllables in spontaneous speech in comparison to read speech as a result of potentially less regulated and less uniform articulatory movements (e.g., De Nil & Abbs, 1991; Illa & Ghosh, 2020) Similarly, Machado also employed LDA to test the discriminatory potential of the measures taken, reporting low speaker classification rates, with the best results being negative measures of means (CR = 4.8%; chance level = 1.9%). This figure for the best-performing measure is one which reinforced by the results reported in this chapter given that the best-performing measure here was positive measures of standard deviations (CR = 14.4%; chance level = 5%), thus the best results from both studies were only around two and a half times above their respective chance levels. Although some of the dynamic intensity measures in the present study performed better than some of the static measures, the improvements were, for the most part, only minor.

Overall, although more effective than measures of f_0 and duration (as a standalone entity), measurements of intensity, whether these be static or dynamic, are shown to have relatively poor discriminatory power when analysing spontaneous speech

utterances. It could well be the case that adopting the methodology of He and Dellwo (2017) for the extraction and measuring of intensity dynamics is the reasoning behind the poor results obtained from the LDA given that their methods were designed for the use on content-controlled, read speech. Therefore, it is probable that this methodology is likely too sensitive to the variation that content-mismatched, spontaneous speech contains.

Does this mean that attempting to utilise measurements of intensity in the assessment of speech rhythm for forensic purposes should be abandoned? Given that intensity has been shown to be the best-performing measure when compared to f_0 and duration within the present chapter (as well as in previous research (e.g., He and Dellwo (2016)), it would seem that discounting intensity altogether would be wasteful. However, it is well-established that analysing intensity in spontaneous speech is extremely delicate given how susceptible the measure is to noise (e.g., the introduction of background noise will likely initiate a subconscious increase in vocal effort from a speaker), and that a slight turn of the head or a hand passed over the mouth will result in a drop in the intensity measured. Similarly, discrepancies in the distance between speaker and microphone and even in the type of microphone used to record the spoken data could have marked effects on measurements. Nevertheless, intensity has been shown to play an important role in marking prominence within spontaneous speech, and, considering the potential importance of prominence in assessing speech rhythm, it seems that intensity should be accounted for. Rather than adopting a stance of disregarding intensity outright as being a measure which is 'too sensitive' for consideration within the forensic context (i.e. when forensically-relevant speech data is under investigation), the experiments carried out in the present chapter have looked to provide an initial exploration into the tenability of analysing intensity measures as a means of assessing speech rhythm patterns. In doing so, these experiments have also taken the natural next step forward from previous studies (which had made use of controlled, usually read, speech data) and have applied these measures to more forensically realistic data.

Despite the results obtained from the present chapter's experiments, it remains that, at present, there is still comparatively less known about the potential forensic application

of intensity measures as opposed to other parameters such as duration and f_0. Research by Kolly and Dellwo (2014) does highlight the potential forensic relevance of intensity measures in observing that intensity patterns may not easily be manipulated by speakers (e.g., as a disguise strategy) due to lack of possible auditory feedback which does provide some support for pursuing the tenability of measuring intensity for forensic purposes. Perhaps, then, if it can be determined that the recording conditions for a known sample and a questioned sample within a given FVC case have been relatively stable, then analysing speakers' intensity patterns as a means of capturing idiosyncratic rhythmic behaviour could be of use to the forensic analyst. Alternatively, measuring intensity over much shorter durations, such as individual speech units (see Chapter 4), where there is less potential for interference from those aforementioned problematic factors, could facilitate more reliable and robust measurements being obtained, and subsequently aid in determining the usefulness of intensity within forensic casework.

3.5. Chapter summary

This chapter provided an examination of spontaneous speech rhythm in terms of measurements of intensity, f_0 and duration. Static syllabic measurements revealed that intensity was the parameter in which most between-speaker variation was evidenced and by which speakers could be differentiated from one another most effectively. However, the discriminatory power of these static measurements for all of the parameters was relatively weak overall with classification rates only marginally surpassing chance level. Dynamic measurements of intensity over the same spontaneous speech utterances yielded marginally improved results in some cases, however, as a collective, these results were also relatively weak. It therefore seems that pursuing measuring speech rhythm parameters for the purpose of speaker discrimination when spontaneous, content-mismatched speech data is under investigation is, for the most part, an unproductive endeavour. What, however, would be the implications within the forensic casework scenario if more promising results were yielded, and the features analysed in the present experiment were able to discriminate between speakers with close to perfect performance? If these parameters

were shown to be useful for speaker discrimination tasks in such scenarios, it could be expected that an x-vector system (i.e., an automatic speaker recognition system) would be able to make near perfect speaker classification assessments. With this being the case, one might ask the question as to why such systems are not utilised within forensic casework. One reason for this is that the use of this technology is, at present, not admissible in court within the United Kingdom. More importantly, however, is the reason pertaining to the explainability of using automatic systems within the criminal justice system. That is, there is concern surrounding how the complexities of automatic speaker recognition systems (e.g., complexities relating to the inner workings of these systems) could be explained within the evidential setting such as to a judge or jury. This is in contrast to the acoustic analytical approach (i.e., the approach taken for the experiments in the present chapter) which assesses features that can be directly linked to speech production and are more easily explained to the layperson (such as judge and/or jury). Given that the aim of the present thesis is to develop the way speech rhythm is analysed within FVC casework, it is the acoustic analytical approach which is explored in the production experiments of the present thesis in order to assess the tenability of using these measures for speaker discriminatory purposes.

In light of the results obtained from the experiments in this chapter, the question might be asked as to whether the approach to measuring speech rhythm in these experiments could be improved through employing different methods.

The experiments carried out in the present chapter have taken the approach of attempting to deconstruct speech rhythm by analysing intensity, f_0 and duration individually to determine whether there was a standout performer in relation to speaker discriminatory potential. It remains the case that more research is needed in order to better understand the complex interactions and interrelations between these parameters as it is likely these processes which give rise to the conceptualisation of speech rhythm. Advances in research along these lines could potentially facilitate multidimensional approaches to measuring speech rhythm in which all associated parameters and their interrelations are accounted for in a unified model. With this being said, it may very well be the case that the complexities which would inevitably

come along with such a model may be such that its application for forensic purposes (i.e., speaker discrimination tasks) would be limited at best. Future research which seeks to investigate this area further will therefore inform on whether deconstructing speech rhythm into individual components is the most effective method for assessing idiosyncratic speech rhythm patterns.

CHAPTER 4

Speech Rhythm in Frequently Occurring Speech Units

4.1. Introduction

This chapter details the findings of the speaker discriminatory potential in relation to the rhythmic characteristics of four types of, so-called, “frequently occurring speech units”. Namely, the four speech units analysed in this chapter are the filled pauses *er* and *erm*, and the common monosyllabic responses *yeah* and *no* (n.b., within the present thesis the terms “frequently occurring speech units” and “speech units” are used interchangeably and refer specifically to the four monosyllabic units being analysed unless otherwise stated).

Previous studies have indicated that filled pauses such as "er" and "erm," along with monosyllabic responses like "yeah" and "no," are relatively common elements in spontaneous speech. Additionally, a limited number of investigations have explored the functions of these speech units, their typical positions within utterances, and the other features with which they are likely to co-occur. For instance, Gósy (2023) examined the frequency and duration of filled pauses in relation to words and silent pauses, revealing a distinct pattern in their occurrence across various positions. While this research did not focus on speaker-specific patterns, the observation that these units display specific positioning patterns within spontaneous speech supports the idea that they could serve as valuable reference points for measuring speech rhythm patterns. Similarly, Braun et al. (2023) discovered that verbal fillers, particularly "yeah" (or

"ja"), demonstrate speaker-specific patterns concerning their discourse position, indicating that speakers tend to avoid pausing before and after using these fillers or pause in both instances. This patterning further suggests that such units may be beneficial as 'anchors' or 'control units' for identifying individual speech rhythm patterns. Consequently, these four speech units were analysed for their rhythmic characteristics, allowing for comparisons with the spontaneous speech utterances discussed in the preceding chapter (see Chapter 2, Section 2.6 for further detail).

It may be argued that analysing single speech units such as these is not germane to a study of spontaneous speech rhythm given that rhythm is usually thought of as being a sequence of a number of different components over a stretch of speech (i.e., of greater length than just a single speech unit). However, the current chapter considers the rhythm measurements obtained from these speech units (that is, measurements of intensity, f_0 and duration) alongside the corresponding measurements of speakers' spontaneous speech rhythm patterns (i.e., the measurements analysed in the previous chapter) through normalising the frequently occurring speech unit data against the corresponding data from the spontaneous utterances (see Section 4.2.3). Therefore, the analysis of these frequently occurring speech units within the present chapter serves to provide an initial exploration as to their acoustic composition and potential speaker-specificity with respect to their rhythmic characteristics. Given that previous research which has looked at the speaker discriminatory potential of the filled pauses *er* and *erm* in relation to their F_1-F_3 measurements yielded promising results (see Chapter 2, Section 2.6.1), measurements of these parameters are also taken for these two speech units. This will allow for comparisons to be made between the formant-based measurements and the rhythm-based measurements with respect to which are able to distinguish between speakers most effectively.

In consideration of the relatively poor results obtained from the experiments in the previous chapter, it is hypothesised that analysing the rhythmic characteristics of the four speech units *er, erm, yeah* and *no* in terms of their intensity, f_0 and durational properties will yield more promising results in relation to speaker discriminatory value. This prediction is also informed in part by the small body of forensic research

that has been carried out with respect to some of these speech units (e.g., Hughes et al., 2016).

Of the three parameters under investigation, arriving at a confident prediction as to which should be expected to show the most speaker-specificity is difficult in light of the lack of previous research focussed on these parameters for these specific speech units. Furthermore, an argument for each of the parameters as being the most useful speaker discriminator can be posed as it is most likely that individual speakers may mark these speech units in different prosodic ways from one another (e.g., due to the functionality of the specific units or due to speaker idiosyncratic behaviour). Given that intensity was shown to be (marginally) the best-performing parameter overall in the previous chapter, a tentative prediction is offered in that this parameter could exhibit the most discriminatory potentially for the experiments in the present chapter.

As there has been much more research which has focussed on the filled pauses *er* and *erm* in comparison to the monosyllabic responses *yeah* and *no*, with such research showing that these filled pauses can be used in idiosyncratic ways by speakers, it is hypothesised here that it is more likely that it is these units which may show greater speaker discriminatory potential than *yeah* and *no*. With respect to the different types of measurements obtained from the four speech units (See Section 4.2.3 below), it is further predicted that dynamic measurements will yield the most favourable results in relation to speaker discriminatory value as opposed to the static measurements obtained. This prediction is also partly informed by the results reported in Hughes et al. (2016) who found dynamic measurements of F_1-F_3 outperformed (for the most part) static measurements.

This chapter is composed in the following way. Section 4.2 outlines the methodological procedures followed regarding the materials and speakers used, the data extraction and editing procedures, the measurements taken, the normalisation procedure and the statistical analyses conducted. Following this, in Section 4.3, the results are then presented. Firstly, a brief overview of the linear discriminant analysis results, the statistical method used to assess the speaker discriminatory potential of the speech units, is presented. This will allow for the speech units which show the most forensic potential to be attributed a more in-depth analysis. Following this, the results

from the filled pauses *er* and *erm* are provided before the results for the monosyllabic responses *yeah* and *no* are presented. The chapter is concluded with an overall discussion and summary of the results from both the filled pauses and monosyllabic responses.

4.2. Methodology

The following subsections detail the methodological procedures followed in relation to the preparation of the data and the how the acoustic measurements were obtained for the frequently occurring speech units. Firstly, in Section 4.2.1, the data on which this chapter is based is introduced. Section 4.2.2 provides details how the data were prepared and how each of the frequently occurring speech units were segmented. For each of the frequently occurring speech units, rhythmic patterning was accounted for by taking measurements of intensity, f_0 and duration. Section 4.2.3 defines each acoustic parameter in turn and explains how each parameter was measured. In Section 4.2.4, explanation is provided with regards to the normalisation procedures applied, before Section 4.2.5 details the statistical methods used to analyse the data.

4.2.1. Data

The data analysed in the present chapter were obtained from the same group of 20 young male speakers of Bradford English as the previous chapter and from the same mock police interviews (see Chapter 3, Section 3.1.1). All occurrences of *er*, *erm*, *yeah*, and *no* were identified within Praat for all 20 speakers and marked on separate interval tiers. The number of occurrences for each speech unit were recorded for each speaker. As there were substantial differences between some speakers (e.g., some speakers had < 5 occurrences of *erm*, whilst others had > 50), it was decided that, in order to maintain a balanced data set, those speakers who did not produce 'enough' tokens of a given speech unit would not be included in the analysis for certain speech units. For *er* and *erm*, 40 tokens of each speech unit were obtained across 12 speakers meaning that there was a total of 480 tokens for each type of filled pause (although there was some crossover, the 12 speakers were not all the same for *er* and *erm*). For

yeah and *no*, 35 tokens of each speech unit were obtained across 14 speakers meaning that there was a total of 490 tokens for each type of monosyllabic response (again, although there was some crossover, the 14 speakers were not all the same for *yeah* and *no*). The numbers reported for each speech unit above were the totals following the removal of tokens which had any erroneous f_0 or F_1-F_3 measurement values (see Section 4.2.3.2 below).

4.2.2. Data preparation

Each of the speech units analysed were manually marked on separate interval tiers within Praat. For *er* and *erm*, boundaries were placed at the onset and offset of periodicity of the vocalic segments, as well as the nasal segment for *erm*. For *yeah* and *no*, boundaries were placed at the onset and offset of periodicity of the entire speech unit. To delimit the onset and offset of periodicity, acoustic cues were drawn from both the waveform and the spectrogram. For example, in order to segment the vocalic from the nasal segment in *erm*, the vowel offset was defined in the spectrogram by a decrease in F_1 and F_2 frequencies and an overall decrease in amplitude (Johnson, 2012). If boundaries could not be confidently delimited, these speech units were not used in the analysis (n.b., the homogenous group of 20 speakers under investigation all speak with a non-rhotic variety of Bradford English meaning the segmentation of *erm* into its vocalic and nasal portions was unproblematic). Examples of the segmented speech units are shown in Figure 4.1 – Figure 4.4.

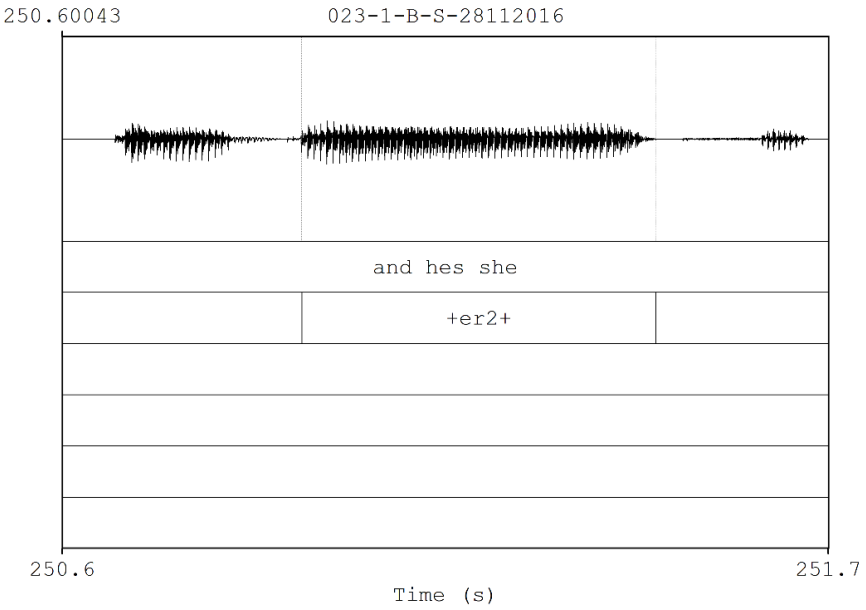


Figure 4.1. Example segmented TextGrid of the speech unit *er* from speaker WY023. Tier 1 contains the orthographic transcription (where 'hes' = hesitation (i.e., filled pauses) and Tier 2 contains the segmentation of *er*.

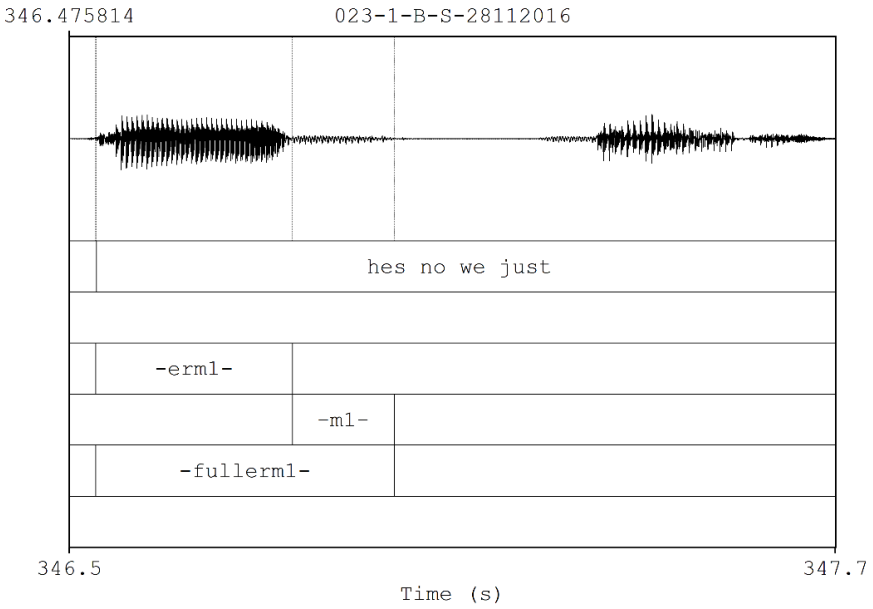


Figure 4.2. Example segmented TextGrid of the speech unit *erm* from speaker WY023. Tier 1 contains the orthographic transcription (where 'hes' = hesitation (i.e., filled pauses), Tier 3 contains the segmentation of the vocalic portion of *erm*, Tier 4 contains the segmentation of the nasal portion of *erm*, and Tier 5 contains the segmentation of *erm* as a whole.

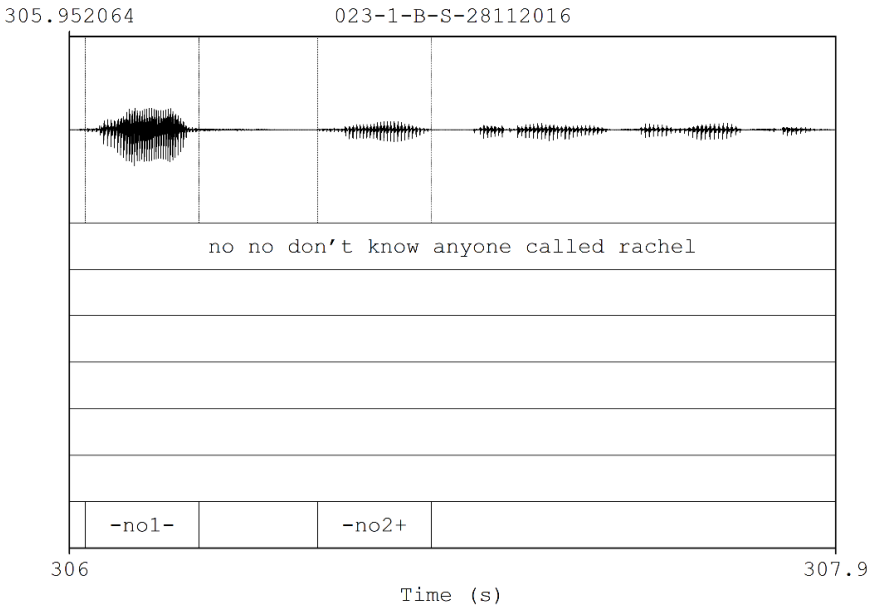


Figure 4.3. Example segmented TextGrid of the speech unit *no* from speaker WY023. Tier 1 contains the orthographic transcription and Tier 7 contains the segmentation of *no* (two occurrences).

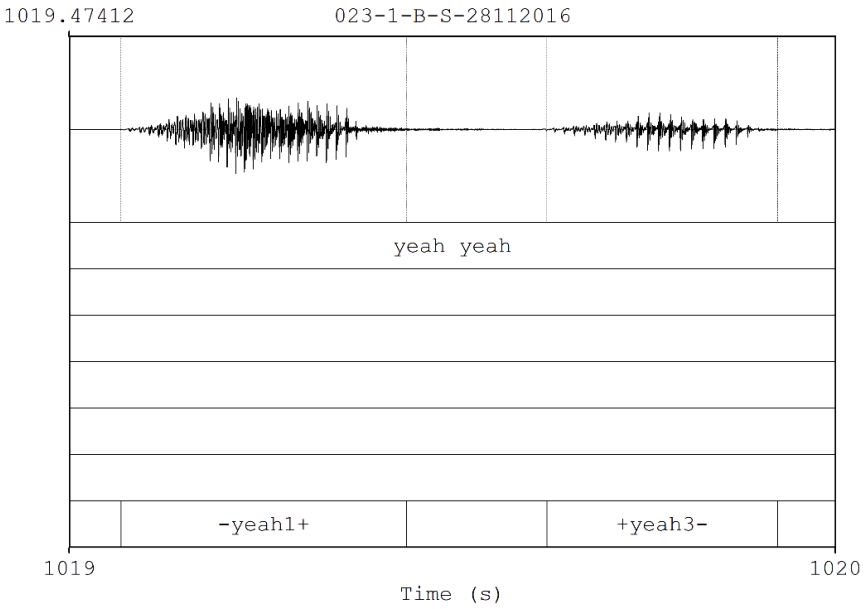


Figure 4.4. Example segmented TextGrid of the speech unit *yeah* from speaker WY023. Tier 1 contains the orthographic transcription and Tier 7 contains the segmentation of *yeah* (two occurrences).

4.2.3. Acoustic parameters

Rhythmic patterning across the frequently occurring speech units was accounted for by taking static measurements of intensity and f_0 as well as measuring duration. Dynamic measurements of intensity and f_0 were also calculated for each speech unit. As mentioned in the opening section of the chapter, F_1-F_3 measurements for the filled pauses *er* and *erm* were also taken in relation to the static midpoint (+50% interval) measurement and dynamic measurements of the formant trajectories. The following subsections detail how each of these parameters was measured.

4.2.3.1. Intensity

Intensity measurements were taken within Praat through the use of a script which was written by the present author. Table 4.1 below details the static intensity measurements obtained through the Praat script.

Table 4.1. Static intensity measurements obtained through the Praat script.

MEASUREMENT	DESCRIPTION
Mean intensity	Mean (in dB) of the intensity values of the frames within the specified time domain (averaging method = "dB").
Maximum (peak) intensity	Maximum value within the specified time domain, expressed in dB (interpolation method = "cubic").
Minimum (trough) intensity	Minimum value within the specified time domain, expressed in dB (interpolation method = "cubic").
Intensity at midpoint (+50% interval)	Value at the midpoint within the specified time domain, expressed in dB (interpolation method = "cubic").

As well as taking static intensity measurements, dynamic measurements of intensity were also calculated for each frequently occurring speech unit. Intensity measurements were extracted from each speech unit at +10% intervals across their trajectories. This meant that for each speech unit the dynamic intensity contour was made up of 9 intensity measurements (interpolation method = "cubic"). See Boersma and Weenink (2020) for the intricacies pertaining to the algorithm Praat uses to

calculate intensity values as well as the details for the averaging method and interpolation method used here.

Figure 4.5 provides an illustration as to where the static and dynamic intensity (and f_0, see Section 4.2.3.2 below) measurements were taken.

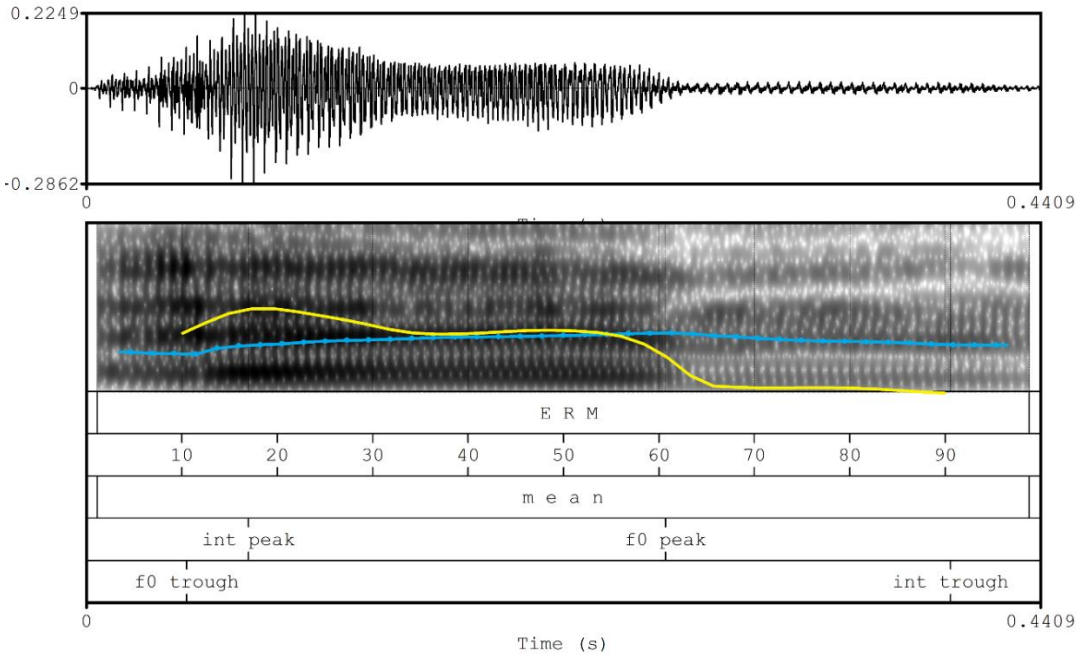


Figure 4.5. Example TextGrid of the speech unit *erm* uttered by speaker WY171. Tier 2 illustrates where dynamic measurements and the midpoint (+50% interval) were taken from across the intensity contour (yellow line) and f_0 contour (blue line). Tier 3, Tier 4, and Tier 5 indicate, respectively, where mean, peak and trough measurements were obtained.

4.2.3.2. f_0 and F_1-F_3

Measurements of f_0 were obtained using VoiceSauce (Shue, 2010) with all measurements being obtained using the same algorithm (i.e., 'STRAIGHT' (Kawahara et al., 1998)) and with the same settings applied as detailed in Chapter 3, Section 3.2.2.2. F_1-F_3 measurements were taken within Praat through the use of a script written by the present author which used the 'get formant' function of the software to obtain the respective formant measurements. The following formant settings were applied:

(1) Algorithm: Burg method (see Childers (1978, pp. 252-255 for full elaboration of the Burg algorithm)

(2) Time step (s): 0.00625 (Praat standard setting)

(3) Maximum number of formants: 5.0 (Praat standard setting)

(4) Formant ceiling (Hz): 5000.0 (Praat recommended setting for adult male speakers)

(5) Window length (seconds): 0.025 (Praat standard setting)

(6) Pre-emphasis from (Hz): 50.0 (Praat standard setting)

Table 4.2 below details the static f_0 and F_1-F_3 measurements calculated.

Table 4.2. Static f_0 and F_1-F_3 measurements calculated.

MEASUREMENT	**DESCRIPTION**
Mean f_0	Mean (in Hz) of the f_0 values of the frames within the specified time domain.
Maximum (peak) f_0	Maximum value within the specified time domain, expressed in Hz.
Minimum (trough) f_0	Minimum value within the specified time domain, expressed in Hz.
f_0 and F_1-F_3 at midpoint (+50% interval)	Value at the midpoint within the specified time domain, expressed in Hz.

As well as taking static f_0 and F_1-F_3 measurements, dynamic measurements were also calculated for each speech unit (F_1-F_3 measurements for *er* and *erm* only). f_0 and F_1-F_3 measurements were extracted from the relevant speech units at +10% intervals across their trajectories. This meant that for a given speech unit the dynamic contour was made up of nine f_0 and F_1-F_3 measurements (Figure 4.5 (above) exemplifies where both static and dynamic measurements were taken).

Occasionally, the automatic extraction of f_0 produced erroneous values due to factors such as creak and voicelessness resulting in tracking errors. Similarly, the automatic extraction of F_1-F_3 measurements would at times result in inaccurate values being returned (e.g., F_1 being measured as F_2). In order to remove such errors, the raw data

were inspected and any tokens which had unrealistic or missing values in relation to the static measurements were manually removed. In relation to the dynamic measurements, in order to preserve as many tokens as possible for analysis (rather than the more reductive approach of removing the entire token from analysis), unrealistic values, missing values or values with improbable shifts from one +10% step to the next were manually replaced with the mean of the two neighbouring values. Where missing values occurred at the +10% or +90% steps, or where there were multiple consecutive missing values, the entire token was removed. This process removed a total of 28 tokens from across the 4 different frequently occurring speech units meaning that for *er* and *erm* 40 tokens of each speech unit were eligible for analysis (across 12 speakers) and for *yeah* and *no* 35 tokens of each speech unit were eligible for analysis (across 14 speakers).

4.2.3.3. Duration

Absolute durations of each frequently occurring speech unit were obtained using a Praat script written by the present author, by calculating the duration of the interval between the marked onset and offset points of each individual speech unit.

4.2.4. Normalisation

For all of the frequently occurring speech units, all of the raw static and dynamic measurements outlined in the previous sections were subjected to z-score normalisation (by-speaker) in order to control for effects such as imprecisions in the distance between mouth and microphone, articulation rate and the likelihood that some speakers will be inherently louder or quieter than others. This normalisation method was deemed appropriate in order to isolate the features in focus for this study. As discussed at the start of this chapter, the analysis of individual speech units may be considered unfitting for a study orientated towards spontaneous speech rhythm given that rhythm is usually thought of as being a sequence of a number of different components over a stretch of speech. The normalisation procedure adopted here is one which serves to alleviate any such apparent discord. That is, all of the raw

measurement values obtained from each speech unit were z-score normalised against the corresponding measurements of the respective speaker's 18 × 9-syllable spontaneous speech data. For example, when calculating the z-score for the *peak intensity* of a given speech unit, this was done in the following way:

$$z_k = \frac{(y_k - \bar{y}_k)}{\sigma_k}$$

Where:

y_k = raw *peak intensity* value from the *speech unit* of a given speaker

$\bar{y}_k$ = the mean of all the raw syllabic *peak intensity* values from the *spontaneous speech* of a given speaker

σ_k = the standard deviation of all the raw syllabic *peak intensity* values from the *spontaneous speech* of a given speaker.

When calculating the z-scores for the dynamic measurements of intensity and f_0, each of the +10% step raw values were z-scored against the respective parameter's *spontaneous mean values*. Normalising the frequently occurring speech unit data against the 9-syllable spontaneous utterance data in this way allows us to capture the rhythmic characteristics of these units relative to the speakers' spontaneous speech patterns.

As the formants F_1-F_3 are only being measured for the purpose of facilitating comparisons between these formant-based measurements and the rhythm-based measurements for the filled pauses *er* and *erm*, and that corresponding measurements were not obtained from the speakers spontaneous 18 × 9-syllable utterances, the z-scores for these measurements were calculated in the following, more typical, manner:

$$z_k = \frac{(y_k - \bar{y}_k)}{\sigma_k}$$

Where:

y_k = the raw value of the point to be measured (e.g., F_1 +50% interval);

$\bar{y}_k$ = the mean of the nine raw values across the speech unit;

σ_k = the standard deviation of the nine raw values across the speech unit.

4.2.5. Statistical analysis

All data in the current chapter were analysed using the same procedures and methods as described in Chapter 3, Section 3.2.6. The discriminatory power of the four speech units is assessed using LDA, whereas GAMMs are used as a means of assessing the significance of interactions between parameters and for visualising intensity and f_0 contours pertaining to the dynamic measurements obtained.

In relation to the LDA setup, for the static measurements of the speech units, predictors are the single values attributed to a given unit (e.g., mean f_0), whereas for the dynamic measurements, predictors are the nine sequential values (+10% values) across the trajectory of the speech (e.g., nine intensity measurements at +10% steps). A 'group' is an individual speaker as represented by the collection of speech units analysed (for further detail, see Chapter 3, Section 3.2.6). It is acknowledged here that the nine predictor variables for a given speech unit are likely to be correlated to a degree and this will be taken into consideration when interpreting the results for the dynamic measurements. The issue of there being highly correlated variables is one shared and accepted by many studies which make use of speech features as it is often the case that such features correlate with one another to some degree. It is also worth acknowledging that in taking dynamic measurements, a given speech unit is subsequently represented by nine values (as opposed to a single value, for example, a midpoint measurement). If improved discrimination performance is obtained through dynamic measurements, it could be argued that this improved performance could be a result of there simply being more data points per speaker, rather than it being that the dynamic information of the speech units is useful. Again, this is a consideration which will be taken into account when interpreting the results, and is a consideration shared by previous studies which have made use of dynamic speech features such as dynamic formant measurements (e.g., Hughes et al. 2016; McDougall, 2004, 2006).

4.3. Results

This section presents analysis of the four frequently occurring speech units *er*, *erm*, *yeah* and *no*. These speech units are hypothesised as potentially being useful markers of a speaker's speech rhythm given that they may frequently punctuate utterances at certain positions (depending on the function of the item such as to initiate a speech turn), and that they might also be more susceptible to prosodic inflections in comparison to the lexical content of the rest of an utterance (for further discussion see Chapter 2, Section 2.6). As discussed above in Section 4.2.3, the rhythmic characteristics (measurements of intensity, f_0 and duration) of these speech units have been analysed relative to the speakers' spontaneous speech rhythm patterns in order to determine to what extent these units are able to distinguish between speakers. As such, comparisons will be able to be made between the results for the speakers' frequently occurring speech unts and their spontaneous speech rhythm patterns. Analysis will also facilitate observation as to whether measurements of a certain parameter are more useful than that of another, or whether it is the combinations and interrelations of these parameters which signal speaker individuality.

The following section provides a brief overview of the LDA results for each of the frequently occurring speech units analysed in order to show which of the speech units harnessed the most speaker discriminatory potential. This will allow for a more in-depth analysis to be attributed to the most forensically promising speech unit. Following this, the filled pauses *er* and *erm* are analysed in Section 4.3.2 to determine whether any speaker-specific patterns are present across the three parameters studied. Lastly, in Section 4.3.3, comparisons are made between the two monosyllabic responses *yeah* and *no* and their discriminatory power is assessed.

4.3.1. Overview of LDA results

Table 4.3 displays the LDA classification rates for each of the frequently occurring speech units in relation to the three parameters measured.

Table 4.3. Summary of results from linear discriminant analyses. Note that the speaker numbers vary between erm/er and yeah/no. Chance = 8.3% for erm/er, and 7.1% for yeah/no.

Speech unit	Classification rate (%)										
	Dynamic		Static								
			midpoint		mean		peak		trough		duration
	intensity	f_0	int	f_0	int	f_0	int	f_0	int	f_0	
erm (full)	51.9	54.2	20.8	46.7	21.5	12.1	21.7	14.2	21.9	38.5	16.7
erm (vowel)	26.5	22.9	21.0	15.4	21.1	14.8	21.7	14.6	15.8	17.7	19.8
er	29.4	16.5	21.0	9.4	19.3	7.7	17.9	10.8	16.7	13.5	10.8
yeah	23.5	10.6	16.3	9.2	18.2	10.4	16.1	7.8	21.2	13.7	12.5
no	30.0	15.5	10.0	15.1	8.8	15.1	8.8	13.1	15.1	16.9	14.3

It can be observed that the speech unit which shows the most speaker-specificity is the filled pause *erm* when the unit is considered as a whole (that is, conjoined vowel and nasal portions). Specifically, it is the dynamic intensity contour (CR = 51.9%) and f_0 contour (CR = 54.2%) of this speech unit which appear to be the most promising.

Table 4.4 provides a summary of the classification rates for each of the frequently occurring speech units when the three rhythmic parameters (intensity, f_0 and duration) are combined.

Table 4.4. Summary of results from linear discriminant analyses with the three parameters (intensity, f_0 and duration) combined. Note that the speaker numbers vary between *erm*/*er* and *yeah/no*. Chance = 8.3% for *erm/er*, and 7.1% for *yeah/no*.

Speech unit	Classification rate (%)				
	Dynamic	Static			
		midpoint	mean	peak	trough
erm (full)	81.3	72.5	29.8	26.9	54.0
erm (vowel)	40.6	32.9	35.4	33.3	32.5
er	31.3	26.3	17.5	17.9	21.7
yeah	27.3	19.8	25.1	23.7	29.4
no	36.3	19.6	16.5	17.3	20.8

Table 4.4 shows that the speaker discriminatory potential for each of the four speech units is improved when intensity, f_0 and duration measurements are combined before being subjected to LDA. Again, it is the filled pause *erm* which yields the most

promising classification rate (81.3%) in relation to the dynamic measurements (that is, the intensity contour + the f_0 contour + duration) and as such this speech unit will be subject to a greater depth of analysis in order to determine reasons as to why this is the case.

4.3.2. Filled pauses

4.3.2.1. Erm: combined vocalic and nasal portion

Table 4.5 summarises the CRs for all of the dynamic and static measurements taken for *erm* and Figure 4.6 shows the classification rates yielded from the linear discriminant analysis results for the filled pause *erm* dynamic contour across both the vocalic and nasal portions.

Table 4.5. LDA results for the dynamic measurements (contour and midpoint + 90% interval) and the static measurements (mean, peak, trough and midpoint) for *erm* (vocalic and nasal portion together). N.b., where 'n/a' is reported this indicates that classification rates were not computed for these specific measurements.

Measure	**Classification rate (%)**				
	Dynamic		**Static**		
	contour	**midpoint +90% interval**	**mean**	**peak**	**trough**
Duration	16.7	n/a	n/a	n/a	n/a
Intensity	51.9	39.2	21.5	21.7	21.9
f_0	54.2	53.8	12.1	14.2	38.5
F_1	36.5	n/a	n/a	n/a	n/a
F_2	40.0	n/a	n/a	n/a	n/a
F_3	32.9	n/a	n/a	n/a	n/a
Intensity + f_0	80.6	77.3	23.8	26.3	47.9
F_1-F_3	64.8	45.0	33.1	30.0	28.5
All	91.0	81.5	45.6	39.4	58.5
Intensity + duration	55.4	42.3	26.7	27.9	23.8
f_0 + duration	53.5	55.6	18.1	18.5	44.2
Intensity + f_0 + duration	81.3	76.9	29.8	26.9	54.0
F_1-F_3 + duration	65.6	47.1	35.4	32.5	31.0
All + duration	90.8	81.5	49.2	41.9	60.0

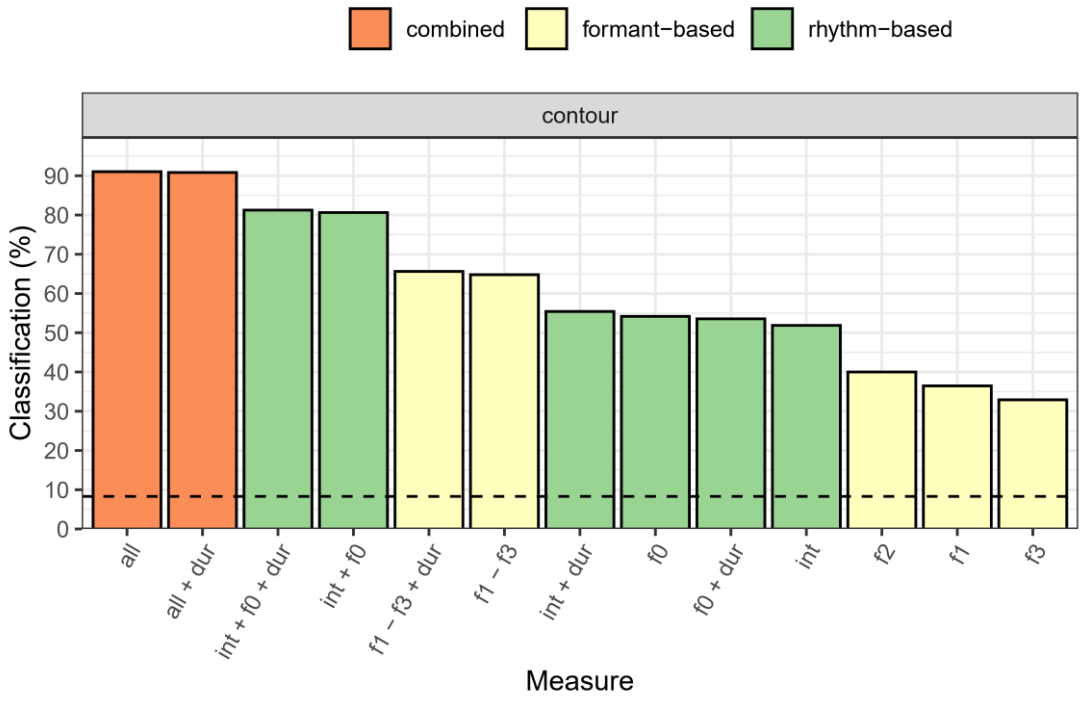


Figure 4.6. Discriminant analysis results for the filled pause *erm* dynamic contour across both the vocalic and nasal portions (ranked from highest to lowest). Chance level is 8.3% (indicated by dotted line) as there are 12 speakers.

In comparison to the results obtained from the spontaneous utterances in Chapter 3 (Section 3.3), the filled pause *erm* is able to distinguish between speakers to a substantially better standard. Where previous studies have segmented the vocalic and nasal portions of *erm* and analysed them separately from the outset (due to the focus of format frequencies of the vocalic portion), here we examine *erm* in its entirety to start with. From Figure 4.6 it is evident that taking a dynamic approach to *erm* is particularly useful with regards to intensity (CR = 51.9%) and f_0 (CR = 54.2%). When factoring in duration to each of these measures, intensity yields a slightly improved CR of 55.4% (interaction: $p = 0.008$) whereas the CR for f_0 drops slightly to 53.5% (interaction: $p = 0.099$ (n.s.)) What appears most promising is the combination of the intensity contour and the f_0 contour which gives a CR of 80.6% (interaction: $p < 0.0001$) and which is further improved (marginally) with the inclusion of duration (CR = 81.3%). Overall, it is observable that these rhythm-based measures perform better than formant-based measures (e.g., F_1-F_3+dur, CR = 65.6%) and that a model which

combines formant-based measures with rhythm-based measures performs especially effectively (all, CR = 91.0%).

Figure 4.7 – Figure 4.10 shows the CRs for the dynamic measurement (midpoint + 90% interval) and the three static measurements mean, peak and trough.

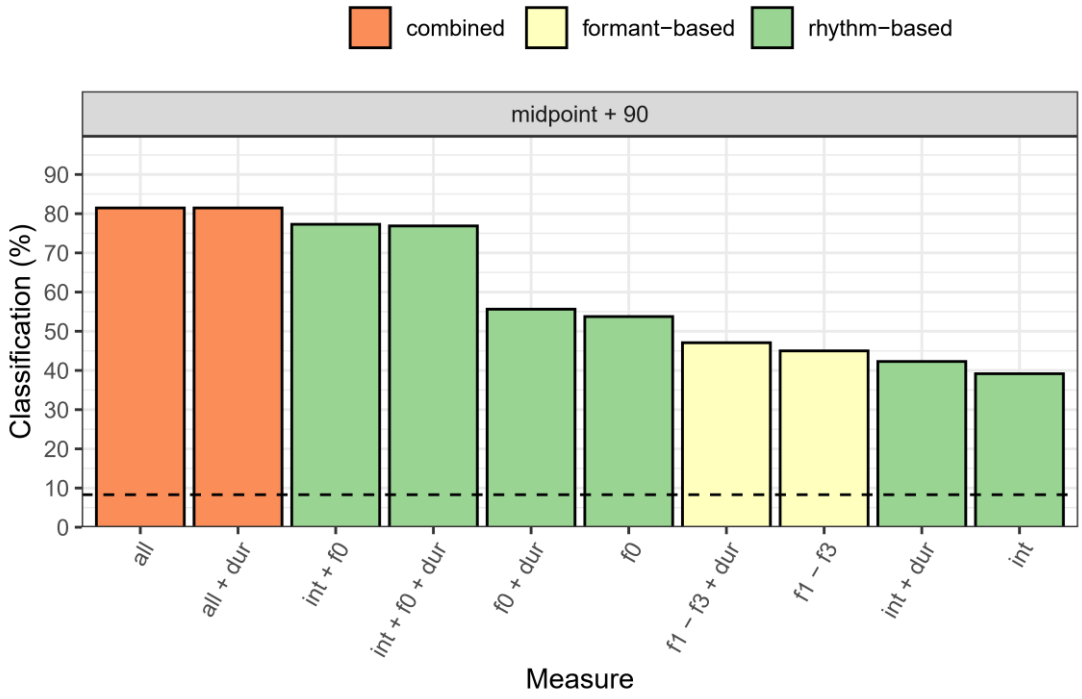


Figure 4.7. Discriminant analysis results for the filled pause *erm* for the dynamic measurement (midpoint + 90) across both the vocalic and nasal portions (ranked from highest to lowest). Chance level is 8.3% (indicated by dotted line) as there are 12 speakers.

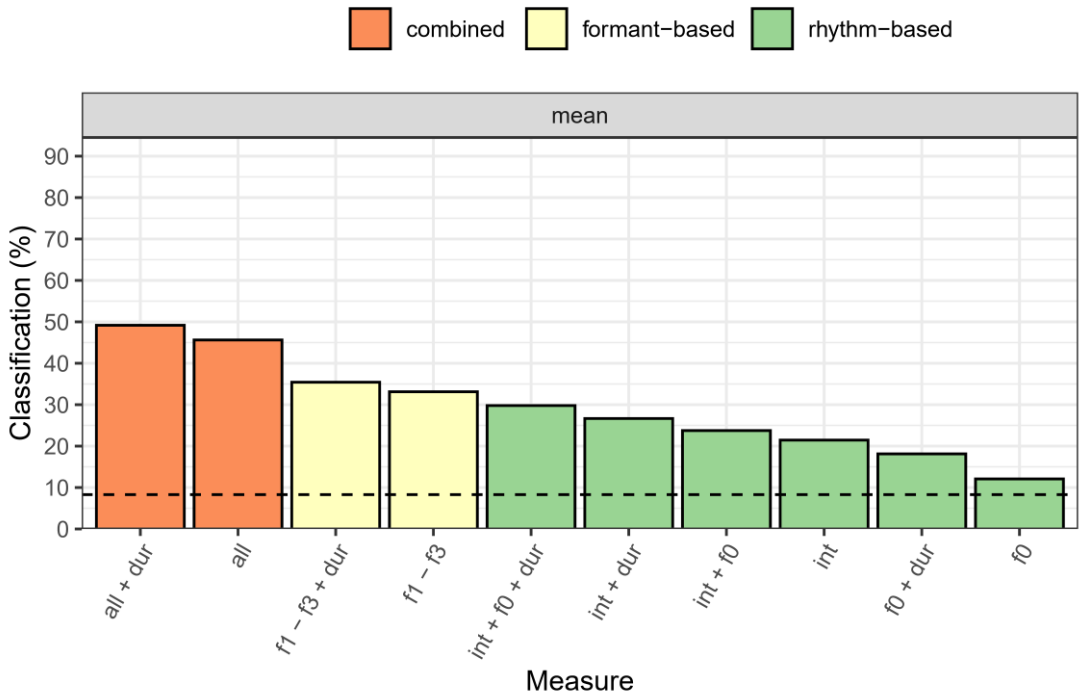


Figure 4.8. Discriminant analysis results for the filled pause *erm* for the mean static measurement across both the vocalic and nasal portions (ranked from highest to lowest). Chance level is 8.3% (indicated by dotted line) as there are 12 speakers.

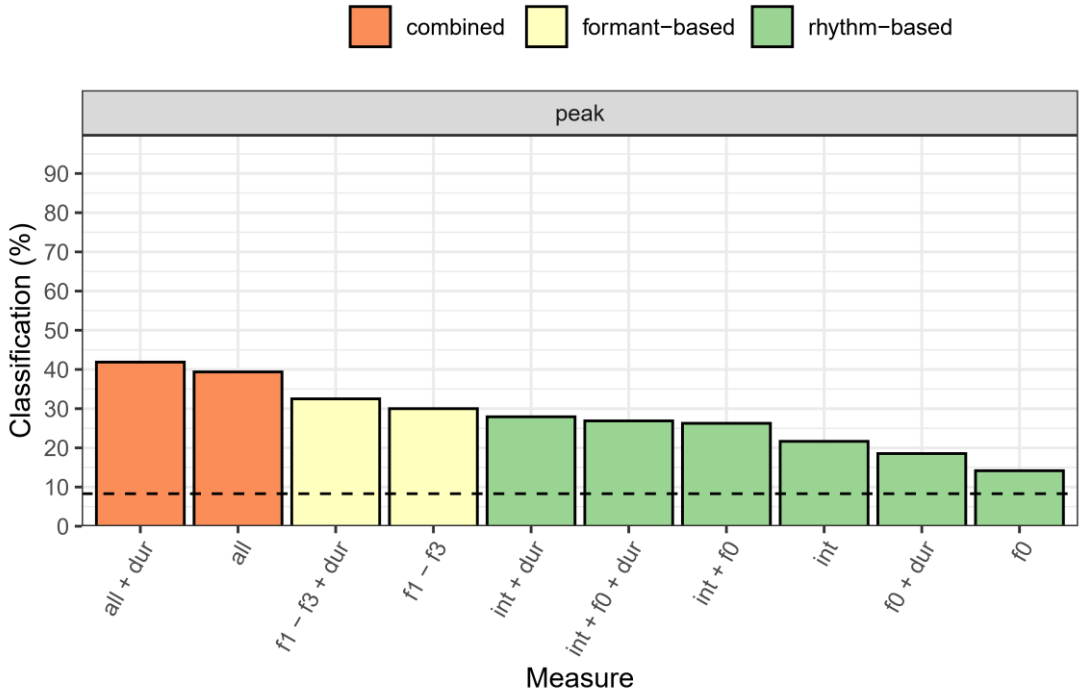


Figure 4.9. Discriminant analysis results for the filled pause *erm* for the peak static measurement across both the vocalic and nasal portions (ranked from highest to lowest). Chance level is 8.3% (indicated by dotted line) as there are 12 speakers.

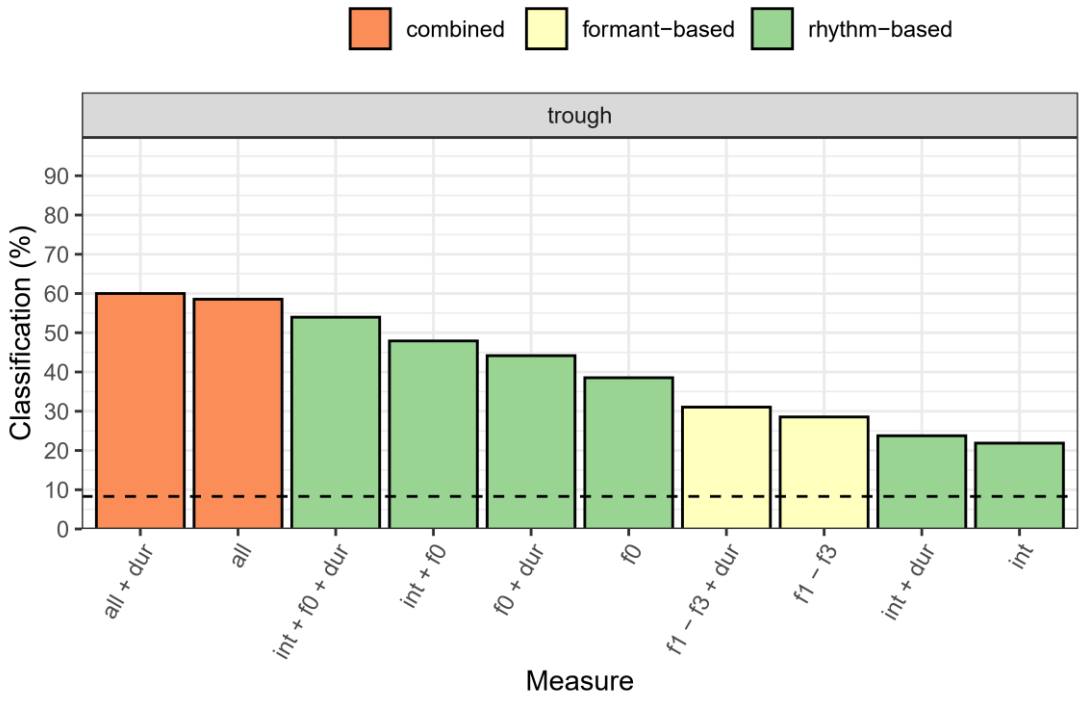


Figure 4.10. Discriminant analysis results for the filled pause *erm* for the trough static measurement across both the vocalic and nasal portions (ranked from highest to lowest). Chance level is 8.3% (indicated by dotted line) as there are 12 speakers.

For this particular filled pause, the decision was made to take a dynamic measurement of just the midpoint and the last dynamic point (+90% interval of the contour) given the anticipated drop in intensity and potential effects on f_0 as a result of the transition from the vocalic portion to the nasal portion of *erm*. During the data editing stage, when labelling the filled pauses *erm*, it was observed that the vocalic portion generally made up about 60-80% of *erm* as a whole and therefore the midpoint (50%) measurement and the final (90%) measurement were deemed appropriate to capture any distinctive intensity and f_0 fluctuations as a result of the vowel to nasal transition. As can be observed from Table 4.1 and Figure 4.7 the two measurements captured as part of the midpoint + 90% interval approach yielded comparatively strong results for both intensity and f_0, particularly when the two are considered together (intensity + f_0, CR = 77.3%). This result is only slightly shy of the CR for the combination of intensity and f_0 the whole contour (i.e., 9 points, CR = 80.6%) which suggests that this transition from vowel to nasal is where a good deal of speaker-specificity lies. To investigate this observation further, Figure 4.11 shows the by-speaker intensity

contours (top panel) and by-speaker f_0 contours (bottom panel) across the whole of the filled pause *erm*.

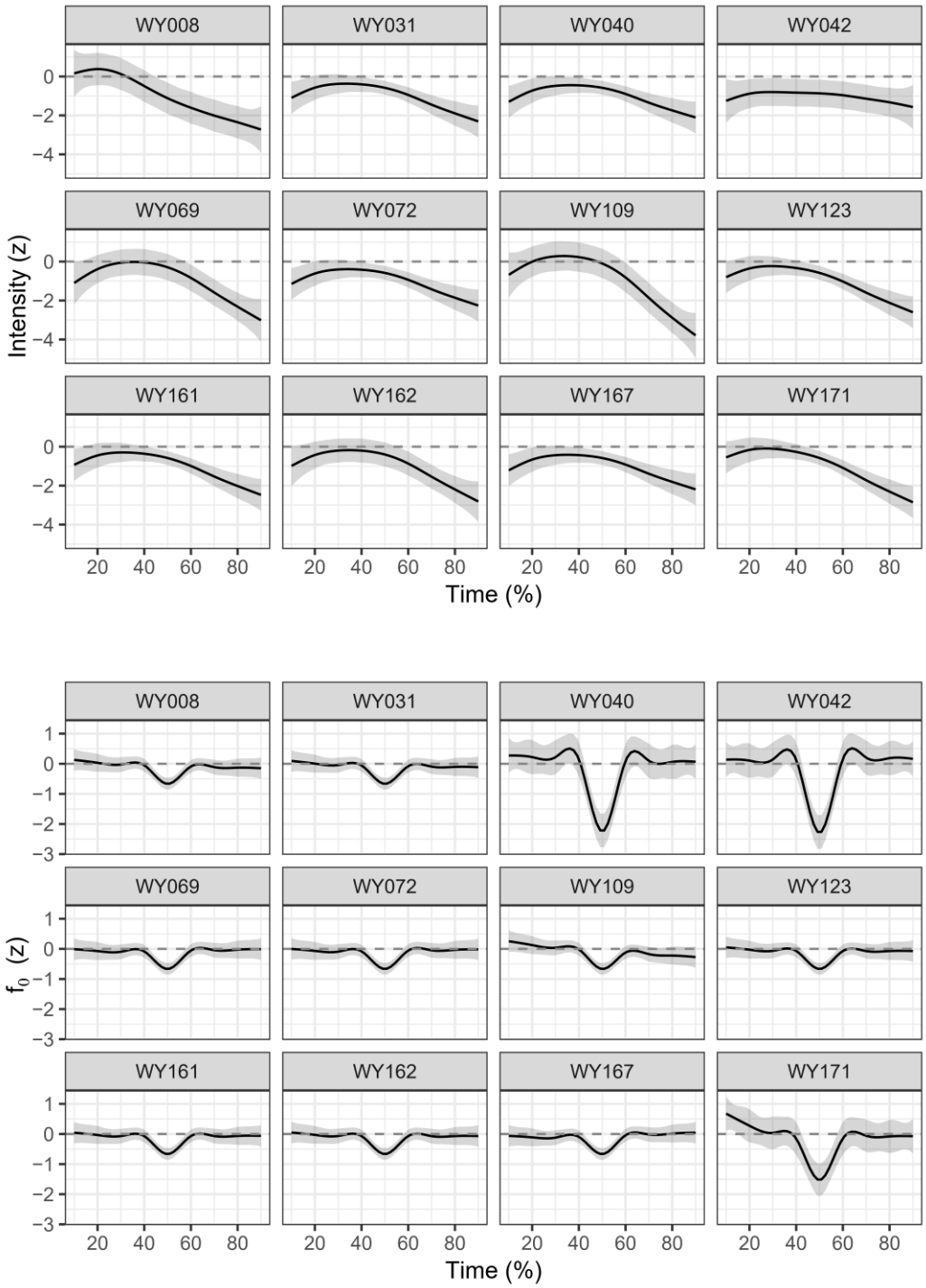


Figure 4.11. GAMM plots of by-speaker syllable-varying intensity contours (top) and by-speaker f_0 contours (bottom) for the filled pause *erm* (vocalic and nasal segments). Higher z-scores correspond to greater intensity and higher pitch respectively.

As expected, there is a general trend across the 12 speakers in relation to a drop in intensity between the 40% - 60% portion of the contour, with some speakers' trajectories exhibiting a more drastic falling slope than others. In comparing the intensity trajectories to the LDA results, it is indeed those speakers whose trajectories indicate the most variation (e.g., a dramatic fall-off in intensity) which were the best-performing (e.g., speaker 109, CR = 87.5%; speaker 008, CR = 75.0%). Figure 4.12 provides an illustration of the LDA results for the dynamic intensity contour where speaker WY109 is represented by the yellow ellipse.

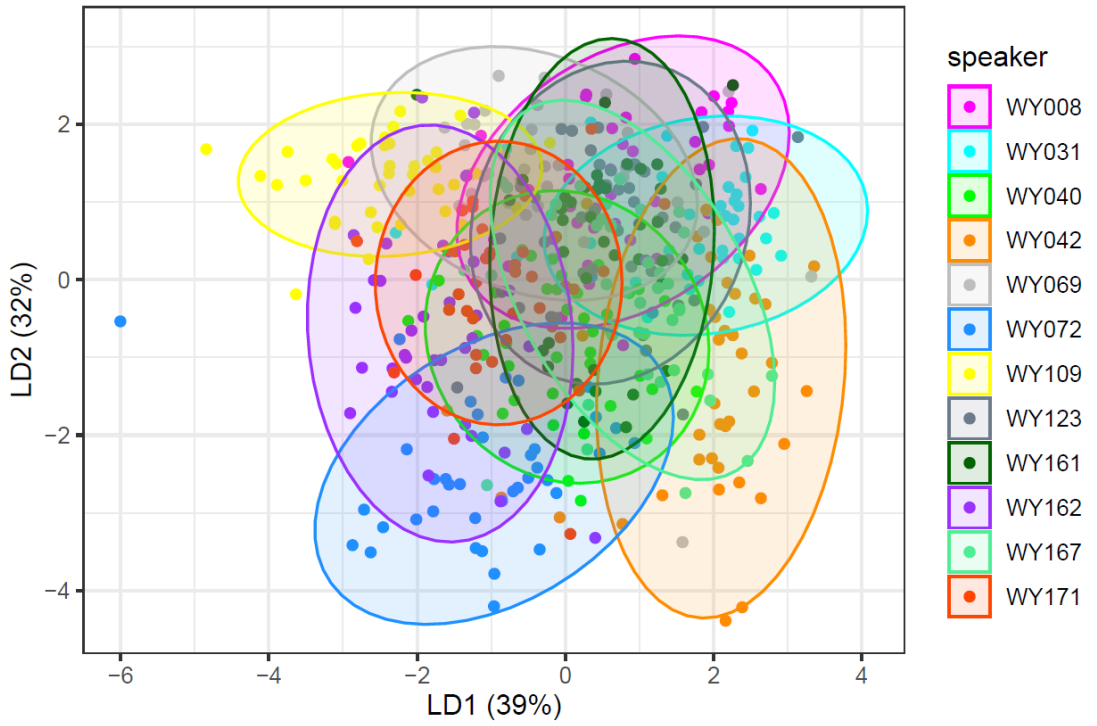


Figure 4.12. Visualisation of the LDA results for the dynamic intensity contour of *erm* (vocalic and nasal portions). Overall classification rate = 51.9% (chance = 8.3% as there are 12 speakers).

Figure 4.11 also reveals a notable trend in the speakers' f_0 contours in that the transition from vowel to nasal appears to cue a temporary drop in f_0, again between the 40% - 60% area. For some speakers, this temporary decrease is more pronounced than others (e.g., speakers 040, 042, 171) and when referencing these observations with the LDA results, these speakers are indeed well-distinguished through their f_0 contours (speaker 040, CR = 62.5%; speaker 042, CR = 62.5%; speaker 171, CR =

92.5%). Given the marked visual differences in the contours of these speakers, their occurrences of *erm* were inspected again within Praat, especially in light of speaker 171's remarkably high CR. In observing this speaker's individual plot, what distinguishes this speaker from the rest is the comparatively higher pitch at the start of the contour with a z-score closer to 1 as opposed to the other speakers whose f_0 contours start closer to 0. Additional auditory and acoustic examination in Praat reinforces the GAMM visualisation in that this speaker does have a raised initial pitch across the majority of his *erm* tokens. Figure 4.13 below serves to exemplify speaker 042's temporary drop in f_0 at the transition from vowel to nasal and Figure 4.14 demonstrates speaker 171's initial raised pitch (f_0 is indicated by the blue line in both panels). Both Figure 4.13 and Figure 4.14 also serve to reinforce the speakers' plotted intensity patterns, with speaker 042 (Figure 4.13) exhibiting a relatively stable intensity contour throughout the vocalic and nasal section of *erm* whereas speaker 171 (Figure 4.14) exhibits a drastic drop in intensity cued by this transition (intensity indicated by the yellow line in both panels).

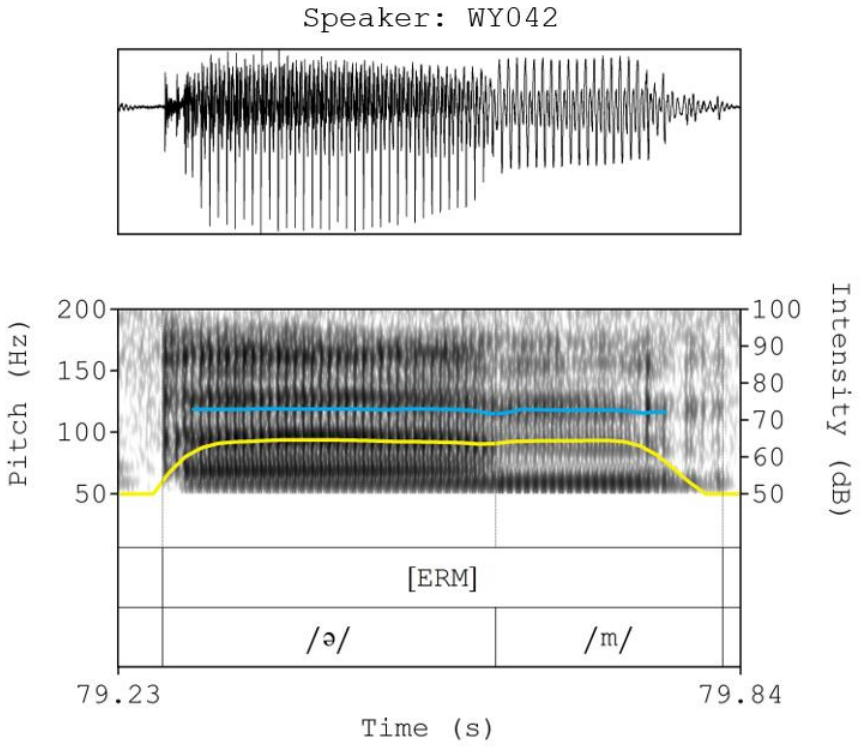


Figure 4.13. Waveform, spectrogram and TextGrid of one of speaker 042's *erm* tokens. Intensity is indicated by the yellow line and f_0 is indicated by the blue line.

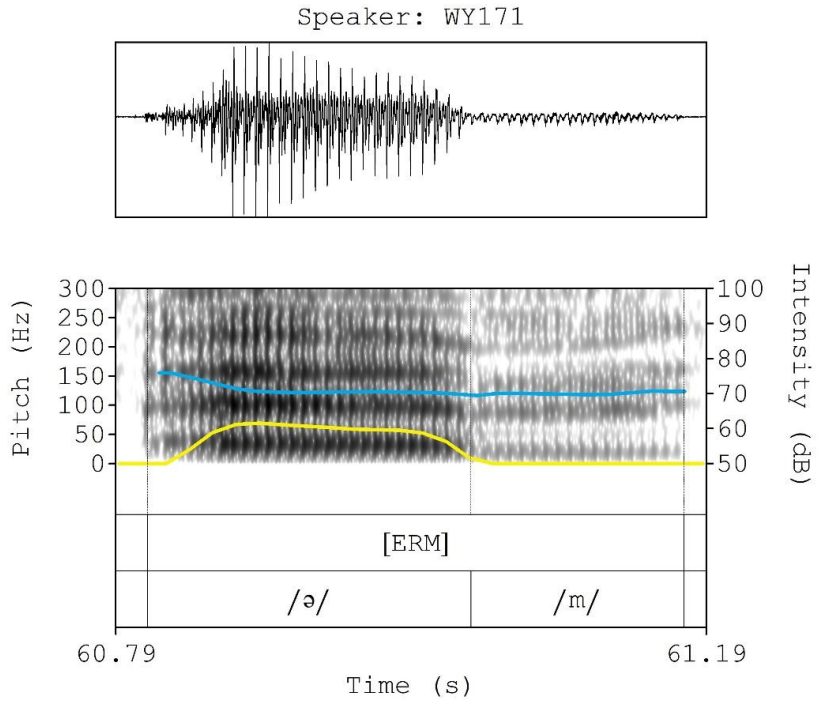


Figure 4.14. Waveform, spectrogram and TextGrid of one of speaker 171's *erm* tokens. Intensity is indicated by the yellow line and f_0 is indicated by the blue line.

Focussing now on the LDA results pertaining to when the three rhythmic parameters are combined together, Table 4.6 details the classification rates for each speaker in relation to the dynamic measurements (that is, the intensity contour + the f_0 contour + duration) and Figure 4.15 provides a visualisation of these results.

Table 4.6. Classification rates for each speaker in relation to the dynamic measurements for the combination of all three rhythmic parameters. Overall CR for the speech unit = 81.3% (chance = 8.3% as there are 12 speakers).

Speaker ID	WY 008	WY 031	WY 040	WY 042	WY 069	WY 072	WY 109	WY 123	WY 161	WY 162	WY 167	WY 171
CR (%)	90	80	85	92.5	55.5	80	85	72.5	82.5	67.5	85	100

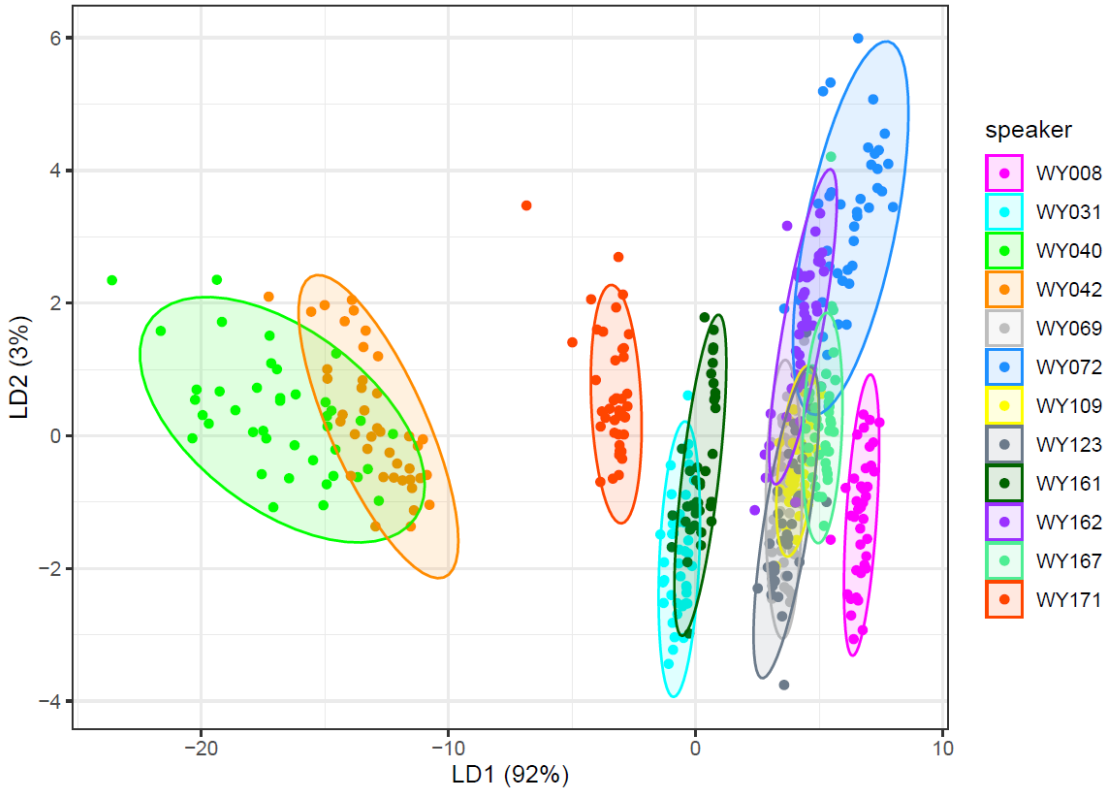


Figure 4.15. Visualisation of the LDA results for the dynamic measurements for the combination of all three rhythmic parameters. Overall CR = 81.3% (chance = 8.3% as there are 12 speakers).

Again, it is speakers 171 (CR = 100%; red ellipse) and 042 (CR = 92.5%; orange ellipse) who demonstrate the most speaker-specificity in relation to the combined dynamic LDA results. In directing focus towards speaker 171, who the LDA was able to discriminate correctly categorically, Figure 4.16 provides an example of this speaker's use of the filled pause *erm* in the context of an utterance.

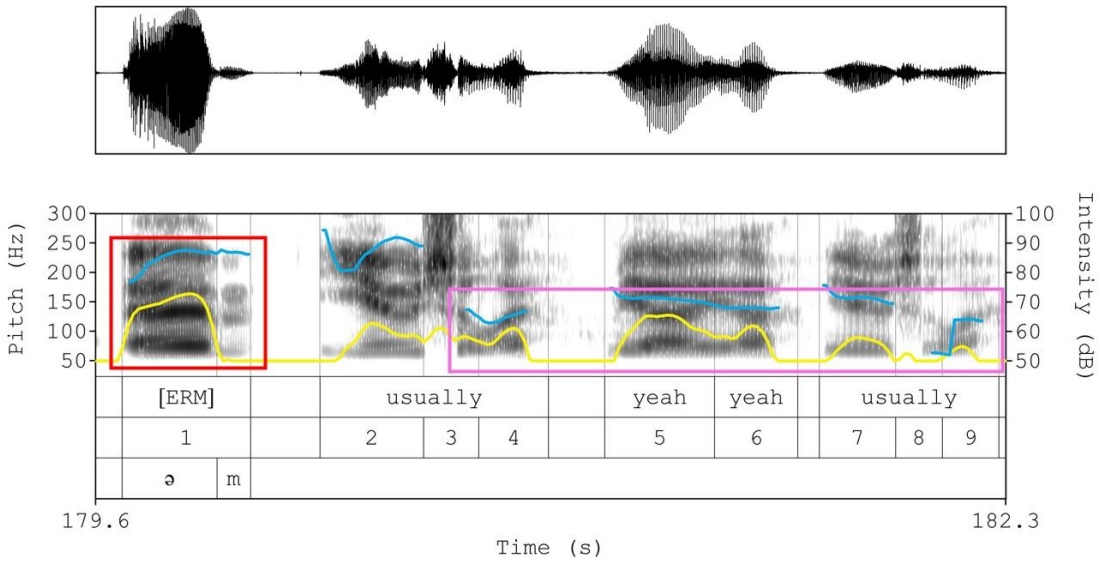


Figure 4.16. Example of speaker 171's use of the filled pause *erm* in the context of an utterance (in this instance, at the start of a response to a question). Intensity contour and f_0 contour are indicated by the yellow line and blue line respectively.

Figure 4.16 highlights the difference in the rhythmic patterns of this speaker's filled pause *erm* in comparison to the rest of the utterance shown, as well as corroborating the patterns shown in Figure 4.11 and Figure 4.14. Highlighted within the red rectangle, the acoustic composition of this speech unit is one of a higher pitch than the rest of the following utterance (highlighted within the pink rectangle), with a marked decline in intensity cued by the transition from the vocalic section to the nasal section. It is also observable that this speech unit has a longer duration than that of the following syllables. Although this visualisation provides good support for the classification rate achieved for this speaker, it must be conceived that a categorically correct classification rate, as promising as it appears, may be subject to some scrutiny. Such scrutiny will likely pertain to the setup of the statistical analysis carried out (see Section 4.4.1 for further discussion). Nevertheless, the results reported here indicate that the rhythmic characteristics of this speech unit do carry a good deal of speaker-specific information.

4.3.2.2. Erm: vocalic portion

Table 4.7 provides a summary of the LDA results for the vocalic portion of *erm* for both the dynamic and static measurements and Figure 4.17 displays the LDA results for the contour of the vocalic portion of *erm.*

Table 4.7. LDA results for the dynamic measurements (contour) and the static measurements (midpoint, mean, peak and trough) for *erm* (vocalic portion). Chance level is 8.3% as there are 12 speakers.

Measure	**Classification rate (%)**				
	Dynamic	**Static**			
	Contour	**midpoint**	**mean**	**peak**	**trough**
Duration	19.8	n/a	n/a	n/a	n/a
Intensity	26.5	21.0	21.1	21.7	15.8
f_0	22.9	15.4	14.8	14.6	17.7
F_1	n/a	20.8	n/a	n/a	n/a
F_2	n/a	25.6	n/a	n/a	n/a
F_3	n/a	20.4	n/a	n/a	n/a
Intensity + f_0	33.8	27.3	28.8	26.7	24.4
F_1-F_3	62.3	48.3	n/a	n/a	n/a
All	74.2	64.4	n/a	n/a	n/a
Intensity + duration	34.2	30.2	29.6	31.0	25.2
f_0 + duration	22.9	26.7	26.7	21.9	30.2
Intensity + f_0 + duration	40.6	32.9	35.4	33.3	32.5
F_1-F_3 + duration	64.6	38.3	n/a	n/a	n/a
All + duration	75.8	29.6	n/a	n/a	n/a

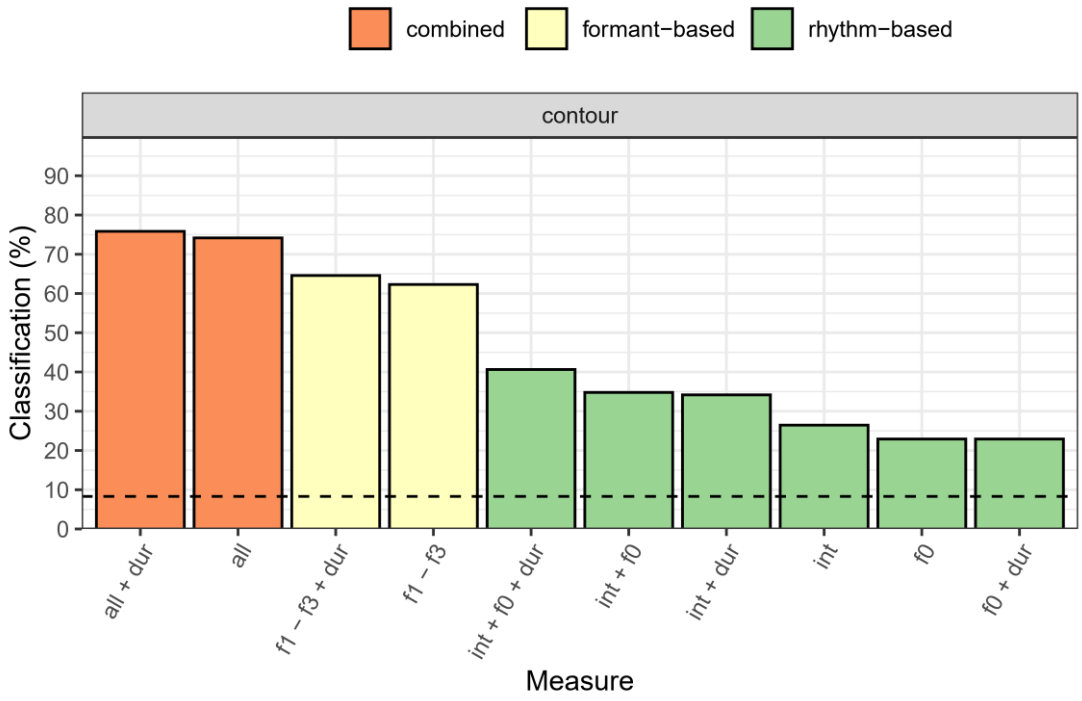


Figure 4.17. Discriminant analysis results for the vocalic portion of the filled pause *erm* dynamic contour (ranked from highest to lowest). Chance level is 8.3% (indicated by dotted line) as there are 12 speakers.

Similar to the results of *erm* when considered as a whole, speakers are best distinguished through a combination of both rhythm-based and formant-based measures, however the vocalic *erm* contour CR is lower (all+dur, CR = 75.8% vs. 90.8%). The most noticeable difference between the two sets of LDA results is that the vocalic portion of *erm* is considered on its own the formant-based measures out-perform the rhythm-based measures (F_1-F_3+dur, CR = 64.6%; int+f_0+dur, CR = 40.6%). This is also true with regards to the static midpoint measurements as exemplified in Figure 4.18. (n.b., the opposite effect is found when the nasal portion is analysed separately in that it is rhythm-based measurements that out-perform formant-based measurements, both in relation to dynamic measurements and static measurements (e.g., nasal contour: int+f_0+dur, CR = 43.3% vs. F_1-F_3+dur = 41.2%; nasal midpoint: int+f_0+dur, CR = 37.3% vs. F_1-F_3+dur, CR = 35.8%.)) For the rhythm-based measurements, it is always a combination of the three parameters (intensity, f_0 and duration) which distinguishes between speakers most effectively

both in relation to the dynamic contour (CR = 40.6%) and the static measurements, for which the best performing measure is the combination of mean intensity, mean f_0 and duration (CR = 35.4%). For all the measures, apart from the static trough measurements (Figure 4.21), intensity distinguishes between speakers more proficiently than f_0, with duration (CR = 19.8%) performing better than all of the static f_0 measurements as well as the trough intensity measurements.

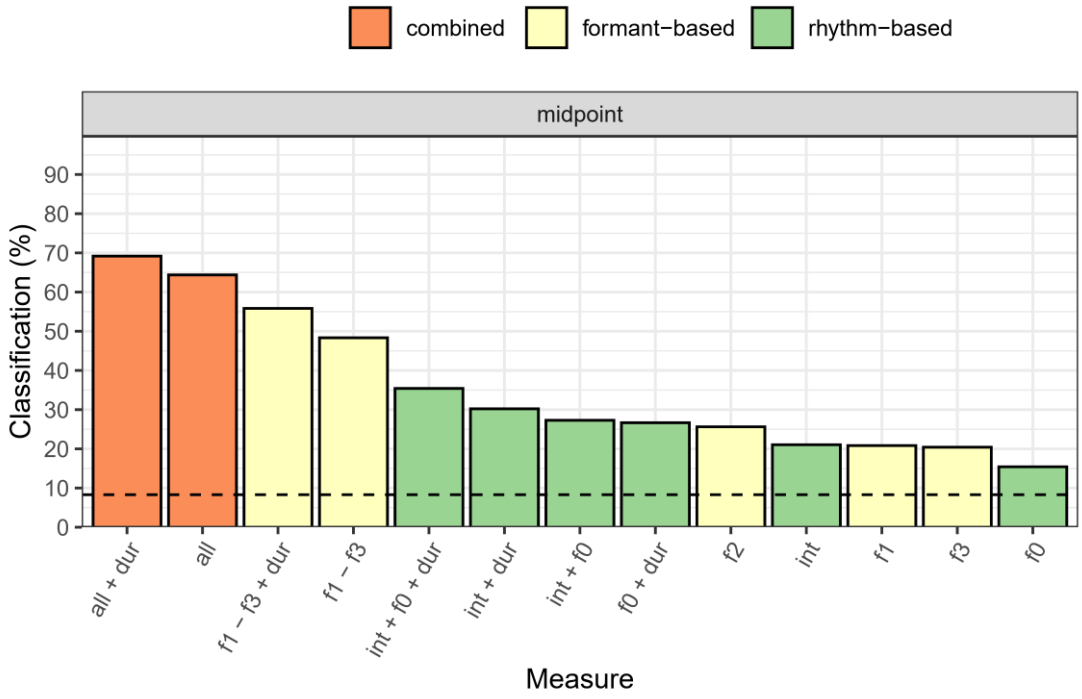


Figure 4.18. Discriminant analysis results for the vocalic portion of the filled pause *erm* for the midpoint static measurement (ranked from highest to lowest). Chance level is 8.3% (indicated by dotted line) as there are 12 speakers.

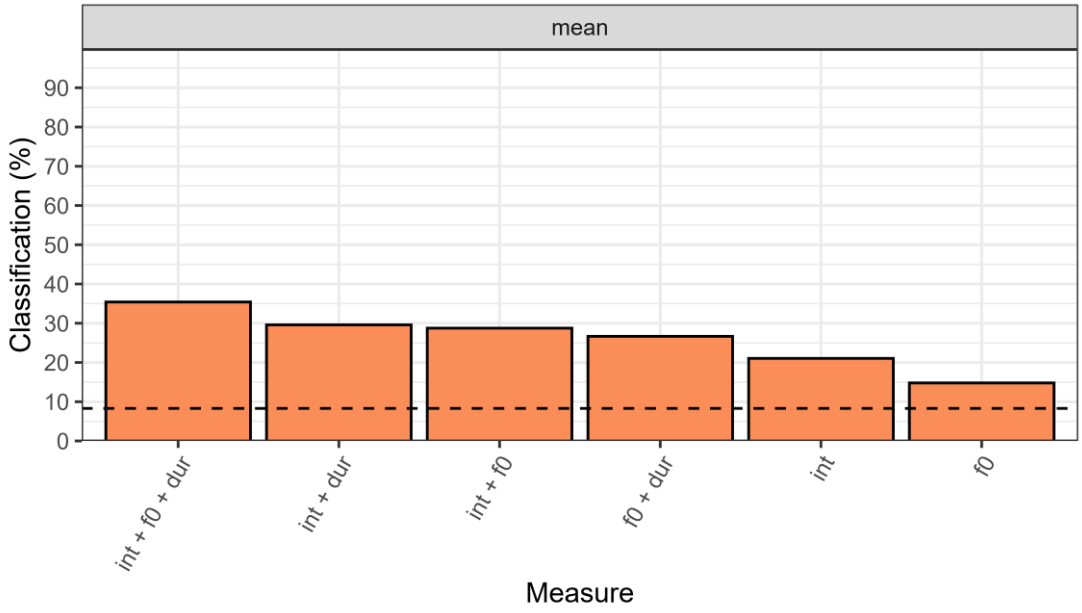


Figure 4.19. Discriminant analysis results for the vocalic portion of the filled pause *erm* for the mean static measurement (ranked from highest to lowest). Chance level is 8.3% (indicated by dotted line) as there are 12 speakers.

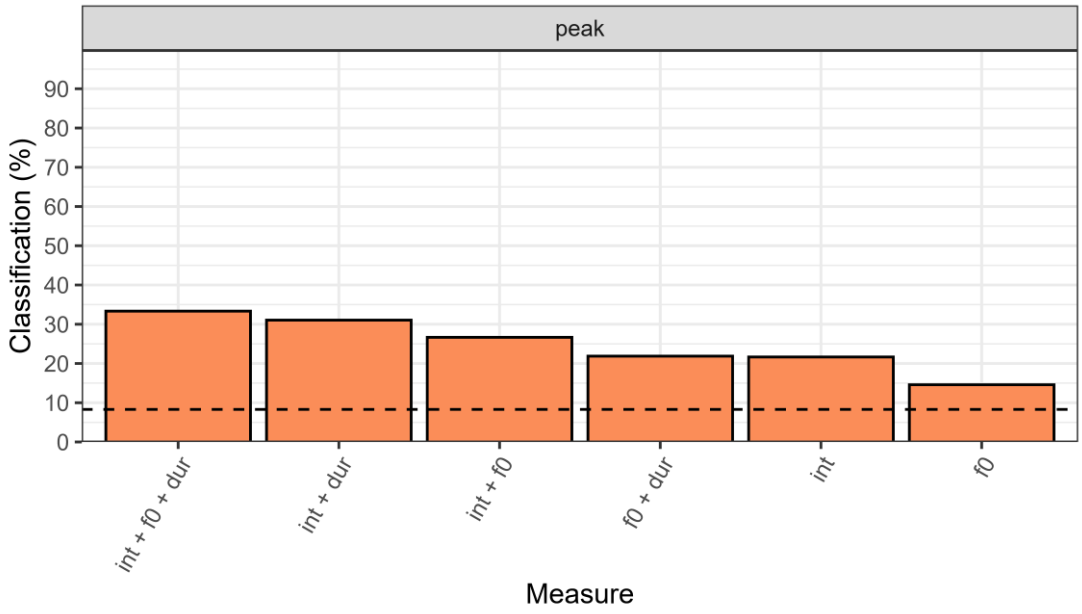


Figure 4.20. Discriminant analysis results for the vocalic portion of the filled pause *erm* for the peak static measurement (ranked from highest to lowest). Chance level is 8.3% (indicated by dotted line) as there are 12 speakers.

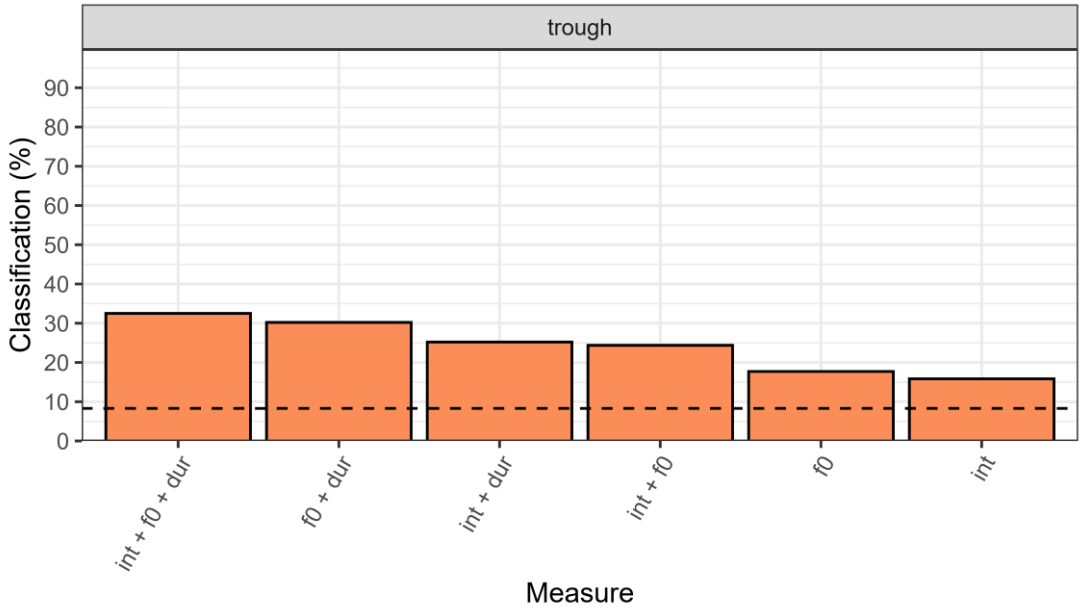


Figure 4.21. Discriminant analysis results for the vocalic portion of the filled pause *erm* for the trough static measurement (ranked from highest to lowest). Chance level is 8.3% (indicated by dotted line) as there are 12 speakers.

4.3.2.3. Er

Table 4.8 provides a summary of the LDA results for the filled pause *er* for both the dynamic and static measurements and Figure 4.22 displays the LDA results for the *er* contour.

Table 4.8. LDA results for the dynamic measurements (contour) and the static measurements (midpoint, mean, peak and trough) for *er*. Chance level is 8.3% (indicated by dotted line) as there are 12 speakers.

Measure	Classification rate (%)				
	Dynamic	Static			
	contour	**midpoint**	**mean**	**peak**	**trough**
Duration	10.8	n/a	n/a	n/a	n/a
Intensity	29.4	21.0	19.3	17.9	16.7
f_0	16.5	9.4	7.7	10.8	13.5
F_1	26.0	18.3	n/a	n/a	n/a
F_2	30.8	24.0	n/a	n/a	n/a
F_3	23.1	19.8	n/a	n/a	n/a
Intensity + f_0	29.4	25.0	18.1	16.2	22.5
F_1-F_3	51.9	44.4	n/a	n/a	n/a
All	56.7	55.8	n/a	n/a	n/a
Intensity + duration	32.5	26.5	17.1	19.6	18.5
f_0 + duration	15.4	12.1	10.6	12.9	12.9
Intensity + f_0 + duration	31.3	26.3	17.5	17.9	21.7
F_1-F_3 + duration	56.7	45.2	n/a	n/a	n/a
All + duration	58.1	56.5	n/a	n/a	n/a

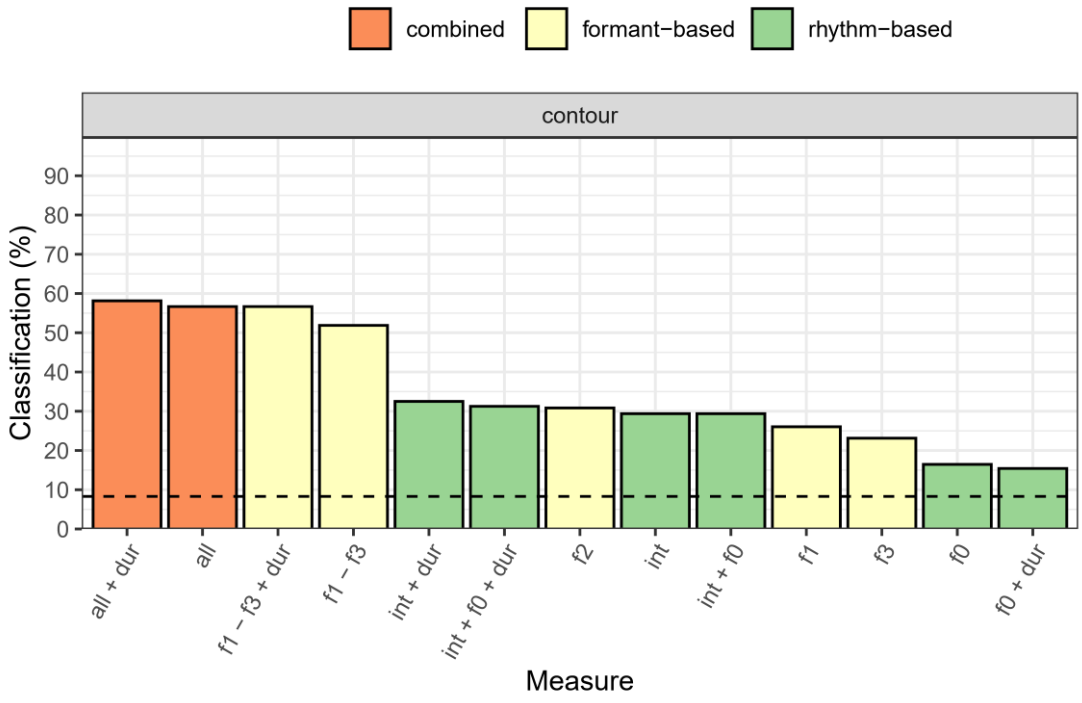


Figure 4.22. Discriminant analysis results for the filled pause *er* dynamic contour (ranked from highest to lowest). Chance level is 8.3% (indicated by dotted line) as there are 12 speakers.

The pattern of the results for the *er* contours is similar to that of the results for the vocalic portion of *erm* in that formant-based measures across the contour perform better than rhythm-based measures. The most notable difference between *er* and the vocalic portion of *erm* is that CRs are generally lower across both formant-based and rhythm-based measures. One exception to this is the static midpoint measure for F_1-F_3+dur (shown in Figure 4.23 below) which for *er* has a CR of 45.2% compared with the vocalic portion of *erm* which has a CR of 38.3%. It is also worth noting here that the midpoint F_1-F_3 measure (without the addition of duration) returned a CR of 44.4% which corresponds with the results obtained by Hughes et al. (2004) who obtained CRs of 37.2% (males, 16 speakers, chance = 6.25%) and 46.6% (females, 15 speakers, chance = 6.67%) For the rhythm-based measures, the best-performing static measurement is midpoint int+dur (CR = 26.5%; interaction: p = 0.0059), although this, along with all the other static measurements for *er*, present as having less speaker-discriminating power than the vocalic portion of *erm*.

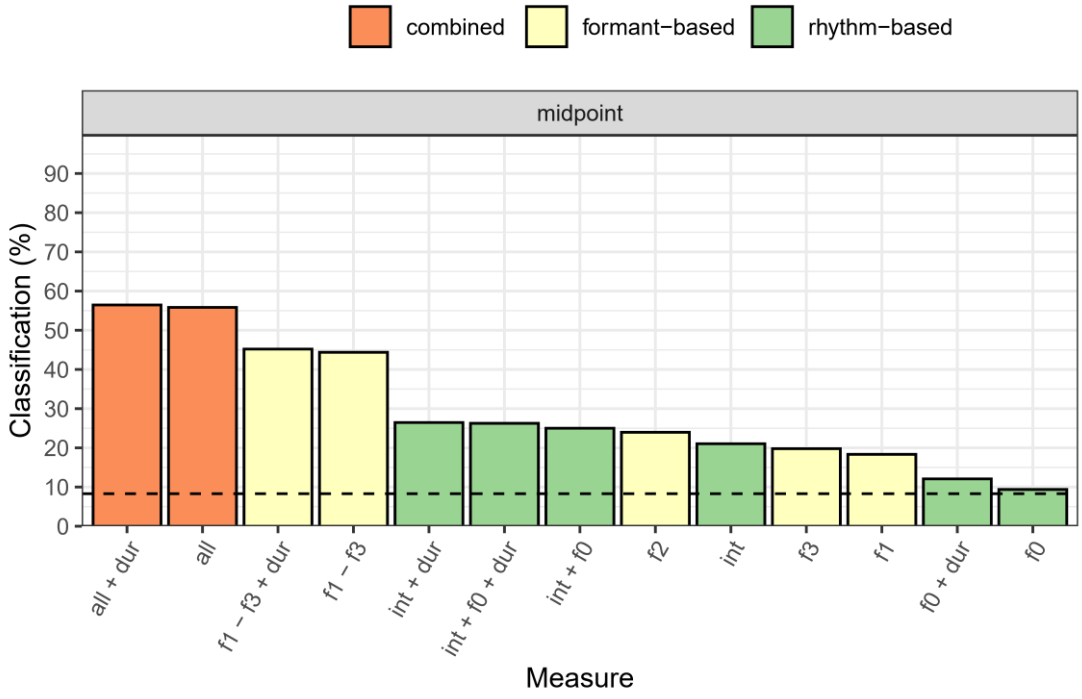


Figure 4.23. Discriminant analysis results for the filled pause *er* midpoint measurements (ranked from highest to lowest). Chance level is 8.3% (indicated by dotted line) as there are 12 speakers.

4.3.3. Common monosyllabic responses: *yeah* vs. *no*

Table 4.9 provides a summary of the LDA results for the dynamic (contour) measurements and the static (mean, peak and trough) measurements for the monosyllabic responses *yeah* and *no*. Figure 4.24 and Figure 4.25 provide a visualisation of the LDA results for the dynamic contours for *yeah* (cyan) and *no* (magenta), respectively.

Table 4.9. LDA results for the dynamic measurements (contour) and the static measurements (midpoint, mean, peak and trough) for *yeah* and *no*. Chance level is 7.1% as there are 14 speakers.

Speech unit	Measure	Classification rate (%)			
		Dynamic	Static		
		contour	**mean**	**peak**	**trough**
yeah	Duration	12.5	n/a	n/a	n/a
	Intensity	23.4	18.2	16.1	21.2
	f_0	10.6	10.4	7.8	13.6
	Intensity + f_0	24.7	18.9	19.7	23.6
	Intensity + duration	27.1	23.5	21.8	25.1
	f_0 + duration	14.7	16.7	15.7	17.1
	Intensity + f_0 + duration	27.3	25.1	23.7	29.4
no	Duration	14.3	n/a	n/a	n/a
	Intensity	30.0	8.8	8.8	15.1
	f_0	15.5	15.1	13.1	16.9
	Intensity + f_0	35.3	12.6	14.7	19.6
	Intensity + duration	31.8	14.3	13.5	19.2
	f_0 + duration	16.3	16.3	16.7	20.0
	Intensity + f_0 + duration	36.3	16.5	17.3	20.8

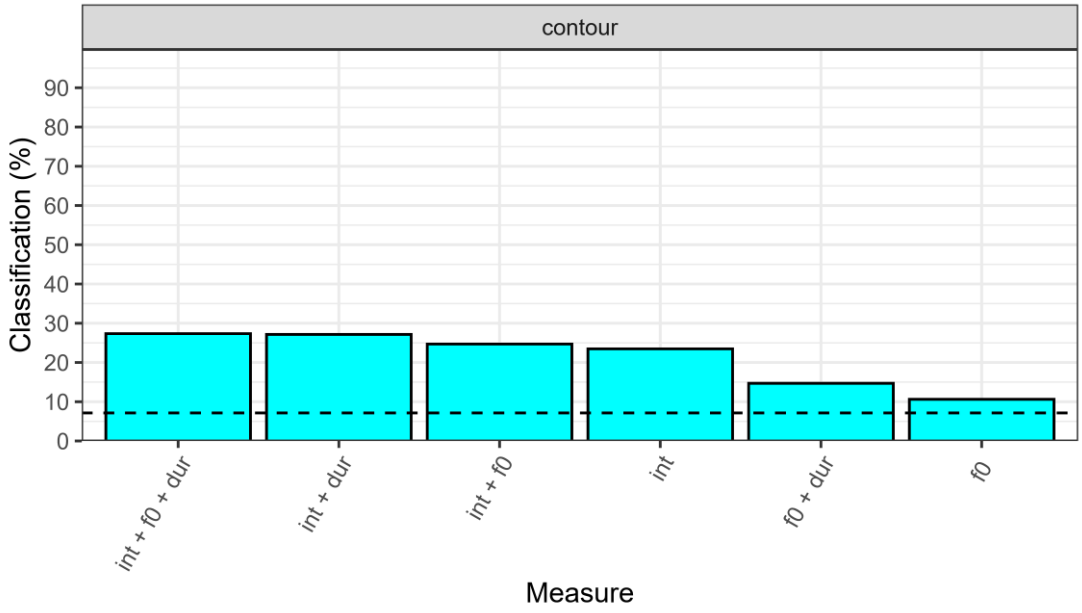


Figure 4.24. LDA results for the dynamic contour measurements of the monosyllabic response *yeah* (ranked from highest to lowest). Chance level is 7.1% (indicated by dotted line) as there are 14 speakers.

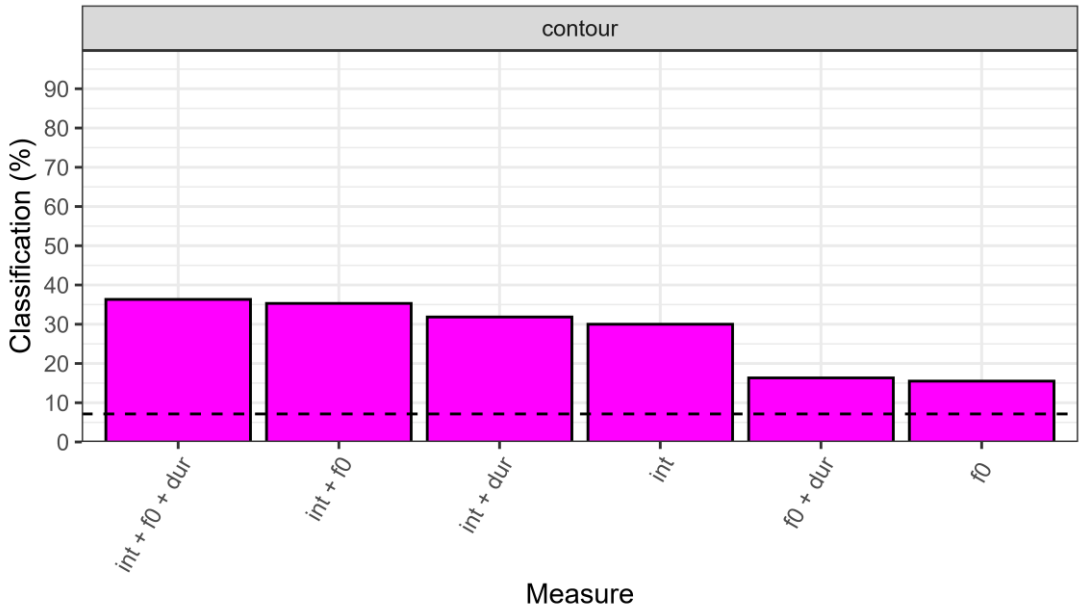


Figure 4.25. LDA results for the dynamic contour measurements of the monosyllabic response *no* (ranked from highest to lowest). Chance level is 7.1% (indicated by dotted line) as there are 14 speakers.

One observable trend for both dynamic and static measurements that is evidenced for both *yeah* and *no* is that it is a combination of the three rhythm parameters (intensity, f_0 and duration) which distinguishes between speakers most effectively. In relation to the dynamic contours, for *yeah* there are significant interactions between intensity and f_0 (p = 0.003), and f_0 and duration (p = 0.004), and for *no* there are significant interactions between f_0 and duration (p = 0.0001) and intensity and f_0 (p < 0.0001). For both *yeah* and *no*, dynamic measurements consistently outperform static measurements, apart from one exception where the static trough measurement int+f_0+dur for *yeah* yields a slightly higher CR than the CR for the dynamic int+f_0+dur (29.4% vs. 27.3%). Of the two types of monosyllabic response, *no* performs better than *yeah* in terms of dynamic contour measurements (e.g., *no*: int+f_0+dur, CR = 36.3%; vs. *yeah*: int+f_0_dur, CR = 27.3%), whereas, for static measurements of mean, peak and trough, *yeah* is seen to be the best-performing (e.g., mean intensity *yeah*, CR = 18.2% vs. mean intensity *no*, CR = 8.8% (see Figure 4.26 for a visualisation of this specific result and evidence of greater between-speaker variation and less within-speaker variation for *yeah*). One exception to this for the static measurements is that *no* performs better than *yeah* for measurements of f_0, and it is the combination of f_0 and duration that performs better for *no*, whereas for *yeah* it is the combination of intensity and duration. It should be noted that despite these differences being present, the differences between classification rates are only minor.

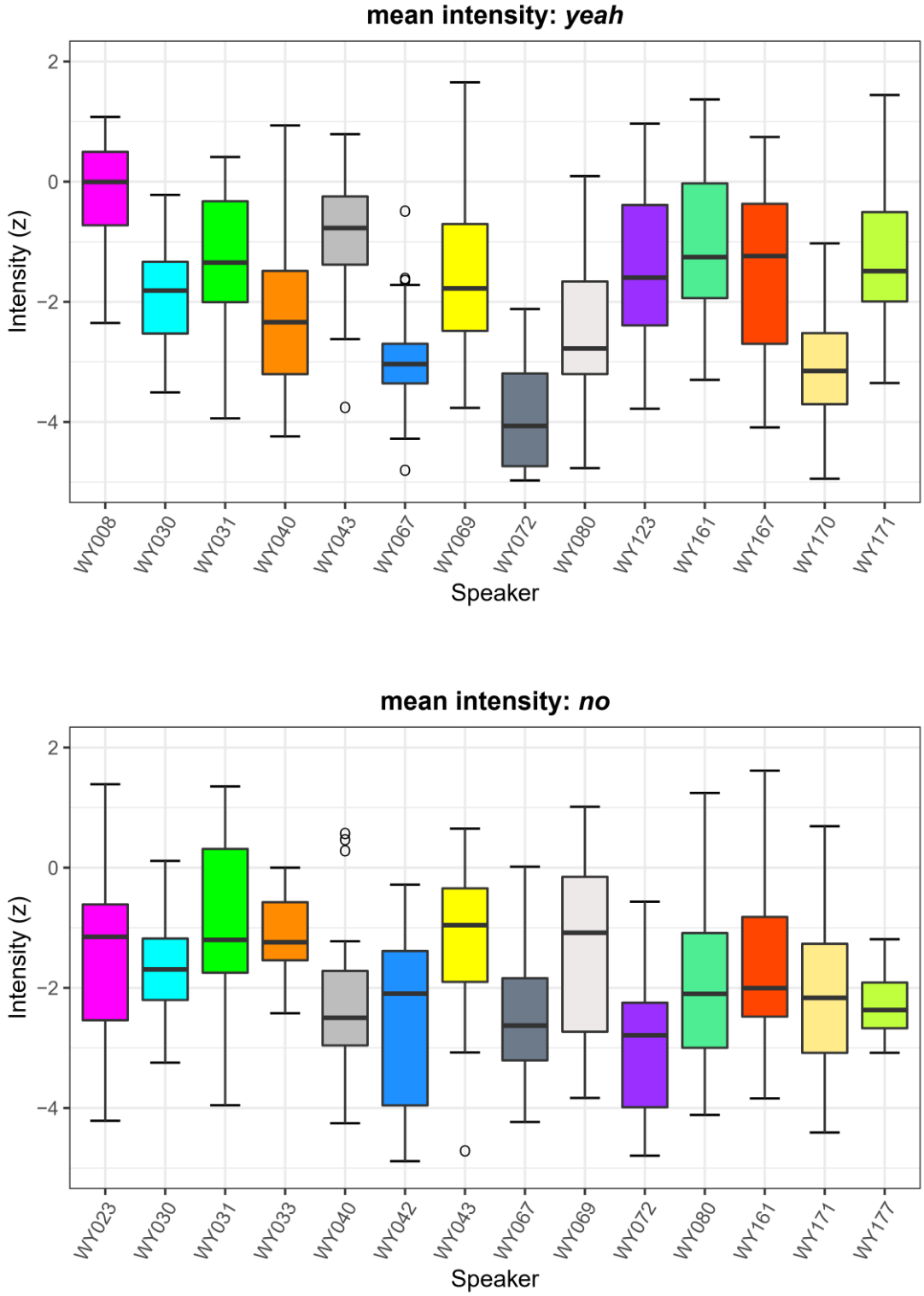


Figure 4.26. Boxplots of the 14 speakers' mean intensity measurements (z-scores) for the monosyllabic responses *yeah* (top panel, LDA CR = 18.2%) and *no* (bottom panel, LDA CR = 8.8%).

In consideration of duration, *no* yielded a CR of 14.3% and *yeah* of 12.5%. Figure 4.27 shows boxplots of the durational variation for *yeah* and *no* (along with the other speech units analysed) where it can be observed that occurrences of *no* tend to have longer duration than those of *yeah*. Also observable is that the majority of z-scores for all of the speech units fall above 0, indicating that these speech units are, for the most part, longer in duration than the speakers' syllables within their spontaneous 9-syllable utterance data (i.e., the data which the frequently occurring speech units were normalised against). It could be suspected that duration could be an influential factor with regards to determining the speaker discriminatory potential of these speech units. In fact, this trend is observable when comparing the durations of the speech units in Figure 4.27 to the LDA results for the dynamic contours of each speech unit (Table 4.3 for combined intensity, f_0 and duration). That is, the longest speech unit *erm* (vowel and nasal) had the highest classification rate of 81.3%, followed by just the vocalic portion of *erm* (CR = 40.6%), followed by *no* (CR = 36.3%), then *er* (CR = 31.3%), and lastly *yeah* (CR = 27.3%).

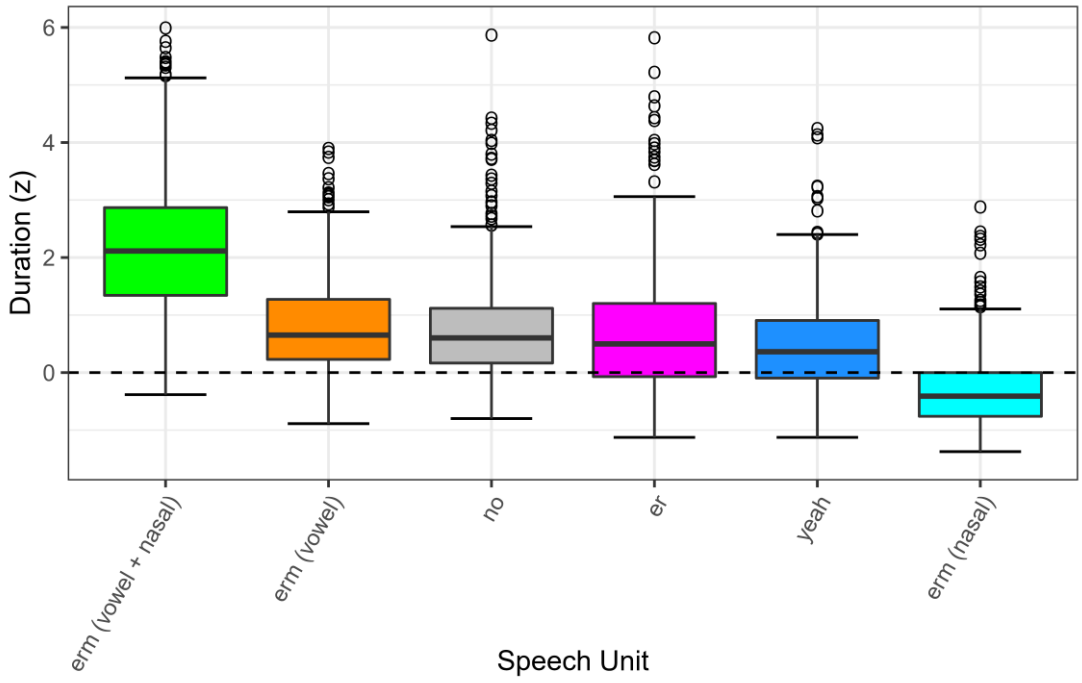


Figure 4.27. Boxplots showing the durational variability of the individual speech units analysed (ranked from highest to lowest). Higher z-scores correspond to longer duration.

4.4. Discussion

This chapter provided analysis as to the acoustic composition of four types of frequently occurring speech unit in relation to their rhythmic characteristics. Previous research has shown that the filled pauses *er* and *erm* as well as the monosyllabic responses *yeah* and *no* are relatively frequently occurring units within spontaneous speech. Furthermore, there have been a handful of studies which have examined the functions of these speech units, as well as where they are most likely to occur within an utterance and what other features which they are likely to cooccur with. For example, Gósy (2023) investigated the occurrences and durations of filled pauses in relation to words and silent pauses and found that the occurrence of filled pauses in various positions demonstrated a specific pattern. Although not concerned with speaker-specific patterns, the finding that these units exhibit specific patterning in relation to their positioning within spontaneous speech adds merit to the notion that such units could serve as useful anchor points from which spontaneous speech rhythm patterns can start to be measured. Similarly, Braun et al. (2023) found that verbal fillers, in which they assign *yeah* (*ja*) as the most relevant example, exhibit speaker-specific patterns with regards to their discourse position – namely, that when using verbal fillers, speakers do not pause before and after or pause in both instances. Again, such patterning promotes units such as these of being potentially fruitful with regards to acting as 'anchors' or 'control units' from which idiosyncratic speech rhythm patterns could be determined. As such, these four speech units were analysed in terms of their rhythmic characteristics, facilitating comparisons to be made between the spontaneous speech utterances analysed in the previous chapter.

The results showed that the speech units analysed were substantially more effective at discriminating between speakers than the 9-syllable spontaneous utterances analysed in Chapter 3. Dynamic measurements (across nine points of the speech units) performed better than static measurements, although the extent to which the dynamic measures outperformed the static measures varied across the different speech units. Of the three rhythmic parameters analysed, it was intensity which, on the whole, proved to be the most effective at distinguishing between speakers followed by f_0 and

then duration. Although it was shown that the combination of these features together harnessed the most speaker discriminatory potential.

As alluded to in Chapter 3, Section 3.4.1, considering intensity as a parameter for analysis within the forensic domain is problematic given the difficulty in obtaining accurate and consistent measurements due to confounding factors such as the distance between speaker and microphone, recording level, background noise, as well as speaker-dependent factors such as their chosen level of speaking. However, rather than altogether excluding intensity as a parameter which can be used for speaker discrimination within spontaneous speech, it is here proposed that measuring intensity over much shorter durations, such as individual speech units, where there is less potential for interference from those aforementioned problematic factors, could be of value to the forensic analyst.

For example, within a FVC case, it could be that a speaker exhibits a specific quality which involves marked patterns in intensity – such as vocal tremor – across certain segments of their speech or certain individual speech units. If the audio quality of the speech sample permitted, then these small segments or units of speech could be analysed in terms of their intensity patterns. Such patterns could be analysed acoustically and also visualised if evident in the spectrographic patterning. As a means of illustration, Figure 4.28 below shows a speaker exhibiting vocal tremor over a 2-second segment of speech.

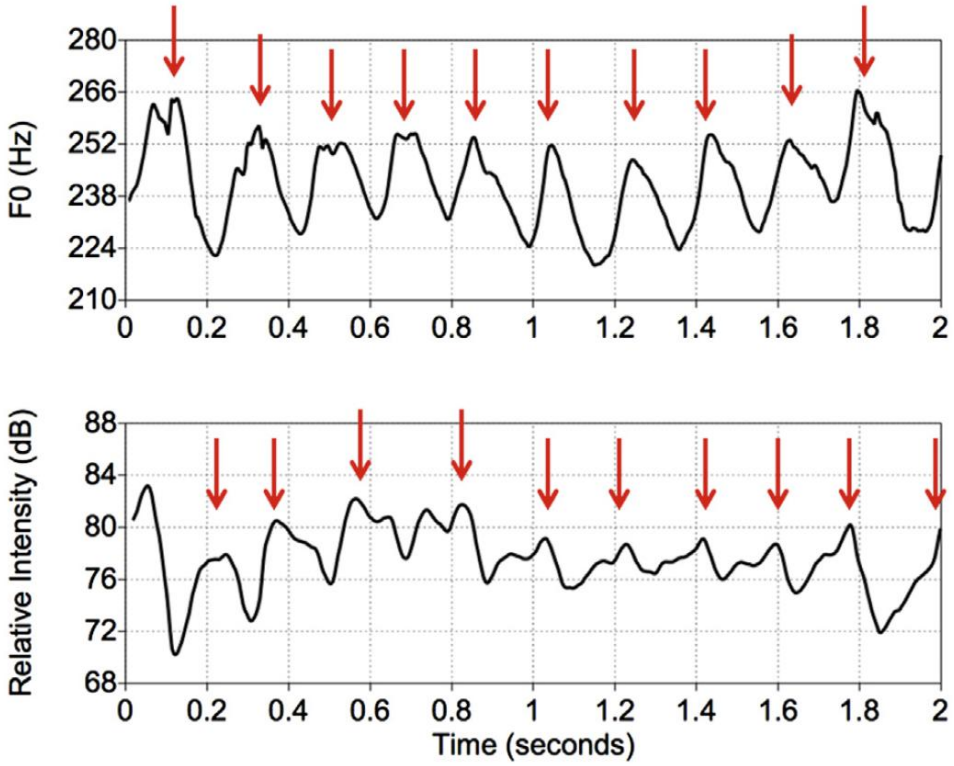


Figure 4.28. Plot of f_0 (top panel) and plot of relative intensity (lower panel) from a 2-second segment of sustained production of /a/ by a speaker with essential vocal tremor. Each red arrow marks the peak of a modulation cycle. Figure taken from Lester et al. (2013, p. 425).

Despite the illustration provided in Figure 4.28 originating from the clinical literature on vocal tremor and showing a pathological speaker with 'essential vocal tremor', it serves as a useful exemplar as to the unusual intensity (and f_0) patterning associated with this particular speech feature. If a non-pathological speaker were to exhibit vocal tremor across certain segments of their speech, this could be a useful feature for FVC as it is likely to be an unusual (and therefore distinctive) feature which many speakers would not exhibit. If such a feature were identified for a given speaker, comparisons of the acoustic intensity (and f_0) patterning could be made between the speech samples under analysis.

Of the four frequently occurring speech units analysed, it was the filled pause *erm* which showed the most speaker-specificity. In the following subsections, the results presented in Section 4.3 are brought together and discussed. This discussion focusses

on the relative speaker-specificity and performance of each speech unit and parameter with a view to highlighting which harness the greatest forensic potential.

4.4.1. Filled pauses: *er* and *erm*

Of the speech units examined, the filled pause *erm* (when considered as a whole) presented as having the most speaker discriminatory power, with measures of intensity and f_0 yielding promising speaker classification rates. When measures of intensity and f_0 are considered alongside each other, speaker classification rates are greatly improved with the interaction between the two parameters for this filled pause type being highly significant ($p < 0.0001$). The marked decrease in intensity (evidenced by *most* speakers) cued by the transition from the vowel portion to the nasal portion is offered as an explanation as to why dynamic intensity measurements perform well here, with it being suggested that there will be between-speaker variation in terms of the extent of this decrease in intensity as a result of both articulatory (e.g., when the /ə/ to /m/ transition occurs) and physiological reasons (e.g., size of a speaker's nasal cavity). The vocalic to nasal transition of *erm* is also put forward as one explanation as to why dynamic measurements of f_0, too, return promising classification rates given the temporary decrease in f_0 as a result of this transition, and that there appears to be between-speaker variation in the extent of this temporary f_0 drop-off. Also, it was shown that dynamic f_0 measurements are useful for distinguishing speakers who may at times mark this filled pause with initial high pitch, with this being an apparently idiosyncratic prosodic feature for some speakers, especially when this filled pause was the initial speech unit when responding to a question (i.e., marked prosody of contemplation). It was acknowledged in Section 4.3.2.1, for the analysis of *erm*, that the markedly promising classification rate for the dynamics of this speech unit (CR = 81.3%) could be subject to scrutiny. Here, the speech unit is considered in terms of its combined intensity, f_0 and duration, meaning that a given speech unit is represented by 19 values (nine intensity values (+10% steps across the unit), nine f_0 values (+10% steps across the unit), and one duration value (the duration of the unit)). As such, it is recognised that any overly encouraging results could be a product of factoring too many variables into the discriminant analysis. In the context of the present work,

however, consideration should also be given to the relatively small dataset being analysed (for *erm*: 12 speakers, 40 tokens per speaker) when weighing up the classification rates obtained.

The coarticulatory effects of the nasal potion of *erm* can also be seen to have an influence on the vocalic portion when it is analysed as a separate entity. There appears to be greater potential for intensity and f_0 fluctuations in the vocalic portion of *erm* than there is for *er*, with this resulting in greater between-speaker variation and therefore the finding that dynamic measurements of the vocalic portion of *erm* outperform *er*.

Where dynamic measures of intensity and f_0 had the most speaker-discriminatory power when considering *erm* as a whole, measurements of F_1-F_3 were more effective when considering just the vocalic portion of *erm* as well as *er*. Of the two, dynamic measurements of the vowel of *erm* proved more useful than for *er* with this finding again being supposed to be due to the coarticulatory effects of the following nasal in *erm*. The results of the present study in relation to measurements of F_1-F_3 correspond with the results obtained by Hughes, Wood and Foulkes (2016), with *erm* performing better than *er*, and also in that the addition of duration improves the speaker-discriminatory potential.

In consideration of duration, this parameter was generally the least effective at discriminating between speakers when taken as a measure on its own (with only measures of f_0 occasionally performing marginally worse). However, its inclusion alongside measures of intensity and f_0 (and F_1-F_3) almost categorically aids in bolstering the discriminatory potential for both dynamic measurements and static measurements, although the extent to which classification rates are improved are only slight in comparison to the advancements when combining intensity and f_0 measures.

Of the four speech units analysed in the present chapter, the filled pauses *er* and *erm* have been subjected to much more research in terms of their acoustics than the monosyllabic responses *yeah* and *no*. Where this research has been focussed on the speaker-specificity of these filled pauses, it has been for the most part centred on vowel formant measurements (e.g., Foulkes et al., 2004; Hughes et al., 2016; Hughes et al., 2023) and to a lesser extent on measuring f_0 (e.g., Braun & Rosin, 2015; Tschäpe

et al. 2005), with some studies analysing accounting for both vowel formants and f_0 simultaneously (e.g., De Boer et al., 2022; De Boer & Heeren, 2020). Where the analysis of vowel formants has incorporated dynamic measurements across the duration of these speech units, f_0 research has for the most part been confounded to static measurements such as the mean or midpoint. The analysis presented in the current chapter investigated both static and dynamic measurements of f_0 as well as intensity – the latter here being a comparatively understudied parameter in the context of the acoustics of filled pauses and as well as its capacity as a speaker discriminatory feature. This initial exploration of these speech units in relation to these parameters has provided further support of the usefulness of analysing the acoustics of filled pauses for forensic purposes and has also shed light on the discriminatory potential of accounting for intensity (to a lesser extent) f_0 dynamics when conducting such acoustic analysis.

4.4.2. Common monosyllabic responses: *yeah* and *no*

Analysis of the monosyllabic responses *yeah* and *no*, showed that dynamic measurements of *no* outperformed those of *yeah*, with the combination of intensity, f_0 and duration showing the most speaker discriminatory potential. Interestingly, in this respect, *no* also outperformed *er* and it could be that the transition from the nasal to vowel in *no* and the associated increase in intensity (and possible interrelations with f_0) allow for greater between-speaker variation and thus increase its discriminatory power. As shown in Figure 4.27, the duration of *no* is generally longer than that of both *yeah* and *er*, and it may be that speech units which are durationally longer allow for more between-speaker variation to manifest and be subsequently captured by dynamic measurements. Also, it is worth noting that the vowel within *no* may be more susceptible to greater between-speaker variation than that of the vowel in *yeah* for this group of speakers. Where some speakers might have a more diphthongal realisation for this vowel (e.g., /əʊ/), others might have a more monophthongal realisation (e.g., /oː/ (or more likely, given the age group of the speakers, a fronted variant such as /ɵː/)) owing to the participants being speakers of Bradford English (e.g., Hughes et al., 2013; Petyt, 1985; Watt & Tillotson, 2001). Thus, it is probable that such realisational

differences could lead to greater between-speaker variation being evidenced with respect to the rhythmic characteristics of the vocalic section of *no*. For example, the monophthongal realisations, as indicated by the diacritics included above, could be longer in duration than more diphthongal realisations, whereas more diphthongal realisations could perhaps evidence greater fluctuations in intensity due to greater articulatory movement. It is more probable, in fact, that it is a combination of all of these aforementioned factors that result in *no* showing greater speaker discriminatory potential than *yeah*.

An alternative interpretation in explaining why *no* outperforms *yeah* could be a consequence of *yeah* being arguably a more polyfunctional word than its notional contrary *no*. In the research literature (mostly pertaining to discourse/conversation analysis), *yeah* has been shown to have many functions within dialogue such as acting as a backchannel (an indicator that the speaker is being listened to and may carry on with the conversation), an assessment item (evaluating something that was said previously), or a marker of speaker incipiency (an indicator of the speaker taking the conversational floor; see e.g., Drummond & Hopper, 1993; Gardner, 1998; Jefferson, 1984).

A small body of research which has investigated the prosodic makeup of *yeah* in relation to its conversational function has shown that these different functions carry with them different prosodic inflections in relation to intensity, f_0 and duration (e.g., Benus et al., 2007; Grivičić & Nilep, 2004; Trouvain & Truong, 2012; Truong & Heylen, 2010). As a result, it could be expected that speakers will exhibit greater within-speaker variation in their use of *yeah* as opposed to their use of *no*, and therefore less speaker-specificity is likely to be evidenced. Although there are a handful of studies which have looked into the potential multifunctionality of *no* (e.g., Jefferson, 2002; Lee-Goldman, 2011), this work has not looked into the associated acoustics. Thus, the explanation offered here regarding *no* outperforming *yeah* is made with a degree of tentativeness.

Overall, it is somewhat challenging to position the results obtained for *yeah* and *no* within the forensic phonetics field given the current lack of research which has been dedicated to these specific units. One recent study conducted by Gibb-Reid et al.

(2022) assessed the vowel formant dynamics of *yeah* in terms of F_1 and F_2 trajectories and found that the formant trajectories varied based on the function of the word as well as its positioning with respect to pauses (mirroring the findings of the aforementioned studies relating to the functional prosodic variation of *yeah*). Moreover, their results showed an indication that *yeah* possessed distinctive formant trajectories across speakers as well as exhibiting low within-speaker variability – both of which are desirable for the forensic analyst. They use these findings to tentatively suggest that word-specific variation is worthy of further investigation with respect to the application to forensic voice comparison tasks.

Even more recently, Braun et al. (2023) investigated the speaker-specificity of a range of different speech disfluencies which included a group of so-called "verbal fillers" for which *yeah* (or rather its German counterpart *ja*) was noted as being the most frequent and relevant example (along with *und*). They examined these verbal fillers in terms of their frequency and positioning within an utterance, as well as in relation to their f_0 relative to their neighbouring context. They highlight how these verbal fillers are overlooked in relation to other disfluency phenomena such as filled pauses, and advocate that they receive more attention in future research given their findings that these items contribute towards speakers' individual disfluency patterns and bolster the speaker discriminatory potential of disfluency behaviour analysis.

Where a good deal of previous research has focussed on static acoustic measurements such as the midpoint of a segment or the mean value of a segment, Gibb-Reid et al.'s findings provide merit for examining the dynamics of specific words. The analysis conducted in the present chapter has examined the dynamic contours of *yeah* and *no* in relation to measurements of f_0 and intensity for the purpose of speaker discrimination, offering an initial exploration into their acoustic makeup and forensic potential. Although, in general, these two units were shown not to possess as much speaker discriminatory power as filled pauses *er* and *erm*, the results obtained offer support to the suggestion that the acoustic analysis of word-specific variation has potential within the forensic domain.

4.5. Chapter summary

This chapter has demonstrated that the rhythmic characteristics of four types of, so-called, "frequently occurring speech units" carry reasonable amounts of speaker-specific information. It was hypothesised that the speech units selected for analysis could potentially be useful markers of an individual's speech rhythm and therefore conceivably a fruitful way to capture speech idiosyncratic rhythm patterns. As such, the rhythmic properties of these speech units were measured, and a novel normalisation method was employed which allowed the rhythmic characteristics of these units to be captured relative to the speakers' spontaneous speech patterns. Results from all four of the speech units analysed were much stronger than the results for the spontaneous utterances. Dynamic measurements of these speech units, for the most part, produced the best results, and it was shown that combining measures of intensity, f_0 and duration resulted in speaker discriminatory power being at its optimal.

CHAPTER 5

Perception Experiments

5.1. Introduction

Where the previous two chapters reported on production experiments which assessed the speaker discriminatory potential of acoustic measurements of speech rhythm, the present chapter examines the contribution of holistic assessments of speech rhythm grounded in perception. This chapter therefore presents the results obtained from two speech rhythm perception experiments. The prior two chapters (along with previous research) have indicated that there is some value in pursuing rhythm for speaker identification, however it is strongly suspected that some rhythmic information will likely be missed using these acoustic methods and it is possible that perception could be used as a tool to draw out further relevant rhythmic information. Furthermore, when turning to the application of such research to real-life forensic casework, comparing the acoustics of speakers' rhythm patterns is reliant upon 'enough' adequate speech data being available to the forensic analyst – a privilege that cannot be guaranteed within the forensic context. The perception experiments presented in this chapter can therefore be seen as a natural next step as they aim to strengthen the auditory analytical potential of rhythm as a speech analysis feature.

The first of these experiments was the Pilot Study which assessed naïve (non-expert) listeners whilst also serving to test the methodological design of the experiment. This Pilot study was also carried out in order to determine the level of difficulty of the tasks which listeners had to complete, along with testing the comprehensibility of the instructions provided to listeners and to get an understanding as to the length of time

the experiment would take to complete. This was followed up by the Main Experiment which was an extended adaptation of the Pilot Study. Where the Pilot Study was made up of eight tasks in each of the three sections, the Main Experiment consisted of five tasks in Section One and 15 tasks in Section Two and Section Three. In both experiments, listeners were invited to discriminate between speakers and evaluate the similarity of speech samples based on primarily rhythmic attributes of speech. Listeners were presented with original (natural) speech samples along with samples which had been subjected to delexicalisation, whereby syllables were represented by schwa-like tones, creating samples which foregrounded rhythmic characteristics (see Section 5.2.3 for the delexicalisation procedure followed). These experiments consisted of three sections. In Section One and Section Two, listeners were required to make a binary decision as to which delexicalised samples contained the same speaker as the original (non-delexicalised) samples whilst, for Section Two, also providing qualitative feedback. In Section Three, listeners had to rate the similarity of pairs of delexicalised speech samples. These experiments therefore looked to determine the extent to which expert and non-expert listeners were able to discriminate between speakers based on speech rhythm. The experiments were structured in such a way so that the level of difficulty increased for each section. Section One of the experiment was crafted to act as a preparatory training stage, where participants could gain exposure to the sound of delexicalised speech samples. This phase was also intended to assist participants in honing their ability to concentrate on rhythmic cues when identifying speakers. In this Section, listeners were provided with an original sample and two delexicalised samples, one of which was a direct reproduction of the original speech segment. This design allowed for straightforward comparisons, enabling listeners to match syllable patterns from the original sample with those in the correct delexicalised sample.

In Section Two, the listeners could no longer adopt a strategy of correlating the direct syllabic patterns from the original message with the delexicalised samples. They were instead tasked with discerning the rhythmic patterns of the speaker in the original speech and determining which of the two delexicalised samples – none of which matched the original speech segment – contained the same speaker by using the rhythmic characteristics present in the delexicalised samples.

In Section Three, only delexicalised speech samples were provided to listeners for their similarity judgments, with no original voice samples available for reference. This design ensured that participants could only utilise the rhythmic features they identified in their similarity assessments. Furthermore, obtaining qualitative feedback from expert listeners was essential in fulfilling an overarching goal of the present work relating to its application to forensic voice comparison casework. Given that within the auditory-phonetic and acoustic approach to forensic voice comparison there is currently no structured framework analysts can use to effectively account for speakers' speech rhythm patterns, this qualitative feedback could provide the basis for the development of meaningful descriptors of speech rhythm which would feed into a perceptual rhythm framework for forensic speech analysis (see Chapter 6).

The present chapter therefore looks to answer the question as to whether certain groups of listeners (e.g., expert vs. non-expert / forensic expertise vs. no forensic expertise) perform better at making correct speaker identification assessments when presented with primarily the rhythmic attributes of speech. Qualitative feedback is also obtained from listener groups in order to ascertain which specific features from the speech samples were being relied upon when making their identification assessments.

There has only been a limited number of forensically-motivated perception studies which have made use of delexicalised speech samples, with this research being reviewed in Chapter 2, Section 2.5.4. Although these studies were not exclusively focussed on speech rhythm, the results obtained showed that listeners demonstrate the capacity to recognise and differentiate among various foreign-accented speech patterns, even when these patterns have been subjected to degradation through diverse signal manipulation methods. This research supports the design of the perception experiments described in the present chapter, wherein listeners are tasked with identifying speakers through delexicalised speech material that foregrounds rhythmic elements.

Additional perception research within the forensic speech science domain has primarily centred on tasks that require listeners to assess the (dis)similarity of typically brief (non-delexicalised) speech samples. For example, Bartle and Dellwo (2015)

investigated the capacity of both expert and non-expert listeners to distinguish between speakers through short utterances presented in both voiced and whispered forms. Similarly, McDougall (2013) tasked naïve listeners with evaluating the similarity of pairs of short speech samples using a rating scale from 1 (very similar) to 9 (very different) with the ultimate goal of assessing the degree of perceived similarity amongst a group of voices for potential inclusion in a voice parade.

Outside of the forensic field, studies which have accounted for speech rhythm have often done so alongside other parameters. For example, Van Dommelen (1987) investigated the influence of speech rhythm, intonation, and pitch on the identification of paired speakers, aiming to determine the factors that contribute to listeners' capacity to differentiate between speakers when the natural quality of the voice is absent from the speech signal. Results from this study showed that listeners employed all three parameters to different extents, depending on the speech sample presented, thereby indicating that the importance of these parameters for differentiating speakers is not absolute but rather varies with the speaker.

In consideration of previous perception studies such as those described above, it is hypothesised that expert listeners, that is, listeners who have received formal linguistic/phonetic training, will perform better than non-expert listeners with regards to making correct speaker identification assessments. Of the expert listeners, it is further predicated that those who have expertise in forensic phonetics will likely perform to a higher standard than those experts who do not have forensic experience. With regards to the qualitative feedback generated, it is predicted that listeners will make use of a variety of different auditory features when making their identification assessments. For example, speaking / articulation rate, pausing behaviour, intonation patterns, and disfluency behaviour are all features which are suspected to be commented on by listeners – albeit to greater or lesser extents.

The results obtained from the experiments in the present chapter will for the most part be evaluated using a qualitative approach – that is, they will be largely descriptive in nature. This is due to the fact that the experiments were not designed to generate large amounts of numerical data for which statistical testing could be carried out.

The composition of this chapter is as follows. Section 5.2 provides the methodological detail of the experiments in relation to the participants (listeners) and the speakers involved, and the procedures involved in data preparation, creating the delexicalised speech samples and the design of the online experiments. Section 5.3 then presents the results of the experiments. Firstly, the results obtained from the initial Pilot Study are reported before a more detailed analysis of the results from the Main Experiment are described. The chapter is subsequently brought to a close with discussion relating to the findings from both experiments and an overall summary of the results.

5.2. Methodology

The following subsections detail the methodology of the chapter. Firstly, in Section 5.2.1, information is provided relating to the subjects who participated in the perception experiments, before Section 5.2.2 provides information for the speakers who contributed to the speech sample data. Following this, Section 5.2.3 details how the data were prepared and edited, and the procedures followed to create the delexicalised speech samples. Section 5.2.4 then explains the experimental design and the nature of the tasks which participants were required to complete. Lastly, Section 5.2.5 describes the statistical analysis carried out for Section Three of the Main Experiment.

5.2.1. Participants

5.2.1.1. Pilot Study

The Pilot Study was made up of 12 participants who were all recruited on the basis that they were naïve (non-expert) listeners who had no formal training in linguistics. All of the participants were native speakers of English. The group was made up of seven male listeners and five female listeners whose ages ranged between 22 and 62 years. None of the participants reported having any significant hearing impairment. Given the purpose and the exploratory nature of the Pilot Study, no additional demographic information was collected from the subjects involved.

5.2.1.2. Main Experiment

The Main Experiment was made up of 45 participants in total: 32 expert listeners and 13 non-expert listeners. The expert listeners were recruited directly (primarily by email) and were deemed 'expert' in that they have all received formal linguistic training within the field of phonetics and/or forensic phonetics. The non-expert listeners were selected on the basis that they had no formal training in the field of linguistics.

Prior to completing the experiment, both the expert and non-expert participants were asked to identify their level of education and were required to state whether they were a native speaker of English. If they were not a native English speaker, they were asked to rate their competence of English on a 4-point scale ranging from ‘low-level competence’ to ‘native-level competence’. The expert listeners were also asked to provide information pertaining to their level of experience/expertise within phonetics and/or forensic phonetics. Based on the responses received, the participants were subsequently categorised into the following six groups:

- Forensic Caseworkers (seven)
- Forensic Phonetics Researchers (six)
- Forensic Phonetics Research Students (nine)
- Phonetics Researchers (five)
- Phonetics Research Students (five)
- Non-expert (thirteen)

Of the 32 expert listeners, 11 were male and 21 were female, and for the non-expert listeners five were male and eight were female. None of the subjects reported having any significant hearing impairment. No further demographic information (e.g., age) was collected. All participants received a £30.00 Amazon Gift Card for the time they invested in the experiment. Table 5.1 and Table 5.2 detail, respectively, how the expert and non-expert listeners responded to the pre-experiment questionnaire.

Table 5.1. Pre-experiment questionnaire responses from the group of expert listeners.

Expert Listeners		
Question	**Response**	**Number of responses**
Level of education (please tick all that apply to you)	I have completed a bachelors degree	16
	I have completed a masters degree	18
	I am currently a PhD student	12
	I have completed a PhD	15
	Other (please specify)	1
Expertise / Experience (please tick all that apply to you)	I currently carry out forensic casework full-time (my current day job)	5
	I have carried out forensic casework in the past (used to be my day job)	2
	I occasionally carry out forensic casework	2
	I am researcher in forensic phonetics	14
	I am currently doing a PhD in forensic phonetics	7
	I am an analyst in a government laboratory	1
	I am an academic phonetician/sociophonetician	15
	I am currently doing a PhD in phonetics/sociophonetics	8
	I am a researcher in phonetics/sociophonetics	17
	Any other relevant information / credentials (please specify)	3
Are you a native speaker of English?	Yes	21
	No – low-level competence	0
	No – moderately competent	0
	No – Highly competent	6
	No – Native-level competence	5

Table 5.2. Pre-experiment questionnaire responses from the group of non-expert listeners.

Non-expert Listeners		
Question	**Response**	**Number of responses**
Level of education (please tick all that apply to you)	Secondary school qualification(s)	4
	College qualification(s)	2
	I am currently studying a bachelors degree at university	5
	I have completed a bachelors degree	4
	I am currently studying a masters degree at university	3
	I have completed a masters degree	2
	I am currently a PhD student	2
	I have completed a PhD	0
	Other (please specify)	0
Are you a native speaker of English?	Yes	10
	No – low-level competence	0
	No – moderately competent	0
	No – Highly competent	1
	No – Native-level competence	2

5.2.2. Speech material

The data used as stimuli in the present chapter were obtained from the same group of 20 speakers from the WYRED corpus as detailed in Chapter 3, Section 3.2.1. The decision to use the same 20 speakers as the production experiments (Chapters 3 and 4) was made so that, where relevant, comparisons could be made across the production and perception experiments (e.g., if speaker X showed speaker-specific patterns in his spontaneous speech rhythm patterns, is this speaker more easily identifiable within the perception tasks?). Rather than using the mock police interview data used in Chapters 3 and 4, the data for the perception experiments were obtained from the answerphone message task. In this task, the speaker was asked to leave an answerphone message in a time-pressured situation and given a rough guide on the information they had to convey. These answerphone messages featured no interlocutor and were generally around two minutes in length. Studio quality recordings were used as opposed to telephone quality recordings which were also afforded. The decision to use the answerphone message task for the perception experiments (as opposed to the mock interview task used in the production experiments) was made in consideration of there being no interlocuter present. In the absence of an interlocutor, speakers lack the opportunity to adjust their speech to accommodate another individual. Consequently, the rhythm patterns of their speech are likely to reflect their personal style more authentically. Furthermore, without a conversational partner, the chances of significant alterations in a speaker's rhythm patterns diminish, as there are no imposed shifts in topic or emotion. This results in a greater consistency of rhythmic patterns compared to those observed in dialogue.

Table 5.3 details which speakers contributed data to the perception experiments and the production experiments.

Table 5.3. Information relating to which speakers contributed to the data across the perception experiments and the production experiments.

Speaker	Production Experiments		Perception Experiments		
	Chapter 3	Chapter 4	Section 1	Section 2	Section 3
WY008	✓	✓	✓	✓	✓
WY023	✓	✓	✓	✓	✓
WY030	✓	✓	✓	✓	✓
WY031	✓	✓	✓	✓	✓
WY033	✓	✓	✓	✓	✓
WY040	✓	✓	×	✓	✓
WY042	✓	✓	×	✓	✓
WY043	✓	✓	×	✓	✓
WY067	✓	✓	×	✓	✓
WY069	✓	✓	✓	✓	✓
WY072	✓	✓	×	✓	✓
WY080	✓	✓	×	✓	✓
WY109	✓	✓	×	✓	✓
WY123	✓	✓	×	✓	✓
WY161	✓	✓	×	✓	✓
WY162	✓	✓	×	×	×
WY167	✓	✓	×	✓	✓
WY170	✓	×	×	×	×
WY171	✓	✓	×	×	✓
WY177	✓	✓	×	×	✓

5.2.3. Data preparation and delexicalisation

Initial editing of the voicemail messages was conducted within Praat and involved extracting the initial 30 seconds and the final 30 seconds from each recording. Delexicalised samples of these edited stretches of speech were then created. The purpose of this delexicalisation was to foreground the rhythmic attributes of the original speech sample whilst removing the lexical content and all aspects of voice quality. As such, a number of different delexicalisation methods were tested in order to determine which would serve the purpose of this perception experiment most adequately. The delexicalisation methods trialled included creating the following:

- **Sasasa-speech** – where all consonants are replaced with /s/ and all vowels are replaced with /a/. Sasasa-speech preserves only characteristics of syllabic rhythm and intonation (e.g., Ramus et al., 1999; Ramus & Mehler, 1999).

- **Noise vocoded speech** – where amplitude envelopes are extracted from several frequency bands and are used to modulate white noise in these frequency regions. This results in the absence of voicing cues as well as the extreme degradation (or absence) of cues to segment durations. In perceptual terms, noise vocoded speech can be described as a succession of syllable beats in the form of white noise pulses (e.g., Kolly & Dellwo, 2014; Shannon et al., 1995).

- **Spectrally rotated speech (with low-pass filtering)** – where the spectrum of low-pass filtered speech is inverted around a centre frequency. Spectral shape and its dynamics are completely altered, rendering speech virtually unintelligible initially. However, prosodic attributes such as intonation, rhythm, and contrasts in periodicity and aperiodicity are largely unaffected (e.g., Blesser, 1972; Green et al., 2013).

In order to determine what would be the most appropriate delexicalisation method for the present study, the delexicalised speech samples (created using the methods described above) were presented to two trained forensic phoneticians, along with a delexicalised sample created manually by the present author (see below for this delexicalisation method). The two trained forensic phoneticians were aware of the nature of the tasks for the perception experiments. In consideration of the feedback obtained from the two trained forensic phoneticians, it was decided that to serve the purpose of the present experiment best, and to ensure consistency, the delexicalisation method devised by the present author should be used. The reasoning for this was twofold. Firstly, the favoured delexicalisation method was deemed the most efficient at conveying the rhythmic characteristics which have been the focus of the production experiments within the present thesis - that is, intensity, f_0 and duration. Secondly, the three other methods trialled were deemed to be unsuited for the present experiment for different reasons. For example, in the case of spectrally rotated speech, this method was deemed to be too ‘distracting’ owing to the fact that, although the original speech is virtually unintelligible, the inverted frequencies were perceptually more prominent than other rhythmic characteristics. This meant that there was potential (at least some)

listeners may direct their focus (intentionally or unintentionally) towards trying to disentangle the delexicalised samples' linguistic content rather than focussing solely on the rhythmic patterns they could perceive. In relation to the noise vocoded speech method and also (albeit to a lesser extent) the sasasa-speech method, it was perceptible that some rhythmic characteristics were being misrepresented. For example, when comparing the delexicalised samples creating using these two methods to the original speech samples, it was sometimes the case that syllable durations were being represented as shorter than they should be (i.e., shorter than in the original speech samples). Additionally, with regards to the sasasa-speech method, there were at times issues with the quality of the /s/ segments which again resulted in some syllables being misrepresented with regards to their duration. As such, delexicalised versions of the original speech samples were created within Praat by the present author. The resulting delexicalised samples rendered syllables from the original samples as being represented by schwa-like tones. The procedure used to create the delexicalised samples is as follows (n.b., many of the descriptions of the processes/functions found below are taken directly – that is, are verbatim – from the Praat manual, accessed through the Praat computer program (Boersma & Weenink (2020). The underlying algorithms for the different processes/functions are all detailed in Boersma (1993)):

1) The Sound Object (i.e., an original speech sample) was selected, from which a Pitch Object was created using the 'Sound: To Pitch' function.

2) The resulting Pitch Object represents periodicity candidates as a function of time. It is sampled into a number of frames centred around equally spaced times. The algorithm used performs an acoustic periodicity detection on the basis of an accurate autocorrelation method. This method is described in full in Boersma (1993). On creation of the Pitch Object, the following settings were applied:

- Time step (s): 0.015 (= auto)
- Pitch floor (Hz): 50.0
- Pitch ceiling (Hz): 300.0
- Max. number of candidates: 15
- Very accurate: off

- Silence threshold: 0.03
- Voicing threshold: 0.45
- Octave cost: 0.01
- Octave-jump cost: 0.35
- Voiced / unvoiced cost: 0.14

3) The Pitch Object was edited using the 'View & Edit' function to correct any pitch tracking errors (e.g., tracking errors caused by octave jumps, instances of creak, etc.). The 'View & Edit' function puts an editor window on the screen, which shows the contents of the Pitch Object. This window allows the Pitch Object to be viewed and modified. Within this window, the following features are observable:

- Digits between 0 and 9 scattered all over the drawing area. Their locations represent the pitch ***candidates***, of which there are several for every time frame. The digits themselves represent the goodness of a candidate, multiplied by ten. For instance, if you see a "9" at the location (1.23 seconds, 189 hertz), this means that in the time frame at 1.23 seconds, there is a pitch candidate with a value of 189 hertz, and its goodness is 0.9. The number 0.9 is the relative height of an autocorrelation peak.

- A ***path*** of red disks. These disks represent the best path through the candidates, i.e. our best guess at what the pitch contour is. The path will usually have been determined by the *path finder*, which was called by the pitch-extraction algorithm, and you can change the path manually. The path finder takes into account the goodness of each candidate, the intensity of the sound in the frame, voiced-unvoiced transitions, and frequency jumps. It also determines whether each frame is voiced or unvoiced.

- A ***voicelessness bar*** at the bottom of the drawing area. If there is no suitable pitch candidate in a frame, the frame is considered voiceless, which is shown as a blue rectangle in the voicelessness bar.

- A line of digits between 0 and 9 along the top. These represent the relative intensity of the sound in each frame.

The editing process involved, where relevant, manually correcting *candidates* which may have been incorrect owing to instances of octave jumping, creak, etc. On occasion, it was also necessary to correct whether a frame was deemed as voiced or

voiceless, with this being accomplished by clicking within the *voicelessness bar*. Both of these processes were accomplished by going back and forth from the PitchEditor window to the Sound Object window (i.e., the visual representation (spectrogram and waveform) of the original sample) and making corrections based on both perceptual and visual information.

4) A PitchTier was then created from this manually corrected Pitch Object using the 'Convert: Down to PitchTier' function. This PitchTier object represents a time-stamped pitch contour that contains a number of (*time*, *pitch*) points, without voiced/unvoiced information.

5) The Sound Object was then selected again, from which a Manipulation Object was created using the 'Manipulate: To Manipulation' function. The creation of this Manipulation Object prompts the following steps to be performed (n.b., the following description is taken from the Praat manual, accessed through the Praat computer program (Boersma & Weenink (2020)):

- A pitch analysis is performed on the original sound, with the method of Sound: To Pitch.... This uses the time step, pitch floor, and pitch ceiling parameters.

- The information of the resulting pitch contour (frequency and voiced/unvoiced decisions) is used to posit glottal pulses where the original sound contains much energy. The method is the same as in Sound & Pitch: To PointProcess (cc).

- The pitch contour is converted to a pitch tier with many points (targets), with the method of Pitch: To PitchTier.

- An empty DurationTier is created.

6) On creation of this Manipulation object, the following settings were applied:

- Time step (s): 0.015 (= auto)
- Pitch floor (Hz): 50.0
- Pitch ceiling (Hz): 300.0

7) From the Manipulation Object, a PointProcess Object was created using the 'Extract pulses' function. This PointProcess Object represents a *point process*, which

is a sequence of *points* t_i in time, defined on a domain [t_{min}, t_{max}]. The index *i* runs from 1 to the number of points. The points are sorted by time, i.e. $t_{i+1} > t_i$.

8) The PointProcess Object and Sound Object were then selected together allowing for the 'View & Edit' function to be used with this subsequently opening a PointEditor window. From the PointEditor window, the *points* (which are the glottal pulses derived from the Praat pitch detection algorithms) are represented by vertical blue lines which run through the waveform of the original sound sample. Here it is possible to edit (add and remove) these *points.*

9) *The* points were edited so that all syllables were attested and represented accurately (e.g., in terms of duration). This editing process involved removing *points* in order to allow for audibly discernible intervals between syllables to be perceived by the listener. Where *points* had to be removed for this purpose, it was decided that two *points* should be removed, with this being consistent across the editing of all speech samples. On occasion*, points* also had to be added in order for syllable duration to be accurately represented and also to ensure unvoiced syllables and unvoiced speech units were audibly present. Figure 5.1 and Figure 5.2 provide an illustration of the *point* editing process.

10) The edited PointProcess Object and Manipulation Object were then selected together allowing for the 'Replace pulses' function to be used. This function replaced the vocal-pulse information (i.e., the points) in the selected Manipulation Object with the selected PointProcess Object (containing the edited *points*).

11) The edited PitchTier and Manipulation Object were then selected together allowing for the 'Replace pitch tier' function to be used. This function replaced the original, unedited pitch (from the original sound sample) with the corrected PitchTier.

12) The Manipulation Object was then selected and the 'View & Edit' function was used which subsequently opened the ManipulationEditor window.

13) From the ManipulationEditor window, the 'Synth' tab was selected from which the 'Pulses – Pitch (hum)' resynthesis method was selected from the dropdown menu. This function converted all of the *points* to a Sound. To do so, a pulse is generated at every point and this pulse is filtered at the Nyquist frequency of the resulting Sound

by converting it into a sampled *sinc* function. This sound is then run through a sequence of second-order filters that represent five formants.

14) From the ManipulationEditor window, the 'File' tab was then selected from which the 'Publish resynthesis' function selected. This function created a new Sound Object ('Sound fromManipulationEditor') which appeared in the main Praat Objects window. This new Sound Object is the delexicalised speech sample.

15) This new Sound Object was then saved as a .wav file.

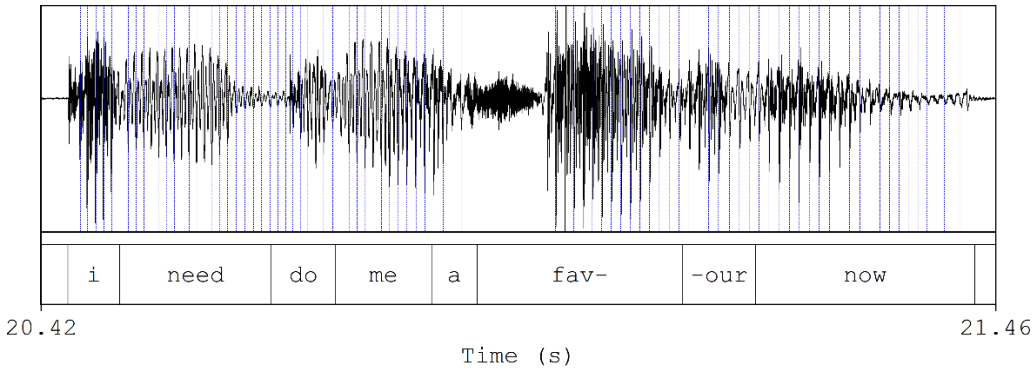


Figure 5.1. Waveform with *unedited* period markers (i.e., *points*; vertical blue lines) from the utterance 'I need do me a favour now'. If the *points* were not edited, it would be perceived as two long schwa-like tones (i.e., only two syllables rather than eight) following the delexicalisation procedure.

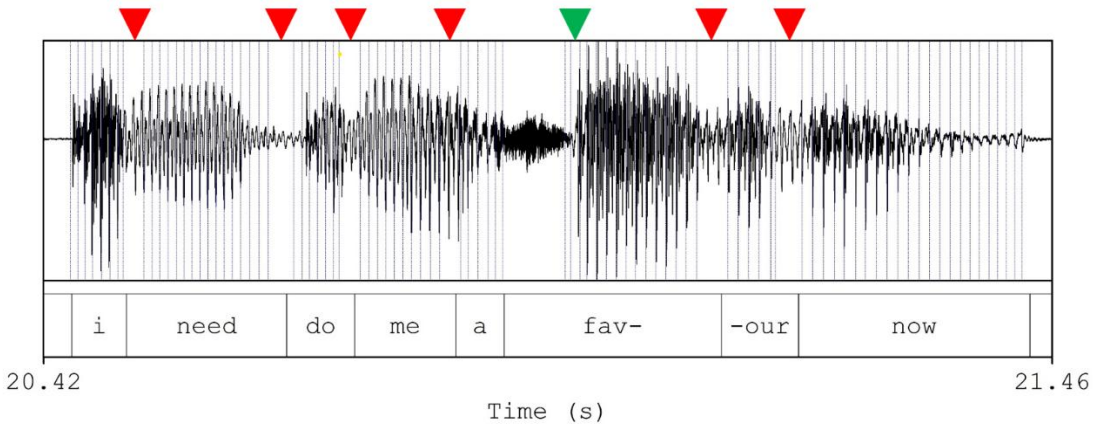


Figure 5.2. Waveform with *edited* period markers (i.e., *points*; vertical blue lines) from the utterance 'I need do me a favour now'. Red arrowheads indicate where *points* have been removed and green arrowheads indicate where *points* have been added.

This delexicalised utterance is now perceived as eight schwa-like tones, corresponding to the syllable structure of the utterance.

So that listeners would not be able to identify speakers solely based on pitch, and to ensure that listeners had to focus on different rhythmic characteristics when making their assessments, it was decided that the delexicalised samples should be normalised. As such an average pitch of 100 Hz (and standard deviation of 20 Hz (average pitch multiplied by 0.2)) was applied to all of the delexicalised samples using a Praat script which consisted of the following processes:

1) The Sound Object (the delexicalised sample) was selected from which the total duration was obtained.

2) A Manipulation Object was created using the 'Manipulate: To Manipulation' function with the following settings being applied:

- Time step (s): 0.015 (= auto)
- Pitch floor (Hz): 50.0
- Pitch ceiling (Hz): 300.0

3) From the Manipulation Object, the PitchTier was extracted using the 'Extract pitch tier' function.

4) From the PitchTier, the mean pitch and the standard deviation were obtained using the 'Get mean (points)' and 'Get standard deviation (points)' functions.

5) From the PitchTier, the number of pitch points was obtained using the 'Get number of points' function.

6) For each of the points, the value was obtained using the 'Get value at index' function.

7) The mean (obtained in step 4) was then subtracted from each value, with the resulting value then being divided by the standard deviation (obtained in step 4).

8) This value was then multiplied by the new standard deviation (20 Hz) with the resulting value then being added to the new average pitch (100 Hz).

9) The resulting value was the new normalised value for that specific point.

10) All of the original point values were replaced with their respective normalised point value using the 'Remove points between' and 'Add point' functions.

11) The normalised PitchTier and Manipulation Object were then selected together allowing for the 'Replace pitch tier' function to be used which replaced the original, unedited pitch tracking with the normalised PitchTier.

12) A resynthesis of the Manipulation Object was then created using the 'Get resynthesis (overlap-add)' function which created a new Sound Object. This resynthesis method (realised by Moulines & Charpentier (1990)) allows for the manipulation of the pitch and duration of an acoustic speech signal. When a Sound is created from a Manipulation Object using this method, the following steps are performed:

- From the PitchTier, new points are generated along the entire time domain, with the method of PitchTier: To PointProcess.
- The period information in the original pulses (available in the Manipulation object) is used to remove from the new pulses all points that lie within voiceless intervals (i.e., places where the distance between adjacent points in the original pulses is greater than 20 ms).
- The voiceless parts are copied from the source Sound to the target Sound, re-using some parts if the local duration is greater than 1.
- For each target point, we look up the nearest source point. A piece of the source Sound, centred around the source point, is copied to the target Sound at a location determined by the target point, using a bell-shaped window whose left-hand half-length is the minimum of the left-hand periods adjacent to the source and target points (and analogously for the right-hand half-length).

13) This new (normalised) Sound Object was then saved as a .wav file.

5.2.4. Experiment design

5.2.4.1. Pre-experiment instructions

Both the Pilot Study and the Main Experiment were comprised of three sections. For the Pilot Study, each section was made up of eight tasks, whereas for the Main Experiment, Section One was made up of five tasks and Section Two and Three were composed of 15 tasks each. Prior to starting the experiment, instructions were provided which explained that they would be required to make judgments on the similarity between speech samples and that some of these samples would be delexicalised. In these instructions, the delexicalised samples were simply described as samples where the lexical content had been removed and an average pitch applied, and it was explained that this was done in order to allow listeners to focus on the rhythmic patterns of the speech samples. In order to get familiarised with what a delexicalised sample sounded like in comparison to an original sample, participants were then provided with two example samples which they could listen to as many times as they wished: an original sample and a delexicalised version of the original sample. Participants were also advised that it would be beneficial to wear headphones throughout the experiment and that they would receive additional instructions prior to starting each section of the experiment.

As mentioned above, and as will be seen over the next three subsections which detail the instructions for each of the three sections of the experiment, listeners are instructed to 'focus on the rhythmic patterns' of the speech samples. When designing the perception experiments and deciding on the wording of instructions which listeners should be provided with, it was decided that no further explanation as to what 'rhythmic patterns' may constitute should be offered. If further explanation was to be provided, this would have instructed listeners as to potential possible cues they might focus on. For example, instructions could have taken the following format (bold text indicates further elaborative instructions):

Focussing on the rhythmic patterns of the speech samples, your task is to decide which delexicalised sample is that of the original Answerphone Message.

You may wish to consider the following features:

- **Pitch / intonation patterns**
- **The lengths of syllables**
- **The rate of the speech**
- **Pausing behaviour**
- **Loudness**
- **How chunks of speech are distributed**
- **(etc.)**

It was decided further elaboration as to what constitutes 'rhythmic patterns' should not be offered as this would have the potential to prime the listeners to focus on – and potentially *only* on – the features which were suggested. Section Two of the experiment required listeners to provide qualitative feedback to explain why they had made the decision that they had (i.e., what cues were they using in making their identification assessments; see Section 5.2.4.3 below), with the purpose of this being to inform the development of a perceptual framework for assessing speech rhythm (see Chapter 6). Providing additional explanation as to what listeners may wish to focus on was deemed to have the potential to influence the qualitative feedback provided by listeners (e.g., feedback being less 'natural' and more constrained to the suggested features).

The following subsections explain the makeup of each of the three sections in turn.

5.2.4.2. Section One

Prior to the commencement of the tasks for Section One, participants were provided with the following instructions:

- For this section you will be presented with the opening 30 seconds of an Answerphone Message.
- You will also be presented with two delexicalised samples: Sample_A and Sample_B.
- One of these samples is the same stretch of speech as the original Answerphone Message.

- Focussing on the rhythmic patterns of the speech samples, your task is to decide which delexicalised sample is that of the original Answerphone Message.
- You may listen to the Answerphone Message and the two delexicalised samples as many times as you wish.
- There are 5 tasks to complete in this section of the experiment (8 tasks in the Pilot Study).

Following these initial instructions, each of the tasks for Section One was presented on its own screen with the following instructions:

- Focussing on the rhythmic patterns of the speech samples, which delexicalised sample do you think most closely resembles the original 'Answerphone Message'?

This initial section of the experiment was intended to serve the purpose of being a 'training stage' of sorts, where participants could become familiar with what the delexicalised speech samples sounded like and also get used to focussing on just rhythmic information when making their speaker identification assessments. The number of tasks in this section was reduced from eight in the Pilot Study down to five for the Main Experiment as it was decided that having participants (particularly the expert listeners) dedicate more time towards Section Two and Section Three was of greater importance to the experiment as a whole.

5.2.4.3. Section Two

Following the completion of Section One, participants were provided with the following instructions for Section Two:

- For this section you will be presented with the opening 30 seconds of an Answerphone Message.

- You will also be presented with two delexicalised samples: Sample_A and Sample_B.
- One of these delexicalised samples is the same speaker as the Answerphone Message, but the sample is taken from later on in the Answerphone Message recording.
- The other sample is a different speaker.
- Focussing on the rhythmic patterns of the speech samples, your task is to decide which delexicalised sample contains the same speaker as the original Answerphone Message.
- Once you have made your decision, you will be asked to explain why you have made that decision.
- You may listen to the Answerphone Message and the two delexicalised samples as many times as you wish.
- There are 15 tasks to complete in this section (8 tasks in the Pilot Study).

Following these initial instructions, each of the tasks for Section Two was presented on its own screen with the following instructions:

- Focussing on the rhythmic patterns of the speech samples, which delexicalised sample do you think contains the same speaker as the original 'Answerphone Message'?

After deciding between Sample_A or Sample_B, participants were then prompted with the following question (on the same screen):

- Please explain why you have made this decision in the text box below.
- Please be as detailed as possible in your response.

For the Main Experiment, Section Two was extended from containing eight tasks to containing fifteen tasks. This decision was made in order to accumulate more

qualitative feedback (specifically from expert listeners), as this feedback would be fundamental in contributing towards developing a perceptual rhythm framework for forensic analysis (see Chapter 6).

5.2.4.4. Section Three

Following the completion of Section Two, participants were provided with the following instructions for Section Three:

- For this section you will be presented with two samples of delexicalised speech which are both 30 seconds in length.
- Both samples are taken from an Answerphone Message recording.
- Sample_A will always be the first 30 seconds of an Answerphone Message and Sample_B will always be 30 seconds taken from later on in an Answerphone Message.
- Focussing on the rhythmic patterns of the speech samples, your task is to rate the similarity of each pair of speech samples on a scale of 1 to 9:

very similar (1) - (9) very different

- You may listen to the two delexicalised samples as many times as you wish.
- There are 15 tasks in this section of the experiment.

Following these initial instructions, each of the tasks for Section Three was presented on its own screen with the following instructions:

- Focussing on the rhythmic patterns of the speech samples, please rate the similarity of this pair of delexicalised speech samples.

very similar (1) - (9) very different

For the Main Experiment, Section Three was extended from containing eight tasks to containing fifteen tasks. This decision was made given that this section posed a different challenge to listeners in that they were only presented with delexicalised speech samples when making the similarity judgements with no original voice sample to consult. The tasks in this section therefore ensured that listeners could only draw upon the rhythmic behaviour they perceived when making their similarity assessments. That is, in comparison to Section Two where listeners could potentially make use of idiosyncratic disfluency behaviour which could have been a more easily identifiable feature in the delexicalised samples (e.g., prolongations of filled pauses). Of the fifteen tasks, eight tasks contained same-speaker pairings and seven tasks contained different-speaker pairings.

As alluded to above, listeners were presented with a nine-point scale from which to rate whether the delexicalised samples sounded similar (1-4) or different (6-9) to one another. In order to visualise the results of this section in a way comparable to the previous two sections, if a given task contained a same-speaker pair, for example, then a response of (1), (2), (3), or (4) would be recorded as 'correct' and a response of (6), (7), (8), or (9) would be recorded as 'incorrect'. A response of (5) (i.e., the middle of the scale, indicating that the listener thought the samples were neither similar nor different from one another) would be recorded as 'incorrect'.

5.2.5. Statistics

For Section Three of the Main Experiments, in order to assess how listeners and listener groups were using the 9-point Likert scale (e.g., were some listeners using the extremes of the scale to indicate their confidence? Where some groups of listeners

more cautious in their use of the scale?), the response given for each task was z-scored. This was done in the following way:

1. Firstly, the ratings provided (i.e., (1) – (9)) were recoded.

2. For tasks with same-speaker pairs, a correct response of (1) on the Likert scale would be coded as [+4], a correct response of (2) as [+3], and so on. An incorrect response of (9) would be coded as [-4], an incorrect response of (8) as [-3], and so on.

3. For tasks with different-speaker pairs, a correct response of (9) on the Likert scale would be coded as [+4], a correct response of (8) as [+3], and so on. An incorrect response of (1) would be coded as [-4], an incorrect response of (2) as [-3], and so on.

4. Each listener would therefore have a score of [+4] – [-4] for each task.

5. From these scores, the mean and standard deviation were calculated.

6. The new scores were then z-scored.

It is worth noting here that z-scoring the recoded scores did not modify the recoded data in any way and just served to project the [+4] – [-4] scores onto a z-score scale. It was decided that visualising results in the form of z-scores (as opposed a new [+4] – [-4] scale) would be preferable in line with the visualisation of z-scored data in previous chapters (i.e., Chapter 3 and Chapter 4). As such, the resulting z-scores could then be visualised using boxplots in which listeners with higher z-scores would be those who were making accurate discrimination decisions the most confidently (i.e., by using the extremes of the Likert scale).

For this section of the experiment, as there were two delexicalised samples which were either same-speaker pairs or different speaker pairs (i.e., a different format to the previous two sections of the experiment), the percentage of correct responses (%C)

for each listener could be calculated in a more formulaic way. This was done following the method outlined by Pallier (2002) which is as follows:

$$\%C = 100*((Hits + CRs)/(Hits+Misses+FA+CR))$$

Where:

Hit = correct identification of a same-speaker pair

CR (Correct Rejection) = correct identification of a different-speaker pair

FA (False Alarm) = wrong identification of a different-speaker pair

Miss = wrong identification of a same-speaker pair

In order to apply the above formula to the nature of the results obtained for Section Three (i.e., Likert scale data), listeners' responses had to be simplified down to either being 'correct' or 'incorrect' (n.b., a rating of (1) is 'very similar' as opposed to a rating of (9) which is 'very different'). This was done in the following way:

- For tasks with same-speaker pairs, a 'correct' response was recorded if a given listener responded with either (1), (2), (3), or (4) on the 9-point Likert scale. A response of (6), (7), (8), or (9) was recorded as an 'incorrect' response.
- For tasks with different-speaker pairs, a 'correct' response was recorded if a given listener responded with either (9), (8), (7), or (6) on the 9-point Likert scale. A response of (4), (3), (2), or (1) was recorded as an 'incorrect' response.
- Where listeners responded with (5) to any task (same-speaker or different-speaker pairing), that is, the middle of the Likert scale (i.e., indicating no degree of similarity or difference), this response was also recorded an 'incorrect response'.

5.3. Results

This section presents the results from both the Pilot Study and the Main Experiment. The Pilot Study aimed to explore how well naïve listeners, with no linguistic training,

could make meaningful speaker identifications based solely on the rhythmic attributes of speech samples. Additionally, the study was designed to test the experiment's methodology, assess the difficulty of the tasks, evaluate the clarity of the instructions provided to participants, and gauge the time required to complete the experiment. The results of the Pilot Study are presented in Section 5.3.1. This was followed by the Main Experiment, which extended the Pilot Study by including different groups of listeners with varying levels of expertise. While the Pilot Study consisted of eight tasks across three sections, the Main Experiment included five tasks in Section One and 15 tasks in both Section Two and Section Three. The results of the Main Experiment are presented in Section 5.3.2. For both the Pilot Study and the Main Experiment, each of the three sections is afforded its own analysis in which results are discussed in terms of participant performance, which sections and tasks produced the most correct responses, and whether any patterns in success rates can be determined.

5.3.1. Pilot Study

5.3.1.1. Overview of results

In Section One and Section Two, listeners were asked to make a binary decision on whether the delexicalised samples contained the same speaker as the original (non-delexicalised) samples. In Section Three, listeners were tasked with rating the similarity of pairs of delexicalised speech samples. To present the results of Section Three in a manner consistent with Sections One and Two, any task containing a same-speaker pair, for instance, would have responses of (1), (2), (3), or (4) recorded as "correct," whilst responses of (6), (7), (8), or (9) would be recorded as "incorrect." A response of (5) (the midpoint of the scale, indicating that the listener considered the samples neither similar nor different) would also be recorded as "incorrect."

Results by task

Figure 5.3 displays the overall results (for all of the listeners) for each of the eight tasks in each of the three sections of the experiment.

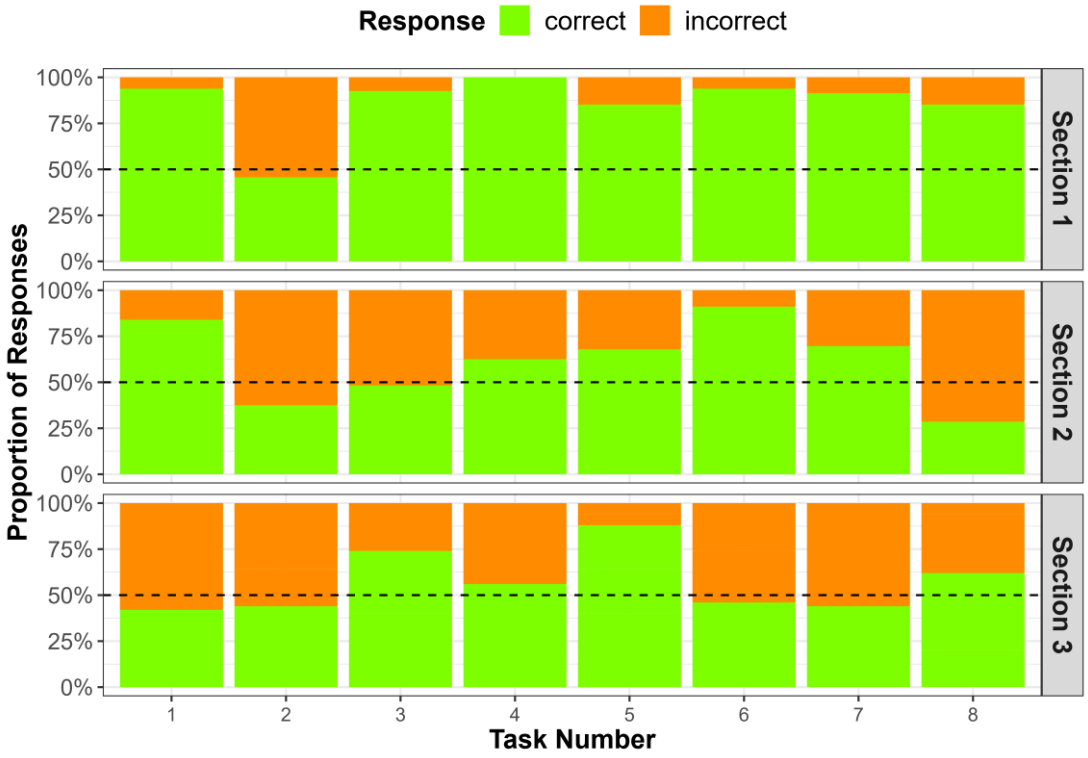


Figure 5.3. Stacked bar plots of the responses (all participants) for each of the eight tasks in each of the three sections of the experiment. The dotted line indicates the chance level (50%) as responses were judged to be either correct or incorrect (n.b., tasks across each section of the experiment are not related to one another).

When collectively assessing the results by task over the three sections of the experiment, it is observable that listeners experienced an increase in difficulty as they progressed through the three sections. In Section One, 81 out of 96 responses were correct (84.4%), in Section Two, 56 out of 96 responses were correct (58.3%) and in Section Three, 50 out of 96 responses were correct (52.0%).

Results by listener

Figure 5.4 provides a visualisation of the same results by listener.

Figure 5.4. Stacked bar plots of the participants' responses across the three sections of the experiment. The dotted line indicates the chance level (50%) as responses were judged to be either correct or incorrect.

From observing Figure 5.4, it is clear that no participants particularly stood out in terms of superior performance across the three sections. Within the group, only three listeners (nep01, nep09 and nep12) achieved above 50% (chance level) of correct responses on each of the three sections.

5.3.1.2. Section One

For each of the tasks in Section One, listeners were presented with an original speech sample and two delexicalised samples, with one of the delexicalised samples being the same section of speech as the original. Listeners had to select which of the delexicalised samples contained the same speaker as the original speech sample.

Results by listener

Figure 5.5 displays the results for each of the participants for Section One.

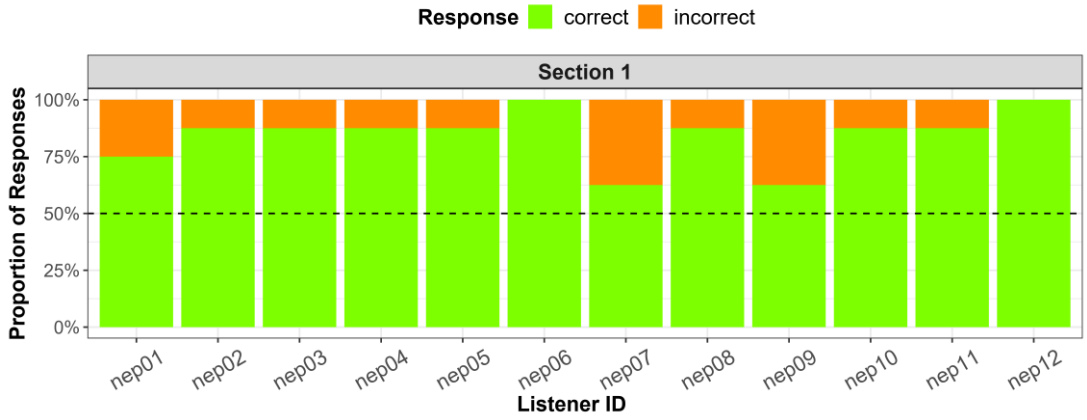


Figure 5.5. Bar plots of the participants' responses for Section One of the experiment. The dotted line indicates the chance level (50%) as responses were judged to be either correct or incorrect.

Out of the 12 participants, only two (nep06 and nep12) identified the correct delexicalised sample correctly for all eight tasks. However, an additional seven of the twelve listeners achieved correct responses for seven out of the eight tasks. The most problematic task for these speakers can be observed (from Figure 5.4) to be the second task which had an overall correct response rate of less than 50% (an explanation for this is provided in Section 5.4.2.1).

5.3.1.3. Section Two

For each of the tasks in Section Two, listeners were presented with an original speech sample and two delexicalised samples, with one of the delexicalised samples containing the same speaker as the original. The difference between the tasks in this section and those in Section One is that neither of the delexicalised samples were the same stretch of speech as the original. In this section, listeners were also required to provide qualitative feedback in which they explained why they had made the choice

that they had made (i.e., what rhythmic cues were listeners focussing in on when making their speaker identification assessments).

Results by listener

Figure 5.6 displays the results for each of the participants for Section Two.

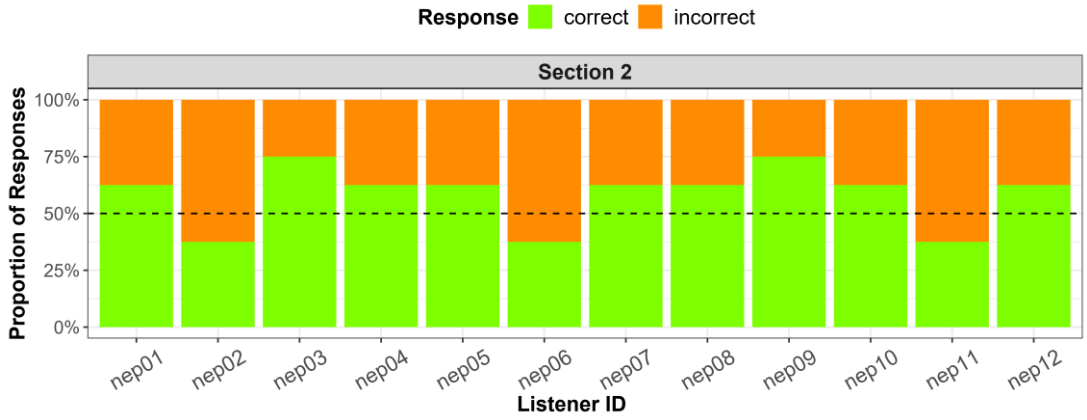


Figure 5.6. Stacked bar plots of the participants' responses for Section Two of the experiment. The dotted line indicates the chance level (50%) as responses were judged to be either correct or incorrect.

It is immediately noticeable that participants encountered a greater degree of difficulty when completing Section Two. Although nine of the twelve listeners scored above chance level (50%) with regards to correct responses, only two of these participants identified the correct sample in six of the eight tasks with the other seven participants recording correct responses in only five of the eight tasks.

Qualitative results

This section of the Pilot Study also required listeners to provided qualitative feedback relating to why they made the decision that they did. That is, what rhythmic features were they focussing on when making their speaker identification assessments. Table 5.4 below provides a summary of these qualitative observations.

Table 5.4. Summary of qualitative feedback obtained from non-expert listeners in the pilot perception study. Examples provided are verbatim.

FEATURE	ASPECT OF FEATURE	EXAMPLE(S)
PAUSING BEHAVIOUR (SILENT PAUSES)	• Frequency • Length • Distribution	• *The rhythm is quite fast, and there are hardly any pauses.* • *the length of the pauses seemed to reflect the pattern of the speaker* • *Sample A has breaks in the right places*
FILLED PAUSES	• Frequency • Length	• *Seemed to have long 'erms' and long pauses followed by frequent short 'erms'.* • *when the man says 'uhh' in pausing in his sentences, the length of that matches closely*
SPEECH RATE / ARTICULATION RATE	• Relative tempo • Variability	• *Seems to be a faster tempo again* • *I believe it has a slower pace, with longer sounds reflecting the 'uhh' pauses in the sentences.* • *the rhythmic pattern is slower pace and there is less intonation*
PITCH / INTONATION	• Movements • Inflections • Rises & Falls • Tone • Monotony	• *there were few words which changed noticeable in pitch…* • *Sample B, as there are periods of higher pitch and there are more rise and falls in the rhythmic pattern.* • *Again, the lifting of the tone at the end of some sentences.* • *Sample B is too monotone*
SYLLABLES / SPEECH UNITS	• Prolongations / lengthening of units • Short / snappier units	• *Rhythm sounds more similar; four words tend to be clustered then a slight pause.* • *Sample A contains a combination of drawn-out notes and short sharp notes* • *Sample B has shorter sounds, whereas sample A has longer sounds which fits with the message*
OVERALL FLOW	• Punchy • Flowing	• *a more continuous flow to the rhythm.* • *…also the notes in Sample A seem to short and punchy.*
STRESS / LOUDNESS	• Loudness	• *Sample B seems to be more punchy and some quite loud start to words*

For discussion pertaining to the qualitative feedback obtained for the non-expert listeners in the Pilot Study, see Section 5.4.2.2.

5.3.1.4. Section Three

For each of the tasks in Section Three, listeners were presented with two delexicalised speech samples and had to rate the (dis)similarity of the sample pairs using a 9-point Likert scale with (1) being very similar and (9) being very different.

Results by task

Figure 5.7 displays the results for each of the tasks for Section Three.

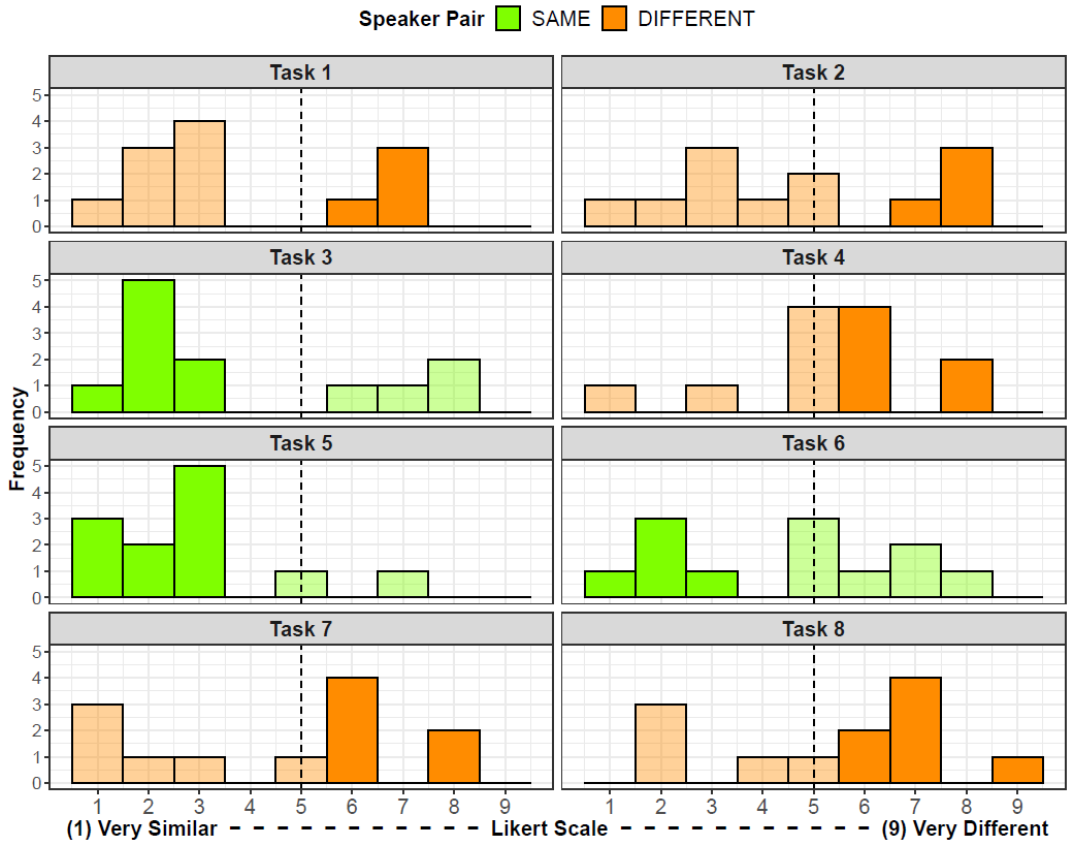


Figure 5.7. Histograms of the listeners' responses for Section Three of the experiment. Tasks with same-speaker pairs are depicted in green and tasks with different-speaker pairs are shown in orange. Full colour bars are indicative of correct responses and transparent bars indicate incorrect responses.

From observing the histograms of the eight tasks it is noticeable that the participants found greater degrees of success with some tasks than with others. Interestingly, of the eight tasks, the two which had the greatest number of correct responses were both tasks which contained same-speaker pairs (Task 3 and Task 5), with only three of the eight tasks containing same-speaker parings. The task which yielded the most correct responses was Task 5 with ten of the 12 participants recording correct responses, followed by Task 3 in which eight of the twelve participants judged the speaker-similarity correctly. Also of note is that for both Task 3 and Task 5 the majority of participants rated these same-speaker pairs correctly using the extremes of the rating

scale. That is, giving a rating of either (1) or (2) which indicates that they found the pair of delexicalised samples to be 'very similar'. This contrasts with the correct responses for the different-speaker pairs as participants were less likely to make use of the extremes towards the 'very different' end of the scale, with most speakers opting to select a rating of either (6) or (7) on the scale when (correctly) assessing different-speaker pairs.

5.3.2. Main Experiment

5.3.2.1. Overview of results

Results by listener and listener group

Figure 5.8 displays the overall results for each of the listeners across all sections of the experiment and Figure 5.9 shows the percentage of correct responses (%C) for each of the listener groups. Listeners have been grouped according to their expertise.

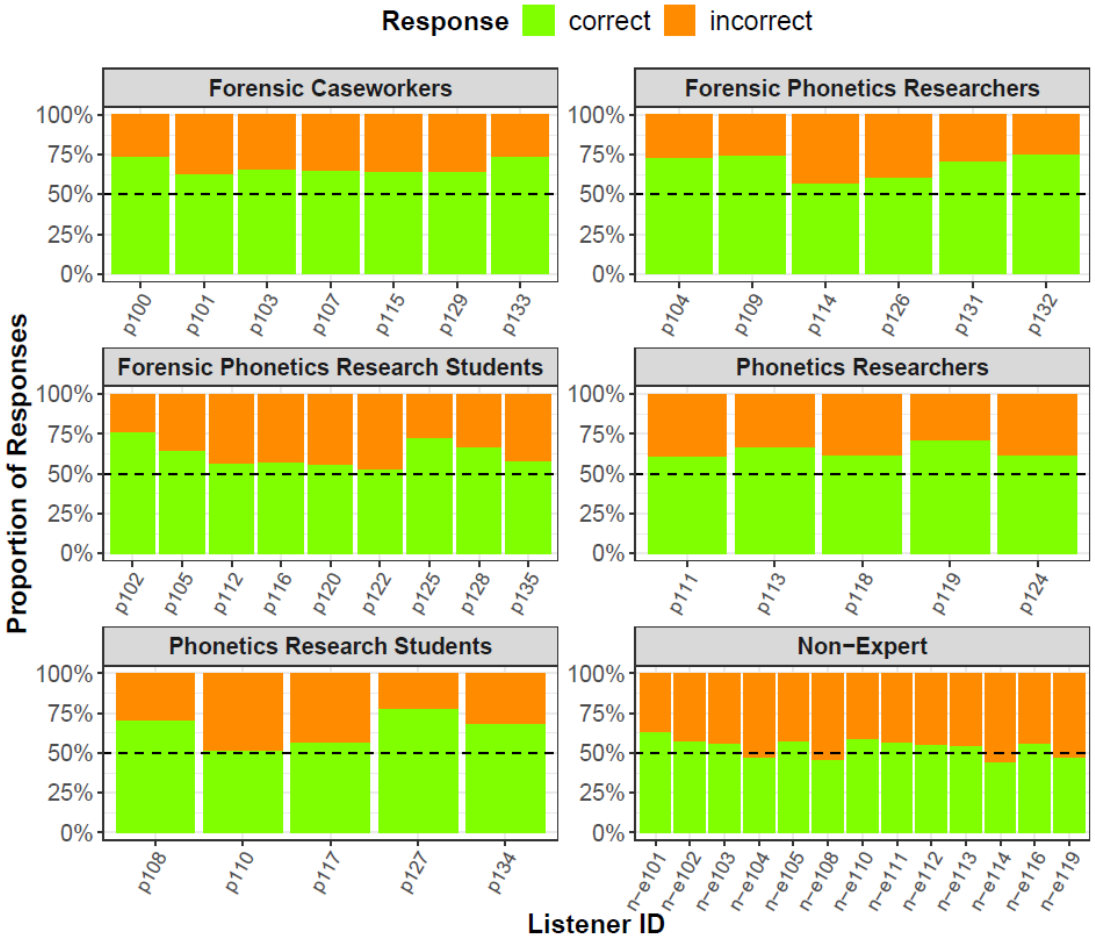


Figure 5.8. Bar plots of the participants' responses across all sections of the experiment. The dotted line indicates the chance level (50%) as responses were judged to be either correct or incorrect.

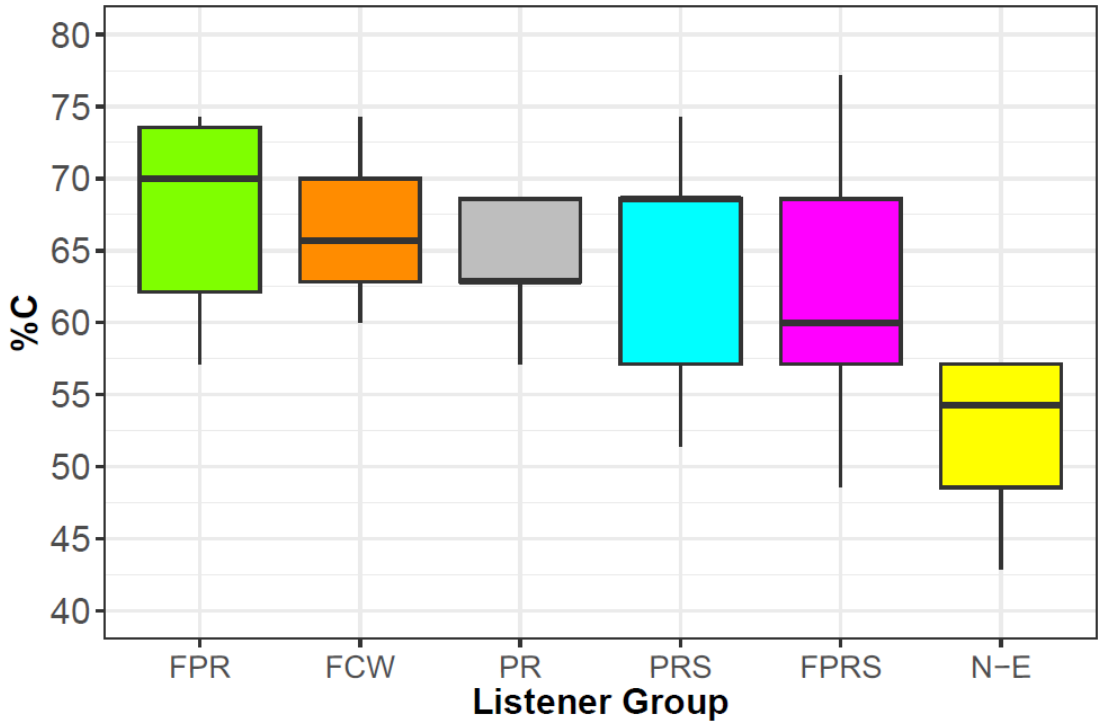


Figure 5.9. Boxplots showing the percentage of correct responses (%C) for each of the listener groups for the experiment as a whole. Boxplots are ordered by mean score, highest to lowest, left to right.

When considering the experiment as a whole, it is evident from the above plots that participants had varying degrees of success with regards to making accurate speaker identification assessments. For some groups, more variation is evidenced between the participants. For example, whilst the participants in the Forensic Caseworkers' (FCW) group were rather consistent with regards to their overall proportion of correct responses, there is more between-participant variation present within the Forensic Phonetics Research Student (FPRS) group and the Non-Expert (NE) group. The group which had the highest proportion of accurate responses for the entire experiment was the Forensic Phonetics Researchers' (FPR) group with 143 out of 210 correct responses (68.1% correct), followed by the FCW group with 163 out of 245 correct responses (66.5%). Both the Phonetics Researchers' (PR) group and the Phonetics Research Students' (PRS) group both recorded 112 out of 175 correct responses (64%), with the FPRS group (the largest of the groups with nine participants) attaining 197 out of 315 correct responses (62.5%). The group with the lowest proportion of correct responses was the NE group with 242 out of 455 correct responses (53.2%).

When considering the expert listeners as a holistic cohort, they recorded a total of 725 out of 1120 correct responses (64.7%) with this being a statistically significant difference in comparison to the non-expert listeners responses (two-tailed t-test: p = 0.004). Amongst the expert listeners, those who had expertise in forensic phonetics (groups FCW, FPR and FPRS) recorded 65.2% of correct responses overall, where experts with no forensic expertise (groups PR and PRS) recorded 64.0% of correct responses with this difference not being statistically significant (two-tailed t-test: p = 0.825).

5.3.2.2. Section One

For each of the tasks in Section One, listeners were presented with an original speech sample and two delexicalised samples, with one of the delexicalised samples being the same section of speech as the original. Listeners had to select which of the delexicalised samples contained the same speaker as the original speech sample.

Results by listener

Figure 5.10 displays the results for each of the participants for Section One. Participants have been grouped according to their expertise.

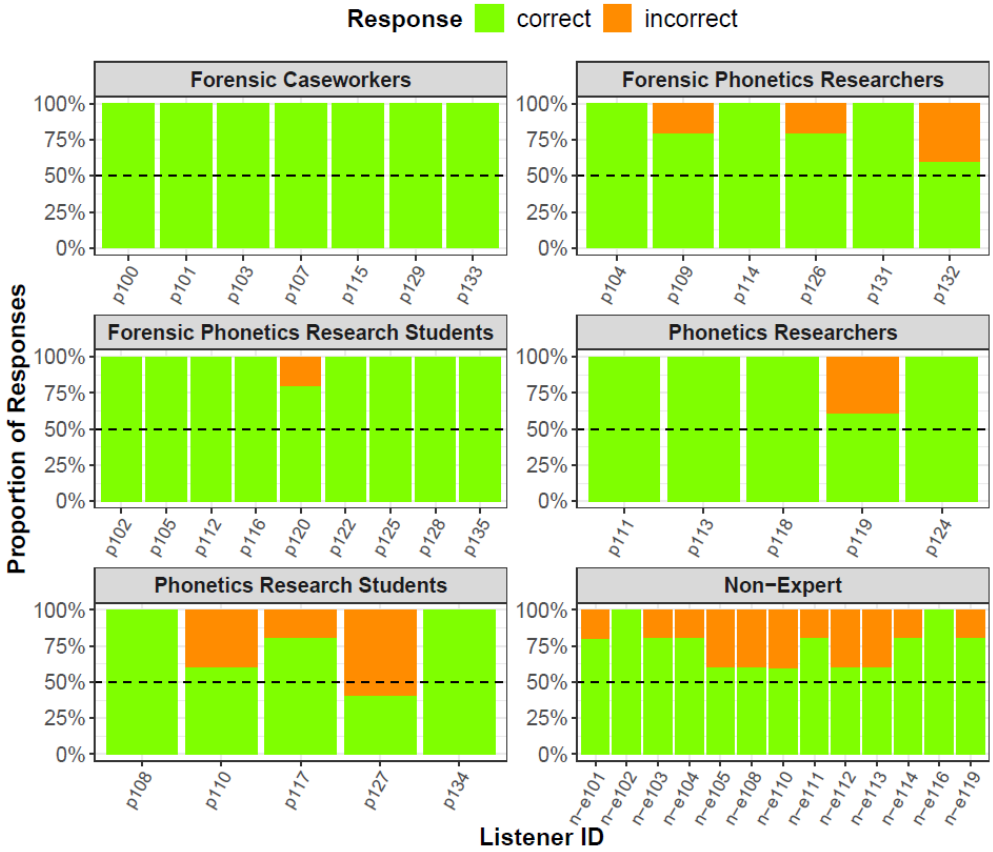


Figure 5.10. Bar plots of the participants' responses for Section One of the experiment. The dotted line indicates the chance level (50%) as responses were judged to be either correct or incorrect.

From observing the plots for the six groups of participants, it is apparent that, on the whole, the tasks within this section were fairly straight forward with regards to correct speaker identifications being made. Of the 45 participants, 26 recorded correct responses across all five of the tasks, and as a collective there were 196 out of 225 correct speaker identification assessments making the overall proportion of correct responses 87.1%. The FCW group has the best performing with all seven participants recording correct responses across all five tasks, followed closely by the FPRS group where there was only one incorrect response recorded from all nine participants. The greatest number of incorrect responses was attributed to the NE group with 16 incorrect responses overall, with only two of the 12 participants recording making correct speaker identification assessments across all five tasks.

Results by task

To determine whether there were any tasks which were specifically problematic for the participants, Figure 5.11 displays the results for each of the tasks for Section One (participants have been grouped according to their expertise).

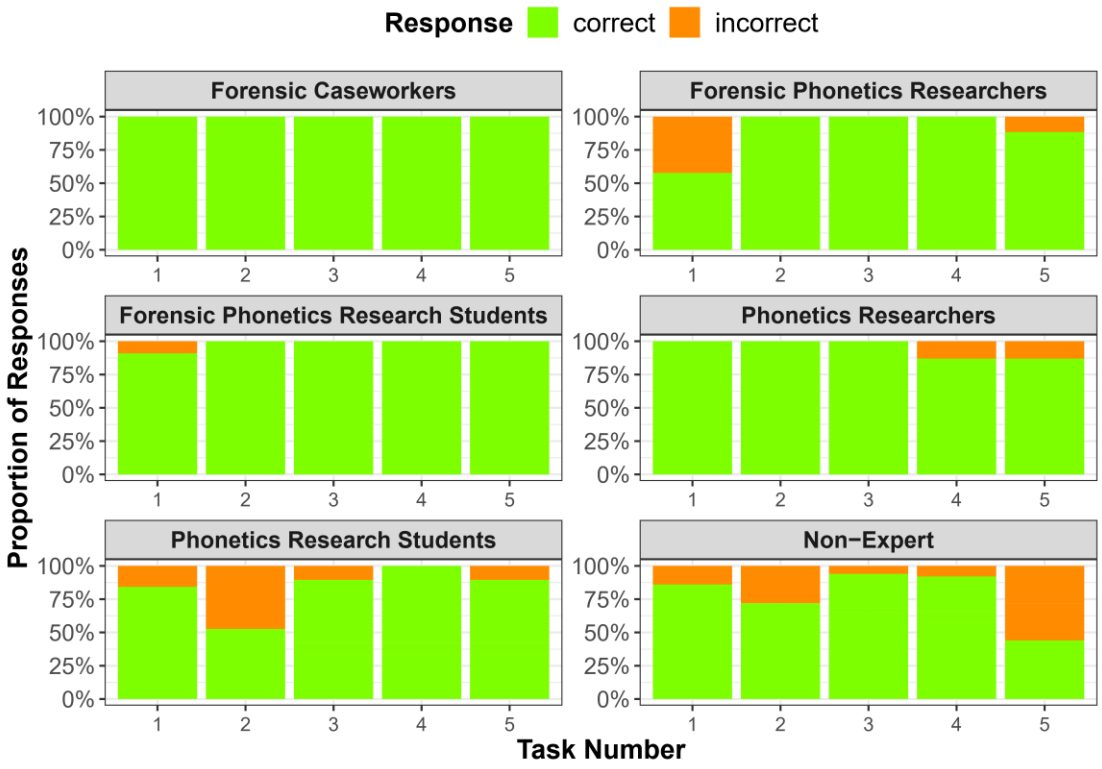


Figure 5.11. Stacked bar plots of the participants' responses for each of the tasks for Section One of the experiment.

From observing the plots, it is evident that there is no specific pattern present with regards to a certain task being overwhelmingly more difficult than the rest (cf. Task 2, Section One of the Pilot Study, Section 5.3.1.1). The task which was most problematic overall was Task 5 with 11 of the 45 participants recording incorrect responses (eight of these responses being attributed to the NE group).

5.3.2.3. Section Two

For each of the tasks in Section Two, listeners were presented with an original speech sample and two delexicalised samples, with one of the delexicalised samples containing the same speaker as the original. The difference between the tasks in this section and those in Section One is that neither of the delexicalised samples were the same stretch of speech as the original. In this section, listeners were also required to provide qualitative feedback in which they explained why they had made the choice that they had made (i.e., what rhythmic cues were listeners focussing in on when making their speaker identification assessments).

5.3.2.3.1. Quantitative results

Results by listener and listener group

Figure 5.12 displays the results for each of the listeners for Section Two and Figure 5.13 shows the percentage of correct responses (%C) for each of the listener groups. Listeners have been grouped according to their expertise.

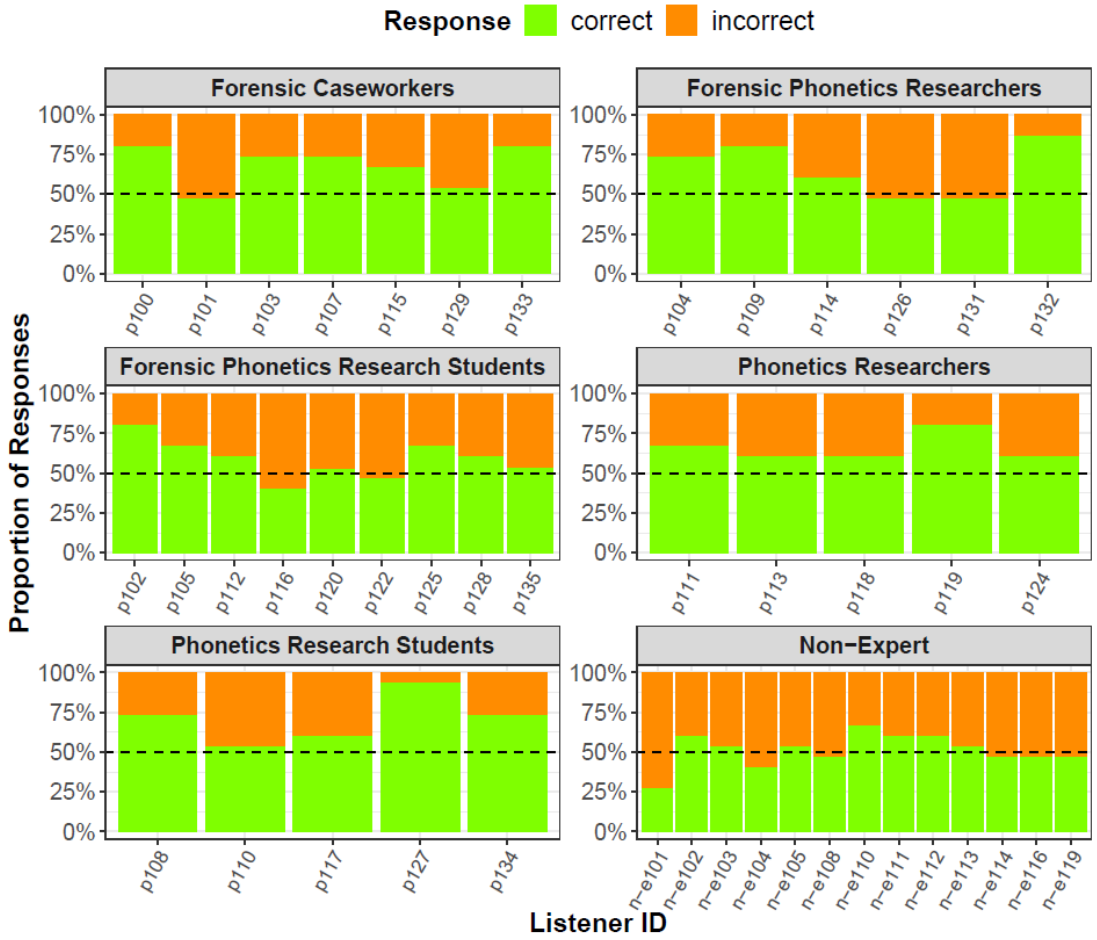


Figure 5.12. Bar plots of the participants' responses for each of the tasks for Section Two of the experiment.

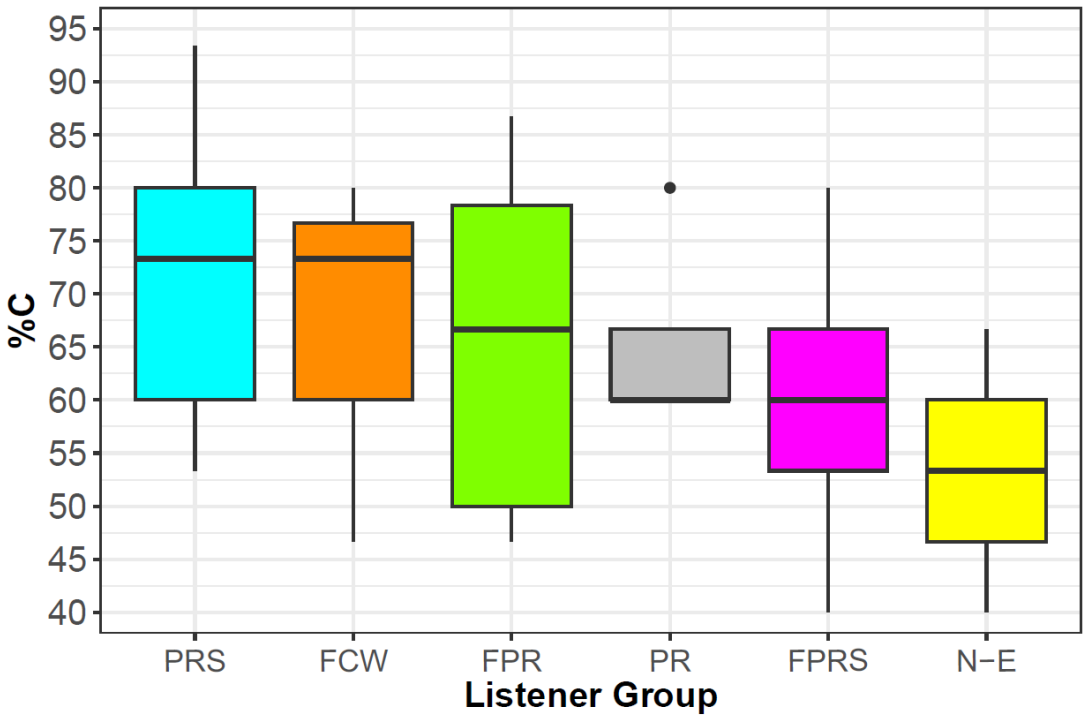


Figure 5.13. Boxplots showing the percentage of correct responses (%C) for each of the listener groups for Section Two of the experiment. Boxplots are ordered by mean score, highest to lowest, left to right.

In comparing the above plots to those analysed in Section One of the experiment, it is clear that the tasks in Section Two posed a greater degree of challenge to the participants as no participant recorded correct responses for all 15 tasks. With regards to the best performing participants, one participant (p127, PRS group) recorded 14 correct responses, one participant recorded 13 correct responses (p123, FPR group) and six participants recorded 12 correct responses (p100 and p133 (FCW group), p109 (FPR group), p102 (FPRS group), p119 (PR group), p108 (PRS group)). For this section, the overall proportion of correct responses was 410 out of 675 (60.7%). The PRS group was the best performing with 53 out of 75 correct responses (70.7%), followed by the FCW group with 71 out of 105 correct responses (67.6%). The group with the lowest proportion of correct responses was the NE group with 99 out of 195 correct responses (50.8%).

Results by task

To determine whether there were any tasks which presented as being more challenging in comparison to others, Figure 5.14 displays the results for each of the tasks for Section Two (participants have been grouped according to their expertise).

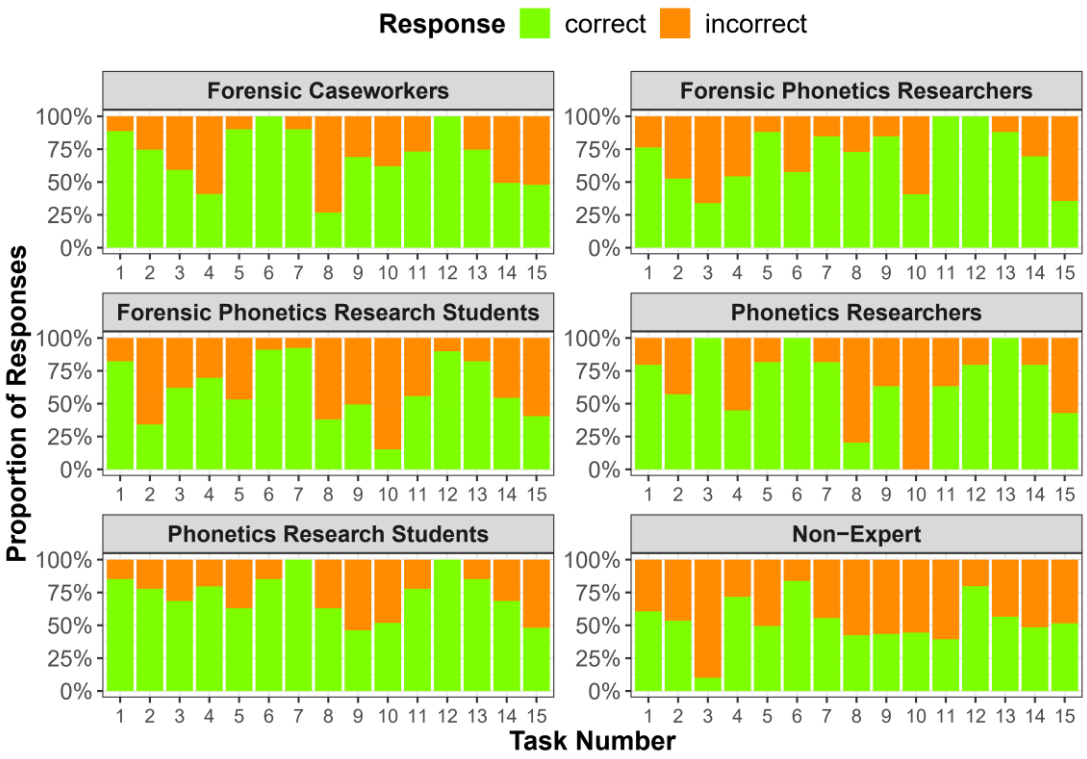


Figure 5.14. Bar plots of the participants' responses for each of the tasks for Section Two of the experiment.

Looking at the proportion of correct responses for individual tasks, there are no outright patterns present whereby all groups of participants have categorically correct or incorrect responses. There are, however, three noteworthy tasks with regards to the proportion of correct responses being recorded, namely, Task 12 (40 out of 45 correct responses (88.9%); speaker pair: WY109 (correct)/WY042), Task 6 (38 out of 45 correct responses (84.4%)) and Task 7 (35 out of 45 correct responses (77.8%)). Task 10 was seemingly the most problematic task with the lowest proportion of correct responses being recorded (15 out of 45 correct responses (33.3%); speaker pair: WY030 (correct)/WY031).

5.3.2.3.2. Qualitative results

For this section of the experiment, participants were required to provide qualitative feedback in which they explained the reasoning behind the delexicalised sample they selected as containing the same speaker as the original message. Given the substantial amount of qualitative feedback generated, a summary of the key features which participants mentioned when making *correct* speaker identification assessments are tabulated below. Further discussion pertaining to this qualitative feedback is provided in Section 5.4.3.2.

Table 5.5. Summary of qualitative feedback obtained from listeners when making correct speaker identification assessments. Examples provided are verbatim.

FEATURE	ASPECT OF FEATURE	EXAMPLE(S)
PAUSING BEHAVIOUR (SILENT PAUSES)	• Frequency • Length • Distribution • Distinctiveness	• *relatively long silent pauses* • *lots of silent pauses* • *some silent pauses but it wasn't frequent.* • *no obvious silent pauses* • *unfilled pauses in between,*
FILLED PAUSES	• Frequency • Length • Distribution • Distinctiveness	• *more filled pauses* • *the answerphone had long filled pauses with a falling intonation.* • *They also have many filled pauses which are longer than their words*
SPEECH RATE / ARTICULATION RATE	• Relative tempo • Variability • Distribution of variability	• *a much faster overall articulation rate.* • *Relatively slow articulation rate with some unfilled pauses throughout* • *speaker varies between phases of slow articulation and phases of fast articulation.*
PITCH / INTONATION	• Range • Variability • HRT	• *appears to contain prolongations and pitch variability.* • *Sample B is too monotonous. (There seems to be no pitch accents).*
SYLLABLES / SPEECH UNITS	• Prolongations • Duration • Regularity • Grouping / chunking	• *speaker has prolongations and filled pauses (often prolonged)* • *syllables are shorter and the speech rate is faster* • *Similar profiles of unstressed syllable reduction and phrase-final lengthening.*
INTONATION PHRASES / CHUNKING	• Length • Openings / closings • Prosodic variability	• *The intonation phrases in Sample_A seem to be much longer in comparison.* • *speech divided into linguistic chunks, e.g. intonation units.*
STRESS PATTERNS	• Pitch • Intensity (loudness)	• *high amplitude initially but then drops.* • *amplitude variation* • *the original has a quite monotone stress/accent pattern, like Sample A.*

5.3.2.4. Section Three

For each of the tasks in Section Three, listeners were presented with two delexicalised speech samples and had to rate the (dis)similarity of the sample pairs using a 9-point Likert scale with (1) being very similar and (9) being very different.

Results by listener and listener group

Figure 5.15 displays the results for each of the listeners for Section Three and Figure 5.16 shows the percentage of correct responses (%C) for each of the listener groups. Listeners have been grouped according to their expertise.

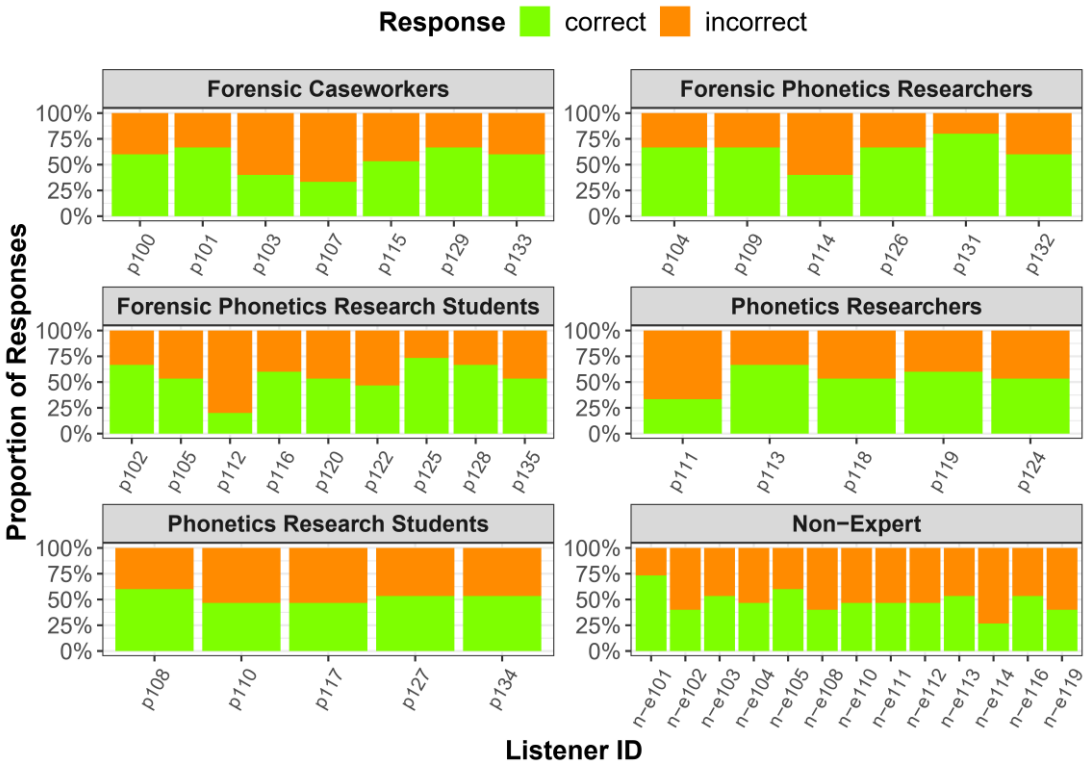


Figure 5.15. Stacked bar plots of the participants' responses for each of the tasks for Section Three of the experiment.

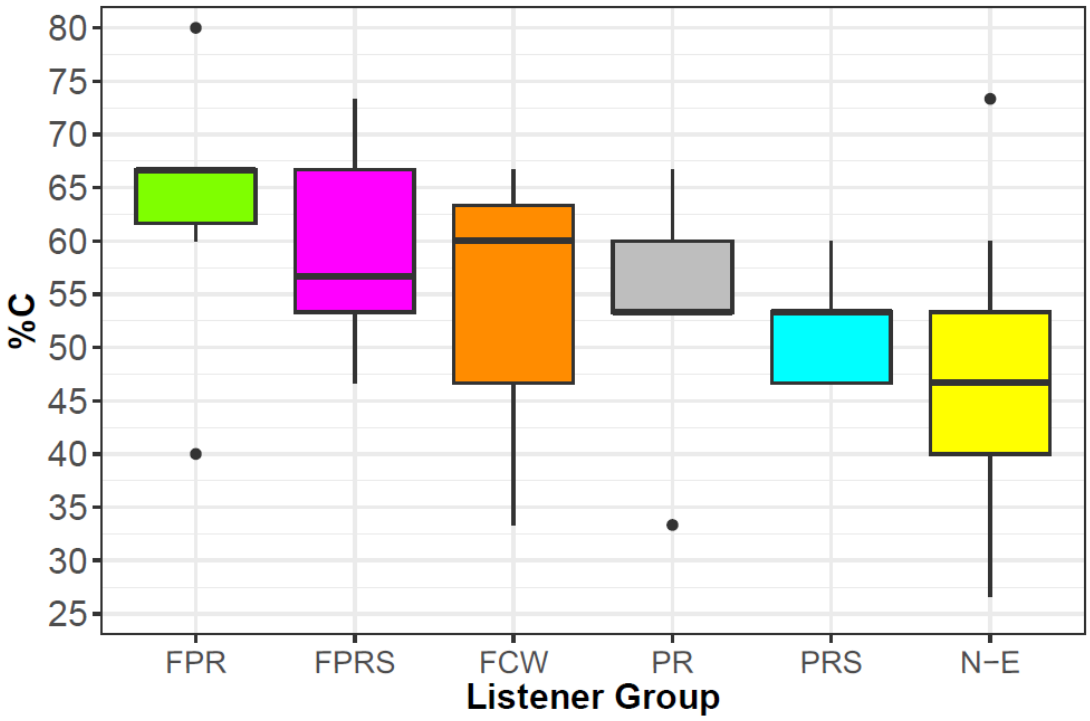


Figure 5.16. Boxplots showing the percentage of correct responses (%C) for each of the listener groups for Section Three of the experiment. Boxplots are ordered by mean score, highest to lowest, left to right.

When comparing the above plots to those of the previous two sections, it is evident that Section Three was the most difficult for the participants overall. The participant who recorded the most correct responses (12 out of 15; 80%) was p131 from the FPR group. Two participants, p125 (FPRS group) and n-e101 (NE group), recorded 11 out of 15 correct responses (73.3%), with eight participants recording correct responses for 10 out of the 15 tasks (66.7%) – seven of which are within groups with forensic expertise. For this section as a whole, there were a total of 361 out of 675 correct responses (53.5%). The best performing group was the FPR group with 57 out of 90 correct responses (63.3%), followed by the FPRS group with 74 out of 135 correct responses (54.8%), and then the FCW group with 57 out of 105 correct responses (54.3%). The group with the lowest proportion of correct responses was the NE group with 94 out of 195 correct responses (48.2%).

Results by task (all groups)

To determine whether there were any tasks which presented as being more challenging in comparison to others, Figure 5.17 displays the results for each of the tasks for Section Two (participants have been grouped according to their expertise).

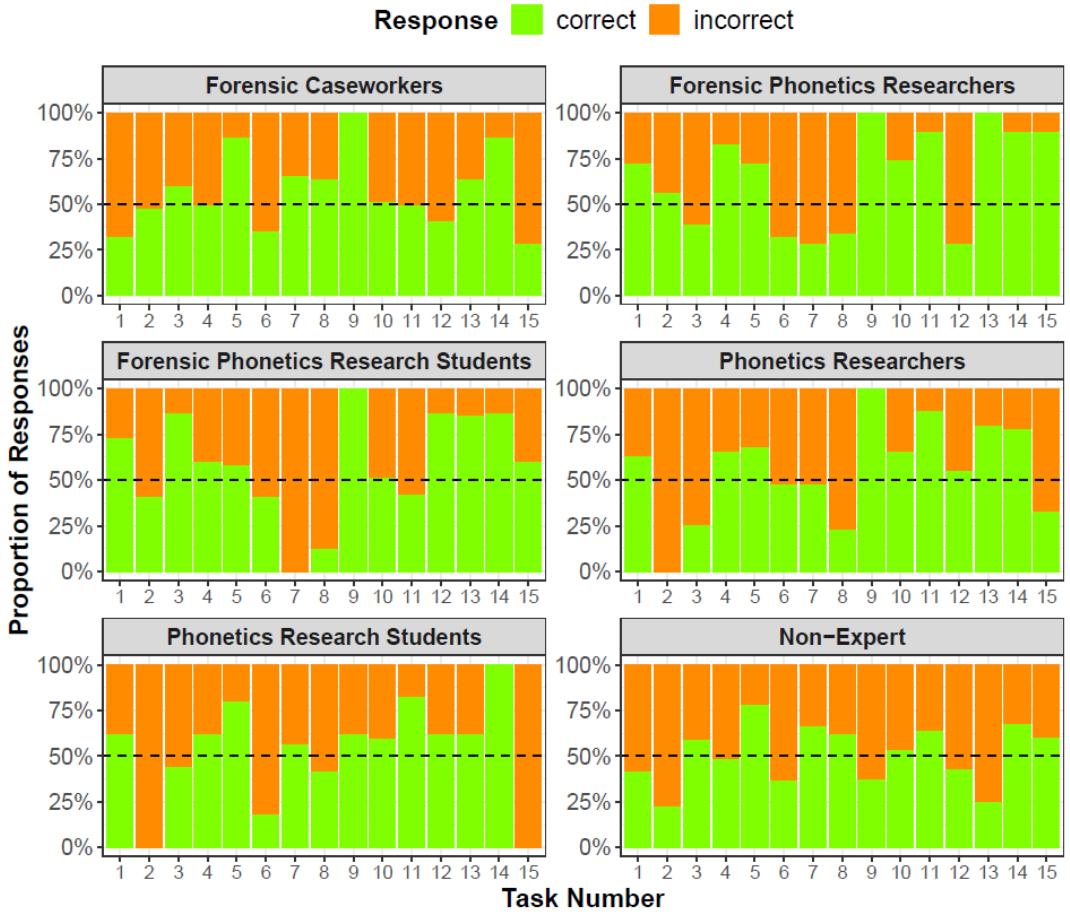


Figure 5.17. Stacked bar plots of the participants' responses for each of the tasks for Section Three of the experiment.

In a similar manner to Section Two, there are no tasks whereby all participant groups have categorically correct or incorrect responses. Again, there are certain tasks which merit further investigation given the proportion of correct responses recorded by the participants. To facilitate a more fine-grained examination of the responses, Figure 5.18 shows histograms for all of the participants' responses as a whole.

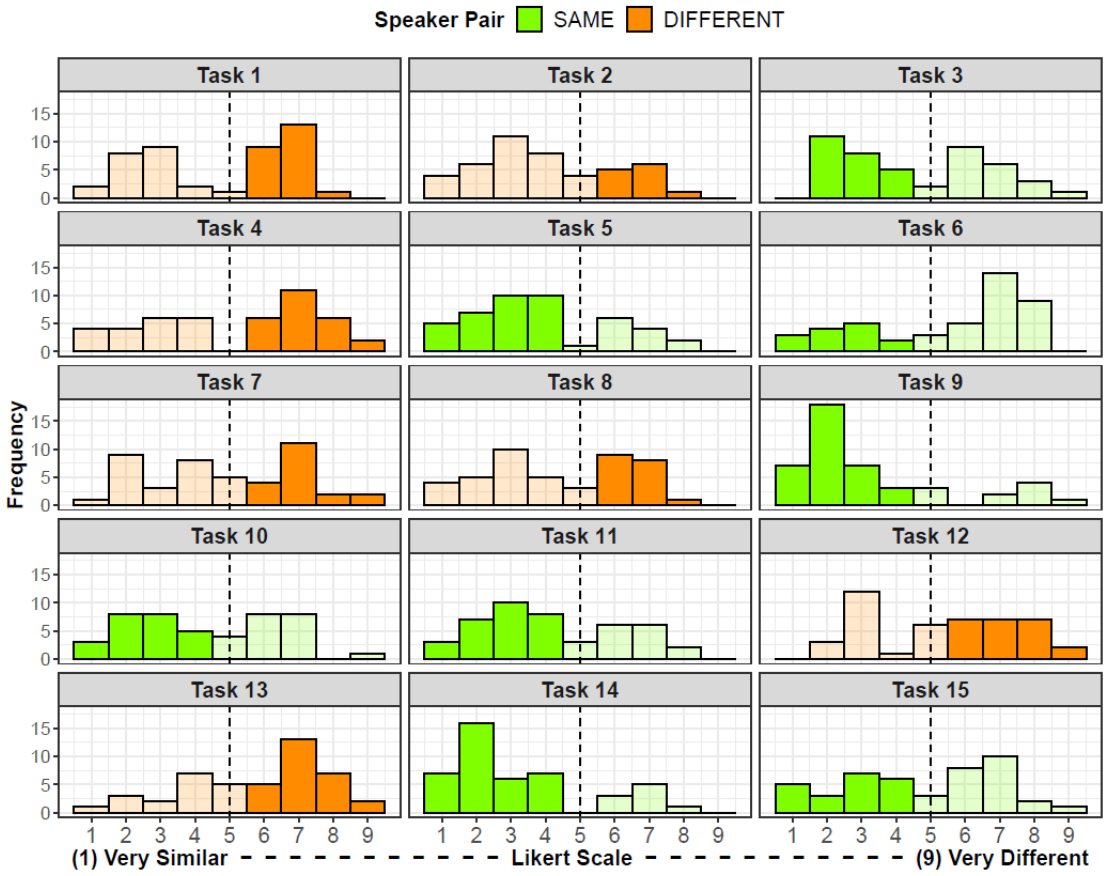


Figure 5.18. Histograms of the participants' responses for Section Three of the experiment. Tasks with same-speaker pairs are depicted in green and tasks with different-speaker pairs are shown in orange. Full colour bars are indicative of correct responses and transparent bars indicate incorrect responses.

When all the groups are considered together, the task with the highest proportion of correct responses is Task 14 with 36 out of 45 correct responses (80%; same-speaker pair (WY109)) closely followed by Task 9 with 35 out of 45 correct responses (77.8%; same-speaker pair (WY177)). The task which had the lowest proportion of correct results was Task 2 with 13 out of 45 correct responses (28.9%; different-speaker pair (WY023/WY030)).

Results by task (expert listener groups)

To examine just the responses of the expert listeners, Figure 5.19 shows histograms for the expert listeners (groups FCW, FPR, FPRS, PR and PRS) where the non-expert group has been omitted.

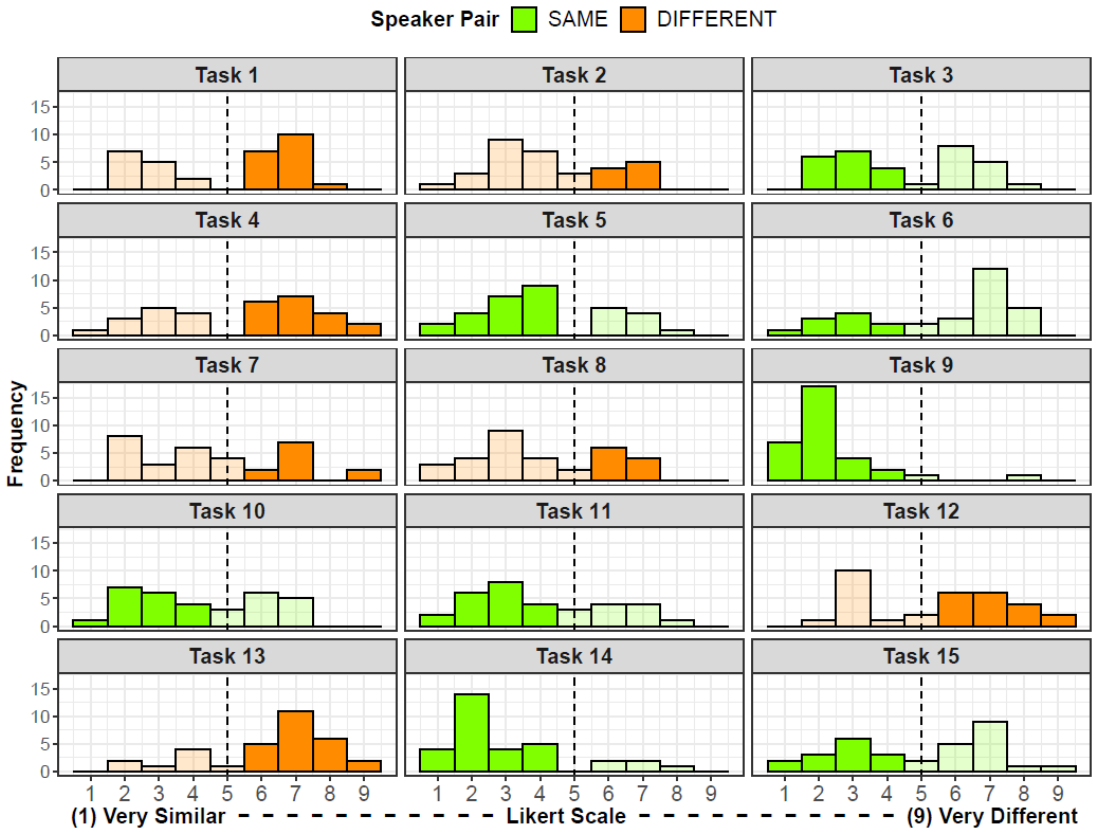


Figure 5.19. Histograms of the expert listeners' responses for Section Three of the experiment. Tasks with same-speaker pairs are depicted in green and tasks with different-speaker pairs are shown in orange. Full colour bars are indicative of correct responses and transparent bars indicate incorrect responses.

In isolating the expert listeners' responses as illustrated in Figure 5.19 above, the tasks which stood out in relation to the proportion of correct responses being recorded remain the same albeit the number of correct responses is more marked. For Task 9, the expert listeners recorded 30 out of 32 correct responses (93.8%) and for Task 14 recorded 28 out of 32 correct responses (87.5%). Task 2 was still the task with the fewest proportion of correct responses (9 out of 32 correct; 28.1%) followed closely by Task 8 and Task 6 which both had 10 out of 32 correct responses (28.1%). With Task 9 and Task 14 being the tasks with the most correct responses and given that both of these tasks contained same-speaker pairs, this was something which was investigated further with regards to whether tasks with same-speaker or different-speaker pairs yielded the best results. Considering just the expert listeners' responses,

same-speaker pairs recorded 158 out of 261 correct responses (60.5%) whereas different-speaker pairs recorded 109 out of 219 correct responses (49.8%). Furthermore, of the top five tasks with the most correct responses attributed to them, four of these were same-speaker pairs (Task 9 (WY177), Task 14 (WY109), Task 5 (WY042) and Task 11 (WY167)), whereas for the bottom five tasks with the fewest proportions of correct responses, three of the bottom five were different-speaker pairs (Task 2 (WY023/WY030), Task 8 (WY069/WY008) and Task 7 (WY067/WY069)).

How listeners used the Likert scale

As explained in Section 5.2.5, as a means of assessing how listeners/listener groups used the 9-point Likert scale (e.g., were the extremes of the scale being used by certain listeners/listener groups more than others?) and the degree to which listeners/listener groups made correct speaker identification assessments, the original ratings provided by listeners were recoded to give a value of [+4] – [-4] and then z-scored. Figure 5.20 shows the z-scores for each of the listeners (grouped by expertise). This figure allows for the inspection of individual listeners in relation to how they were using the 9-point Likert scale, whilst also providing a visualisation as to whether listeners were 'more correct' when assessing same-speaker pairs or different-speaker pairs. Figure 5.21 shows the z-scores for each of the listener groups. Results have been grouped according to whether they were obtained from tasks which contained same-speaker or different-speaker pairs.

Figure 5.20. Boxplots showing the z-scores for each of the listeners. Listeners have been grouped according to their expertise. Boxplots are ordered by overall mean score, from highest to lowest, left to right. Results have been grouped according to whether they were obtained from tasks which contained same-speaker or different-speaker pairs.

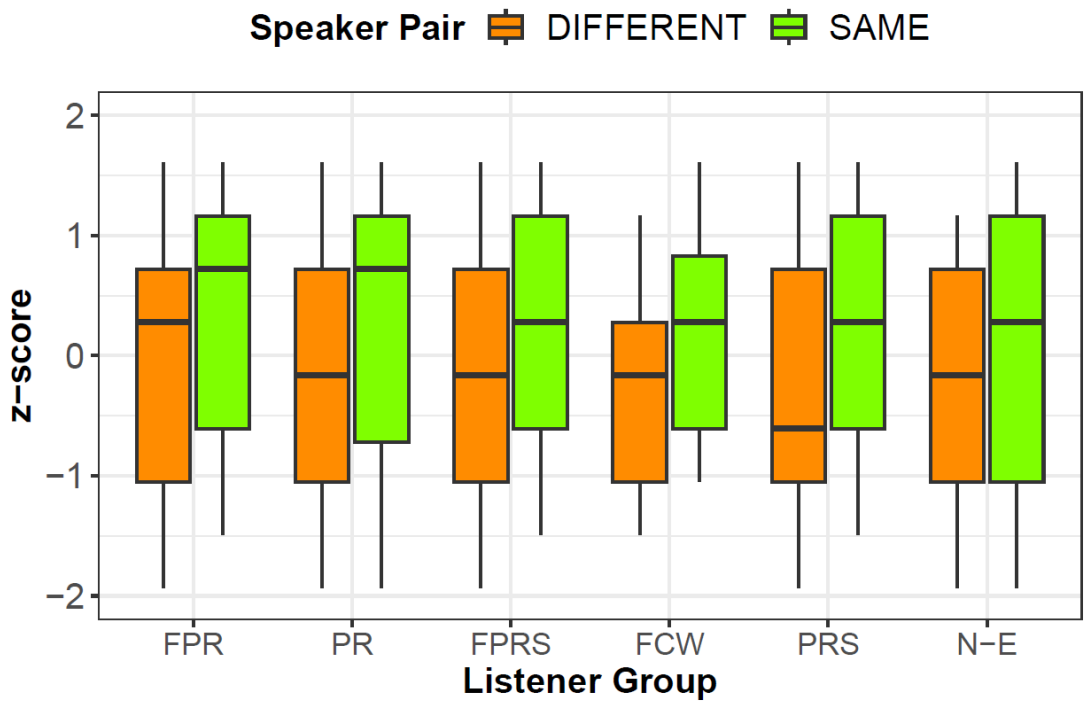


Figure 5.21. Boxplots showing the z-scores for each of the listener groups. Boxplots are ordered by overall mean score, from highest to lowest, left to right. Results have been grouped according to whether they were obtained from tasks which contained same-speaker or different-speaker pairs.

In examining the two figures above, it is evident that the vast majority of listeners performed better in tasks which contained same-speaker pairs. It is also apparent that where some listeners exhibit a good deal of variation with regards to the degree in which they recorded correct and incorrect responses (e.g., p116), other listeners recorded more consistent scores (e.g., p125).

Expert listeners vs. non-expert listeners

In order to assess whether the expert listeners with expertise in forensics performed significantly better than the non-expert listeners, the FCW group and the FPR group were combined into one group (for the remainder of this section referred to as 'forensic phoneticians'). This group now contained 13 listeners with the non-expert group also containing 13 listeners. Figure 5.22 Shows the percentage of correct responses for

both groups (n.b., the Likert scores obtained were simplified into a binary 'correct' or 'incorrect' identification as described in Section 5.2.5).

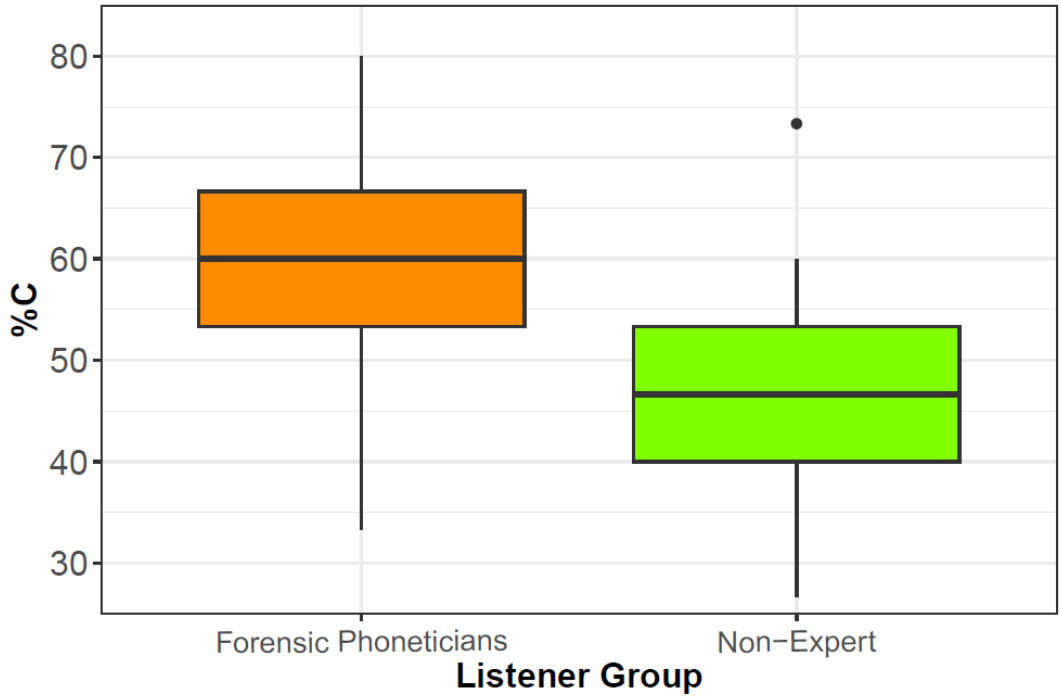


Figure 5.22. Boxplots showing the percentage of correct responses for the Forensic Phoneticians and the non-expert groups.

It can be observed from the boxplots that the forensic phoneticians have a higher percentage of correct responses than the non-expert listeners. In order to test whether this effect was statistically significant, a two-tailed t-test was performed which confirmed the significance of the finding ($t[23] = 2.11$, $p = 0.04$). The boxplots also reveal that there was a good deal of variation within both groups. Although there is only one outlier identified on Figure 5.22, with this being attributed to an apparent 'overachiever' within the non-expert group, a manual inspection of the data revealed that there were other individual results for both groups which were divergent from the majority of the others. In light of this, outliers were investigated using the Tukey outlier method (Tukey, 1977) which defines outliers as values that are more than 1.5 times the interquartile range (IQR) from the quartiles. This means that any data point that is below Q1 – (1.5 × IQR) or above Q3 + (1.5 × IQR) is considered an outlier.

Applying the Tukey method resulted in one outlier being identified from both the expert group and non-expert group (with both of these being listeners who achieved the highest scores). Table 5.6 Below shows the percentage correct (%C) means and standard deviations for both groups for the data when outliers are included and when they have been removed.

Table 5.6. Percentage correct means and standard deviations for the forensic phoneticians and non-expert listener groups.

	With Outliers		**Without outliers**	
% Correct	**Mean**	**Std. Dev.**	**Mean**	**Std. Dev.**
Expert	58.5	12.9	56.7	11.7
Non-Expert	48.2	10.8	46.1	8.4

When the two outliers are removed, and a further two-tailed t-test is conducted, the difference in the level of significance between the expert and non-expert groups is slightly higher ($t[20] = 2.43$, $p = 0.02$).

Also of interest was determining whether the forensic phoneticians were more confident than the non-expert listeners in making the discrimination decisions. Figure 5.23 shows how each of the groups made use of the 9-point Likert scale irrespective of whether or not they were correct in determining whether it was a same-speaker or different speaker pair.

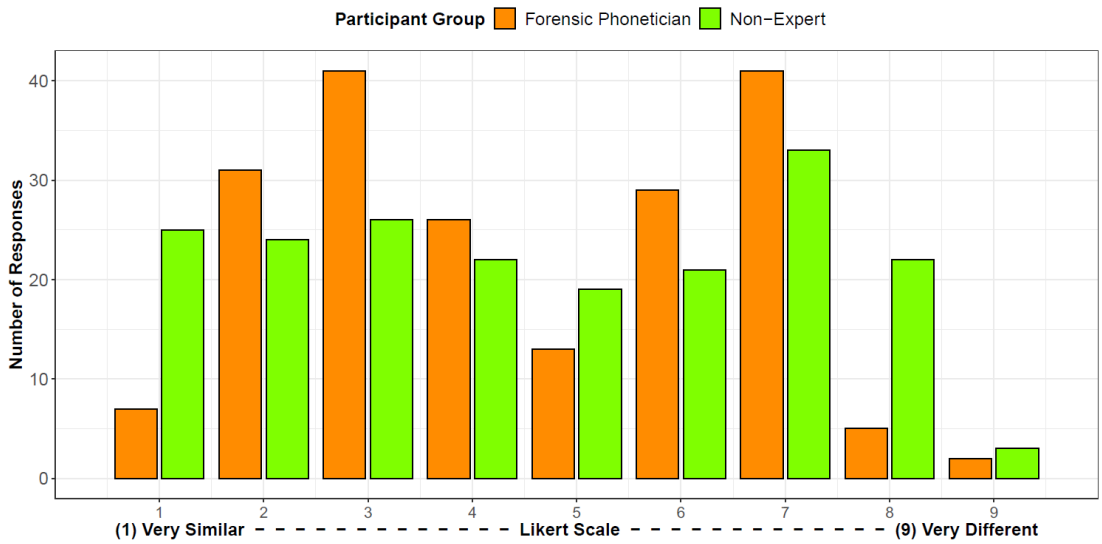


Figure 5.23. Bar plot showing how the forensic phoneticians and the non-experts made use of the 9-point Likert scale.

The above bar plot shows that, irrespective of 'correctness', the non-expert group made greater use of the extremes of the Likert scale than the expert group. This trend can be seen to be true both when rating speaker-pairs as very similar or very different (e.g., (1) rating used to a much greater extent by the non-expert group as was the case with (8) at the opposite end of the scale). If taking these ratings as an indicator of listeners' confidence in assessing whether tasks contained same-speaker or different-speaker pairs, this indicates that the non-expert group were more confident in their decision making. Figure 5.24 shows how the two groups used the Likert scale, however the results have been grouped based on whether listeners were correct in assigning a specific rating to a speaker pair.

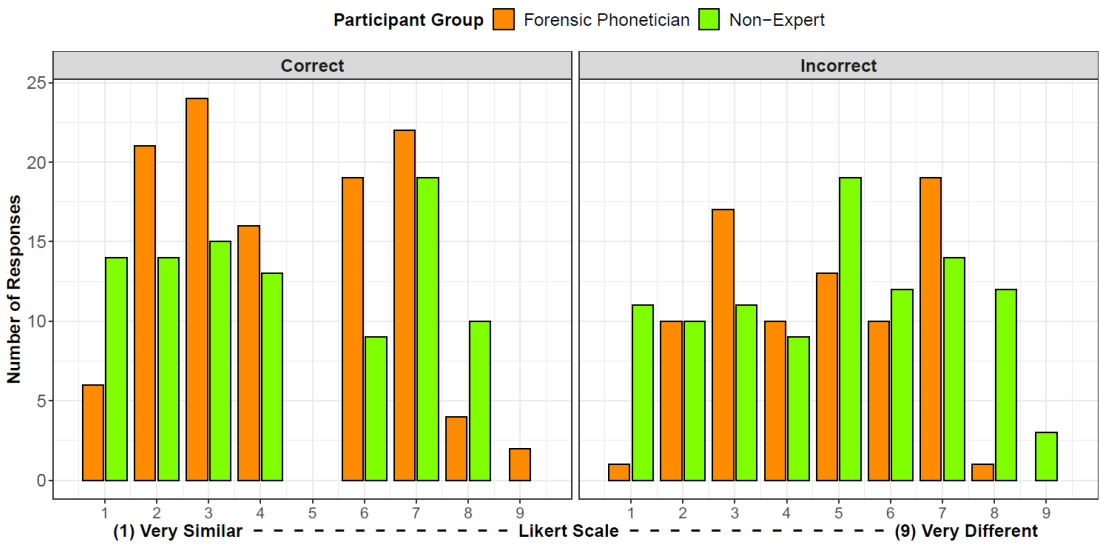


Figure 5.24. Bar plot showing how the forensic phoneticians and the non-experts made use of the 9-point Likert scale. Results have been grouped according to whether the ratings were correct (left panel) or incorrect (right panel).

Observing how the two groups made use of the scale when 'correctness' is factored in shows that the overall trend identified from the previous plot still stands in that it is the non-expert group who present as being more confident in their decision making. However, looking at the right-hand panel of Figure 5.24, it is evident that this apparent confidence is more frequently misplaced for the non-expert group in that they are seen to rate different-speaker pairs as being very similar (shown by using (1)), and rate same-speaker pairs as being very different (shown by using (9) and (8)). Overall, it is apparent that the non-expert listeners have a tendency to use the extremes of the Likert scale to a greater extent than the forensic phoneticians, with the expert group seemingly being more cautious when rating the pairs of the speech samples.

5.4. Discussion

5.4.1. Overview

This chapter presented the results from two perception experiments which sought to determine the extent to which listeners are able to discriminate between speakers

based on speech rhythm. The first of these experiments was the Pilot Study which assessed whether naïve (non-expert) listeners were able to make accurate speaker identification assessments when presented with delexicalised speech samples which foregrounded the rhythmic attributes of the speech signal. Results from this initial Pilot Study revealed that these non-expert listeners were able to make *some* accurate identification assessments, albeit these were mostly attributed to Section One of the experiment in which the tasks at hand were less complex. Section Two and Section Three of the experiment proved more challenging for these listeners with the proportion of 'correct' speaker identification assessments being substantially reduced.

The main perception experiment tested the ability of expert listeners (and a cohort of non-expert listeners) in making accurate speaker identification assessments following the same methodology as the Pilot Study. Expert listeners were grouped into five expert group categories according to their expertise and experience. Expert listeners performed better than non-expert listeners and experts who had expertise in forensic phonetics generally performed better than those experts who did not. The overall trend from the Pilot Study was mirrored in that the number of 'correct' responses generally declined as the listeners progressed through the three sections of the experiment. Across all three sections of the experiment, there were certain tasks which yielded markedly better proportions of correct responses than others, with the nature of the tasks in Section Three showing that same-speaker pairings of delexicalised samples yielded a greater proportion of correct identification assessments. Within Section Two of the experiment, listeners also provided qualitative feedback in which they explained why they had selected a specific sample as containing the same speaker as the original sample. These qualitative results highlighted a number of key features which listeners were tapping into when making (correct) speaker identification assessments.

The following subsections provide a greater depth of discussion pertaining to the results obtained from both of the perception experiments mentioned above.

5.4.2. Pilot Study

5.4.2.1. Section One

The first section of the Pilot Study was composed of eight tasks. Two listeners identified the correct speech sample in all of the tasks and an additional seven listeners identified the correct delexicalised speech sample seven out of eight times. One listener had recorded correct responses for six of the eight tasks and two listeners recorded correct responses for five out of eight tasks. The overall success of the listeners in making accurate speaker identification assessments here can be attributed to the nature of the tasks involved.

This section of the Pilot Study was designed to act as a 'training stage' for the listeners. Here it was possible for speakers to essentially 'match up' the original speech sample with the correct delexicalised sample as the correct sample was indeed the same stretch of speech which had been subjected to delexicalisation. Of the eight tasks, the proportion of correct responses was, on the whole, distributed evenly, except for one task (Task 2) which had an overall correct response rate of less than 50% (substantially less than the other seven tasks). The reason for this was discovered following the analysis of the results and was attributed to a delexicalisation error. This error took the form of the opening three syllables of the correct delexicalised sample being merged into one 'long syllable' (or schwa-like tone) as a result of pulses not being manually removed correctly during the delexicalisation process (see step (9) of Section 5.2.3). It is evident that this error occurring at the very beginning of the sample was a factor in seven of the 12 listeners selecting the incorrect delexicalised sample (the opening of this sample was not similar to the original sample). This is an indicator that, at least for this specific task, some listeners may have been using a strategy of 'matching up' the *openings* of the delexicalised samples to the openings of the original sample as opposed to considering the sample as a whole. Nevertheless, this initial section of the experiment seemed to fulfil its purpose in allowing listeners to adjust and become accustomed to delexicalised speech in preparation for the greater level of challenge which the following two sections posed.

5.4.2.2. Section Two

Quantitative Discussion

The challenge in this section of the Pilot Study was increased as listeners were no longer able to effectively 'match up' specific (identical) parts of the correct delexicalised sample with the original sample. Instead, listeners were now presented with the opening of an answerphone message (original sample) and two delexicalised samples taken from later on in an answerphone message recording, with one of these being the same speaker (from the same message) as the original. The results reinforced this added degree of challenge with the proportion of correct responses being significantly reduced in comparison to Section One. Although nine of the twelve listeners scored above chance level (50%) with regards to correct responses, the vast majority of these listeners only did so marginally (i.e., correct responses in five out of the eight tasks), with three of the listeners falling below chance level with correct responses for only three of the eight tasks.

Qualitative Discussion

Qualitative feedback was also obtained from this section whereby listeners were asked to provide an explanation as to why they selected the sample that they did. When making correct speaker identification assessments, by far the most frequently mentioned feature listeners commented on was pausing behaviour. The frequency, length, regularity, and placement of pauses were all commented on to some degree. The second most frequent feature speakers relied on was the tempo of the speech. Other specific speech features mentioned included pitch/tone, speech unit/syllable length, and the composition of phrases/chunks of speech. For example:

- '*variations in pitch*'
- '*sample 'X' is too monotone'*
- *'elongations for 'emphasis'*
- '*short bursts of words rushed out followed by up to a second of silence*'
- '*four words tend to be clustered then a slight pause*'

Some holistic assessments which commented on rhythm in a more general sense were also present. For example:

- '*continuous flow to the rhythm*'
- '*a constant beat*'
- '*more continuous speech*'
- '*sample 'X' sounds too disjointed*'

Given that the listeners in this Pilot Study were naïve listeners, much of the feedback obtained was expressed without the use of linguistic terminology as opposed to making direct reference to similar sounding patterns of intonation. For example:

- '*[sample] 'X' has the same up and down as the speaker*'

There were also some rather generic comments which do not offer much insight into what the listener was specifically focussing on when making their correct selection. For example:

- '*pace and tone*'
- '*tone and rhythm*'
- '*pauses and rhythm*'

Nevertheless, the qualitative feedback obtained from these non-expert listeners will allow for a comparison to be drawn with the expert listeners in the Main Experiment and provides a useful insight into what listeners are relying upon when assessing speaker similarity when presented with delexicalised speech.

5.4.2.3. Section Three

The final section of the Pilot Study is the section which the listeners found the most challenging. Eight of the listeners failed to record correct responses above chance level, with five of these falling below chance level. This section differed from the preceding sections in that tasks were composed of just two speech samples, both of which were delexicalised, and it was the task of the listeners to rate the similarity of the two samples with some tasks containing same-speaker pairs and other tasks different-speaker pairs.

Same-speaker pairs vs. different-speaker pairs

Of the eight tasks within this section, three contained same-speaker pairs and five contained different-speaker pairs. Of interest here is that the two tasks which had the highest proportion of correct responses attributed to them (Task 3 and Task 5) were both tasks which contained same-speaker pairings. In addition, listeners rated these same-speaker pairs correctly using the extremes of the 9-point rating scale with a rating of either (1) or (2) indicating that they found the pair of delexicalised samples to be 'very similar'. This contrasts with the correct responses for the different-speaker pairs as participants were less likely to make use of the extremes towards the 'very different' end of the scale when making correct speaker identification assessments. As this section did not ask for qualitative feedback from listeners with regards to the sample similarity rating they selected for each task, proposing reasoning as to why these same-speaker pairs were seemingly easier to attribute a correct response to is not a straightforward task.

On personal inspection of these two same-speaker pairings, it does manifest that the delexicalised pairings are similar in their make-up and this is attributed to a number of different features within the samples (e.g., the 'chunking' of phrases and phrase length, pause distribution, prolongations of specific syllables between pauses, quick bursts of speech units/syllables, intonation/pitch patterns). It is, however, the specific combinations of these features and their cooccurrences which lend to the overall perceived similarity of the samples. This trend of same-speaker pairings generating a greater proportion of correct speaker identification assessments is mirrored in the Main Experiment and further suggested explanations as to why this may be the case are put forward there (see Section 5.4.3.3).

A search of the literature reveals that attempting to compare these results directly to any previous research is not possible. There has been no previous research which has sought to test listeners' perceptions regarding the (dis)similarity of pairs of delexicalised speech samples. There has, however, been previous research by Nolan et al. (2011) which has asked naïve listeners to rate the (dis)similarity of paired same-accent voice samples on a nine-point scale. However, the purpose of this research was to determine which acoustic features (f_0, F_1, F_2 and F_3) correlated with perceived

voice (dis)similarity. The findings of this study showed that pitch (f_0) was the most important acoustic feature, confirming f_0's key role in voice similarity, followed by F_3, F_2, and F_1. Within the present study, the delexicalised samples were normalised to an average pitch of 100 Hz meaning that determining the similarity of samples based solely on pitch alone could not be used as an identification strategy.

5.4.3. Main Experiment

5.4.3.1. Section One

The initial section of the experiment served the purpose of proving a training stage for the listeners in which they could get accustomed to the nature of the delexicalised samples which they would be presented with throughout the remainder of the experiment. As listeners were faced with an original sample and two delexicalised samples, one being the same section of speech as the original, this section allowed for listeners to make direct comparisons between the original sample and the correct delexicalised sample (i.e., they could essentially directly 'match up' syllable patterns from the original sample to the correct delexicalised sample). It was therefore unsurprising that the results showed that 75% of the expert listeners recorded correct responses for all tasks within this section, with the remaining 25% of the expert listeners recording incorrect responses in only one or two of the tasks.

In relation to specific tasks which had incorrect responses attributed to them, unlike the Pilot Study in which there was one standout problematic task for the listeners (see Section 5.4.2.1), incorrect responses here were distributed over the five different tasks within the section (i.e., the problematic delexicalised sample from the Pilot Study was corrected for the Main Experiment).

Comparisons to previous research

The results obtained from this section of the experiment can be seen to mirror results from a study similar in design conducted by Remez et al. (1997). In the first of three experiments, naïve listeners were presented with a natural speech sample and two

sinewave samples. One of the sinewave samples was derived from the natural speech sample (i.e., the same speaker saying the same utterance) and the other sinewave sample was derived from a different speaker (saying the same utterance). Like the tasks in the present experiment, listeners were therefore faced with a binary decision, meaning that guessing would have produced results nearing 50%. The experiment was made up of ten speakers (and 13 listeners), and the overall results showed that on average listeners matched eight out of the ten natural samples to the correct sinewave sample better than chance.

Although there are some methodological discrepancies between the tasks in the present section of the present experiment and that of Remez et al.'s study (e.g., the nature and composition of the delexicalised/sinewave samples), both sets of results indicate that listeners are able to make correct speaker identification assessments when presented with replica delexicalised speech samples which are devoid of acoustic attributes of natural voice quality. In offering an explanation for their findings, Remez et al. suggest two possible interpretations. The first of these is that the correct delexicalised (sinewave) samples contained speaker-specific information in relation to the original (natural) samples despite the absence of the acoustic correlates of voice quality. The second explanation bares testament to the nature of the experimental design in that the sinewave samples were direct replicas of the original sample and therefore listeners could be making their decisions based upon more superficial auditory attributes that are not relevant to any speaker-specific characteristics (i.e., listeners could have made a correct sample match based upon similarities of the overall pitch, the duration of tones and the overall duration of the samples without actually perceiving any phonetic differences between the speakers).

The second of the explanations offered by Ramez et al. can be seen to align with the explanation put forward for the overall high proportion of correct responses in the tasks of Section One – that listeners could rely upon a strategy of directly 'matching up' specific patterns from the original sample to the correct delexicalised sample without having to focus on any speaker-specific acoustic characteristics. Nevertheless, this section of the experiment served its purpose as a training stage for listeners and demonstrated that expert listeners and non-expert listeners were able to disregard

dissimilarities between natural (original) speech samples and their delexicalised replicas (e.g., the absence of voice quality) and detect more abstract, phonetic similarities.

5.4.3.2. Section Two

Quantitative discussion

This section of the experiment increased the level of challenge posed to the listeners in comparison to the previous section and this was evidenced in the overall results. Overall, expert listeners outperformed non-expert listeners and those expert listeners who had expertise in forensic phonetics generally performed better than those who did not. Reasons for this latter trend could be attributed to the notion that those with expertise in forensic phonetics could be supposed to have engaged in speaker discrimination tasks and/or speaker discrimination research previously and would therefore be better equipped to discern between competing speech samples.

As already alluded to, this section posed a greater degree of challenge to the listeners. Listeners were no longer able to employ a strategy of 'matching up' direct patterns of syllables from the original message to the delexicalised samples. Rather, listeners had to discern the rhythmic patterning of the speaker in the original speech sample and then determine which of the two delexicalised speech samples (none of which were the same stretch of speech as the original) contained the same speaker as the original by drawing upon the rhythmic patterning present in the delexicalised samples. Other than the main trends mentioned above regarding the results, it is clear that some listeners were more adept at accomplishing these tasks successfully than others, and also that some of the tasks were easier for listeners than others (i.e., they had more correct responses attributed to them).

In relation to some tasks having more correct responses than others, this would indicate that the speaker(s) who featured in those tasks may potentially have comparatively marked idiosyncratic rhythmic patterning that made it relatively straightforward for listeners to correctly match the original sample with the correct delexicalised sample. On the other hand, it could have been that the foil delexicalised

sample presented alongside the correct delexicalised sample could have been markedly different from the original, or, in its own way, highly distinguishable and evidently not the same speaker as was in the original sample.

Comparisons to previous research

In comparing the findings of this section to previous research, it is again possible to draw some comparisons to the work of Remez et al. (1997). In the second of three experiments conducted, the methodological design bares similarities to the nature of the tasks in Section Two of the present experiment. That is, listeners were presented with a natural speech sample and two delexicalised (sinewave) samples – one of which was a sinewave sample produced by the same speaker and the other was a sinewave sample derived from a different speaker. In comparison to their initial experiment where the sinewave samples presented were derived from the same utterance as the original (natural) sample (i.e., a sinewave replica), in this experiment the natural samples used were different from the utterances that were used as models for the sinewave replicas. The results that they obtained were similar to those they obtained from their previous experiment – overall, listeners matched eight of the ten natural samples to the correct sinewave samples better than chance.

Making any direct comparisons between Remez et al.'s results and the results from Section Two is problematic given some rather substantial discrepancies in relation to the number of speakers and listeners involved, as well as the acoustic composition and duration of the delexicalised samples. However, in suggesting explanations as to why listeners were able to make correct speaker identification assessments despite the challenge brought by the experimental setup of this section, similarities can be drawn between the two studies. Remez at al. suggest that when correct assessments were made, this showed that listeners were able to register the phonetic properties of the natural and sinewave samples equally, and were able to compare them without recourse to the auditory correlates of the natural products of vocalisation. This is because the experimental setup dictated that the samples presented to listeners (i.e., the natural sample and two sinewave samples) were never derived from the same identical speech material.

The same premise can be proposed as an explanation for correct speaker identification assessments for the tasks of Section Two in the present experiment. That is, that when making correcting assessments, listeners were able to comprehend the rhythmic properties of the original speech samples and the delexicalised speech samples equally and were able to compare them without the possibility of making any comparisons in relation to lexical content. This conjecture, that the perceptual recognition of individual speakers can be supported by the rhythmic properties in a speech sample, although seemingly holding a degree of weight here, is something which can be subjected to more stringent testing. The reason for this is that if we were to accept this conjecture, testing the assumption about speaker identification would still rely on evidence that the comparison being made is attributed to a listener's implicit ability in discerning rhythmic variation specific to individual speakers. The final section of the experiment therefore provided further opportunity for this conjecture to be evidenced with listeners being tasked with comparing two delexicalised speech samples with no access to any natural speech material (see Section 5.4.3.3).

Qualitative Discussion

The qualitative feedback obtained from this section of the experiment provided a substantial amount of information as to why listeners selected a specific sample as containing the same speaker as the original sample. It was decided that the best way to tackle this vast amount of written feedback was to first assess the qualitative feedback in relation to when listeners made correct identification assessments.

The most frequently mentioned feature by listeners was the pausing behaviour evidenced across the samples which were being compared. Comments in this regard focussed mainly on the frequency, duration and distribution of pauses within the samples. For example:

- '*[b]oth […] seems to produce pauses before and after hesitation markers. These pauses are quite long*'

Similarly, the use of filled pauses were also a frequent focal point with listeners again drawing upon the frequency, duration and distribution of these items. For example:

- '*Bursts of fast speech with short filled pauses in between*'; '*but [the] flow is interrupted regularly by long filled pauses.*'

Other disfluency phenomena were also commented on albeit to a lesser degree with listeners reporting occurrences of word/part-word repetitions, hesitations, false starts and interruptions. For example:

- '*quite a lot of part-word/word repetitions*'
- '*The original speaker has a hesitant rhythmic pattern, i.e., pauses, hesitation markers, false starts and corrections*'

Articulation/speech rate was also a characteristic which listeners considered when making their assessments, with feedback focussing on the variability of articulation rate within a given sample and also the specific distribution of this variability. For example:

- '*short bursts of fast speech, followed by some slow speech*'
- '*The original speaker varies between phases of slow articulation and phases of fast articulation*'

Listeners also made use of pitch and intonation patterns in terms of the range and variability of intonation used, with some listeners picking up on idiosyncratic patterns pertaining to the use of high rising terminals as well as other noteworthy intonation patterns. For example:

- '*Some high rising intonation patterns and fairly staccato style*'
- '*rising intonation at end of some phrases*'
- '*falling intonation through utterances*'
- '*short rising intonation in the first 10 seconds*'
- '*long filled pauses with a falling intonation*'
- '*Boundary tones are either level or rising*'
- '*sample X contains slightly more pitch/amplitude modulation than sample X*')

Listeners also paid attention to specific syllables/speech units and the grouping of syllables/chunking of intonation phrases. For example:

- '*some prolonged vocalic components*'

- '*prolongations at start and end of utterance*'
- '*Bursts of fast speech with short filled pauses in between*'
- '*speech divided into linguistic chunks, e.g., intonation units*'
- '*Sample X has a less staccato rhythmic pattern and similar elongated syllables at the end of intonation units*'
- '*Tendency for shorter phrasal groups, with elongations towards the final of the phrase*'

It should also be noted that listeners at times chose to provide more holistic comments which grouped a number of these features together as a means for making their correct speaker identification assessment. For example:

- '*The original message has quite a disfluent rhythmic pattern (pauses, hesitation markers, prolongations), resulting in alternating fast and slow passages*'
- '*Sample [X] has a less staccato rhythmic pattern and similar elongated syllables at the end of intonation units*'
- '*I paid attention to the occurrence of silence and hesitations*'

Comparisons to previous research

A search of the literature indicates that attempting to make comparisons with regards to the feedback obtained from this section of the experiment with previous research would be a fruitless endeavour due to the novel experimental design and purpose of the present experiment. In relation to all of the features mentioned by listeners, each of these has, to a greater or lesser extent, been subject to previous forensically-motivated research. The vast majority of this research has been in the form of production experiments in which a given feature has subsequently been shown to show a certain degree (whether high or low) of speaker-specificity, within/between-speaker variation, or speaker discriminatory potential.

Forensically-motivated perception experiments are more sporadic within the literature with the majority of these involving tasks which require listeners to rate the (dis)similarity of (usually very short) speech samples in which voice quality is present (i.e., without the inclusion of delexicalised speech samples). Examples of such work

include Bartle and Dellwo (2015) who assessed listeners' (expert and non-expert) ability to discriminate between speakers in short utterances in voiced and whispered speech, and McDougall (2013) who tasked naïve listeners with rating the similarity of pairs of short speech samples on a scale of 1 (very similar) to 9 (very different). There has, to date, only been a handful of studies with potential forensic implications which have made use of different types of delexicalised speech samples, with these being reviewed in Chapter 2, Section 2.5.4. There is a body of earlier perception research, without forensic motivation, which has sought to assess which factors play a part in listeners' ability to distinguish between speakers when natural voice quality has been removed from the speech signal. Van Dommelen (1987) assessed the role of speech rhythm, intonation and pitch in paired speaker identification tasks and found that listeners made use of all three parameters to varying degrees depending on the speech sample, indicating that the relevance of these parameters for speaker discrimination is speaker-dependent rather than absolute.

Despite the fact studies such as those mentioned above bare relevance to the present experiment, albeit in different ways, it remains that none of these experiments required respondents to provide qualitative feedback. Therefore, determining what specific features listeners were actually attending to is not possible. What is apparent after close scrutiny of the qualitative feedback is that some listeners provided much more detailed comments than others, and in such circumstances, it was often the case that the listeners would comment on multiple features when giving their explanations. That is, they would take a holistic approach and consider all of the notable rhythmic features from the original speech sample and then try and marry up as many of these features to those present in the delexicalised samples. For example:

Example 1

Participant ID: p100 (FCW)

"original speaker = nothing particularly distinctive, a bit of everything - silent pauses, filled pauses, prolongations, false starts; did not pick sample A as this included many relatively long silent pauses; also sample B contains slightly more pitch/amplitude modulation than sample A and thought that better matched original sample"

Example 2

Participant ID: p134 (PRS)

"The original speaker holds stressed syllables longer than unstressed. The speaker has more frequent and steeper changes in pitch, making his speech sound bouncy."

Example 3

Participant ID: p108 (FPRS)

"Speaker in Answerphone Message 4 used loads of fillers, had a relatively slow speech rate and sounded relatively monotonous. […] Sample_B sounded like there was much more pitch movement and the speech rate sounded quicker with somewhat less/shorter fillers, so I chose Sample_A"

Example 4

Participant ID: p109 (FCW)

"original speaker = few hes/pauses, no marked prolongation; sample B chosen as two long pauses in middle, like in original sample; also sample A appears to contain prolongations and pitch variability which hasn't been of note in original sample."

Example 5

Participant ID: p115 (FCW)

"AM6 - v. short utterances / frequent mid-utterance pausing. falling intonation through utterances. A - staccato-y rhythm, prolonged onsets to utterances (or longer than within utterance) | B - v. short utterance units - frequent breaking of turn. B more similar to AM6 based on the features described."

On the other hand, some listeners would look to hone in on one or two particularly distinctive features within the original sample and look to compare that with the delexicalised samples. For example:

Example 6

Participant ID: p127 (FPRS)

"The larger pitch range was more similar to Sample A"

Example 7

Participant ID: p132 (FPR)

"pattern of hesitations and silences"

Example 8

Participant ID: p112 (FPRS)

"The original seems to have short intonation phrases, separated by unfilled pauses. This also seems to be the case in Sample_B. The intonation phrases in Sample_A seem to be much longer in comparison."

Considering the speech material used

It is worth highlighting here, after having described the features which were commented on the most by the listeners, that the nature of the speech samples will have had an impact with regards to which features were most predominant in the records. That is, that the stimuli were voicemail messages in which the speakers had no interlocuter, rendering the samples to be essentially monologues. It would therefore be expected that phenomena such as silent pauses and filled pauses would readily feature within these recordings than they would in, for example, a telephone conversation (e.g., turn-taking would be taking place in which any marked (silent) pause by a speaker would likely indicate the end of a conversational turn rather than a unique/idiosyncratic rhythmic strategy used to segment intonation phrases). Therefore, it is acknowledged here that listeners' qualitative comments on the speakers' rhythmic patterns may not be wholly representative of what the speakers' conversational rhythm patterns look like.

Nevertheless, utilising this type of stimuli can be seen as advantageous for the very same reason. In having no interlocuter, speakers would not have the potential to accommodate to any interlocuter and therefore the speech rhythm patterns exhibited by the speakers can be expected to be true to the individual (at least within the speaking style). In addition, the lack of a conversational partner reduces the likelihood of any drastic change in a speaker's rhythm patterns as there will be no forced change of topic and no forced change in emotion meaning that the rhythmic patterns exhibited are likely to be more consistent (as opposed to conversational dialogue).

It might be the case that the nature of the stimuli (e.g., containing just one speaker) has led to features such as silent pauses and filled pauses being frequently mentioned. Nevertheless, the fact that it is features such as these being commented on provides further justification and merit for carrying out perception experiments like the one currently at hand. This is because features such as these would not normally be accounted for in speech rhythm production experiments. For example, in the case of silent pauses, there are obviously no acoustic correlates to measure, and in the case of filled pauses, the vast amount of previous speech rhythm research would omit these from analysis (for the want of 'pure' speech material). It could be that the qualitative feedback obtained from this section of the experiment provides some support for the proposition that if speakers do use idiosyncratic speech rhythm patterns, then perhaps more is to be gained in understanding, capturing and describing these patterns if we turn our attention towards the silences and the subconscious, unplanned disfluency behaviour of the individual.

Overall, it appears that listeners employed different strategies in completing these speaker discrimination tasks and had varying degrees of success with regards to making correct identification assessments. Where previous work has shown that an individual's pitch is a predominant factor in speaker discrimination tasks (e.g., Foulkes & Barron, 2000; McClelland, 2008; McDougall, 2013; Nolan et al., 2011), this parameter was normalised to 100 Hz in the present experiment and therefore listeners were unable to rely solely on pitch as a strategy for speaker identification.

Those expert listeners with experience in forensic phonetics, especially the forensic caseworkers, will have been exposed to speaker discrimination tasks previously. In such tasks, these individuals would likely use a range of analytical techniques such as auditory phonetic analysis of voice quality, pitch, intonation, segmental features, articulation rate, speech fluency, as well as acoustic phonetic analysis (e.g., of spectrograms, fundamental frequency, etc.). The naïve listeners in this experiment (and the Pilot Study) have no training in any kind of speech analysis and therefore the decisions they made with regards to speaker identification assessments were to a large degree non-analytical. In a similar way, those expert listeners without expertise in forensic phonetics and voice comparison examinations would likely have less explicit

and direct experience in comparing speech samples in this way. This goes some way to providing an explanation as why the overall trend of forensic experts > non-forensic experts > non-experts was evidenced in the present experiment.

The length of the speech samples in all sections of the experiment, being 30 seconds, were much longer than stimuli generally presented to listeners in previous speech perception research. In addition, listeners were able to listen to all of the stimuli as many times as they wished (pause, replay, rewind, fast-forward, etc.) which is in contrast to previous studies in which speech samples are usually only played once (or perhaps twice) before the listeners are automatically moved on to the next task. Both of these factors were a necessity in the present experiment given that listeners were instructed to focus on the rhythmic characteristics of the speech samples and thus to assess and make discrimination decisions based on the speech rhythm patterns they perceived. This is not something which a listener could accomplish through a 3-second or even 10-second speech sample, given that speech rhythm is something which is established through the interrelations and accumulative build-up of different prosodic attributes over a stretch of an individual's speech.

Within this section of the experiment, it could be argued that having three 30-second speech samples (one original, two delexicalised) to contend with in each task could be overly laborious for listeners and that having 15 tasks of the same nature to complete could lead to participant fatigue and potentially reduced performance. This is something which was not evident in the results however as there was not a decrease in the overall proportion of correct responses as listeners worked their way through the tasks in Section Two.

5.4.3.3. Section Three

The final section of the experiment was ultimately the most challenging for both the expert and non-expert listeners overall. Although only a minority of listeners performed better in this section than they did in the previous section, the overall trend evidenced was that ranking the (dis)similarity of two delexicalised speech samples was a more arduous task than identifying which of two delexicalised samples

contained the same speaker as an original sample. One reason for this could be that in the previous two sections of the experiment, having access to an original (natural) speech sample served as a focal point for listeners – that is, listeners were better equipped to discern rhythmic attributes from natural speech than they were from delexicalised speech samples. Another reason could be that listeners were able to focus in on the rhythmic patterning of specific syllables/speech units when afforded a natural speech sample and this made identifying similar patterns within the delexicalised samples an easier task.

For example, in the previous section of the experiment (Section Two), it could have been possible that some speakers demonstrated marked/idiosyncratic patterning in using filled pauses (e.g., being prolonged, having falling pitch, etc.), or that they would have other noteworthy dysfluency patterns such as starting intonation units with false starts (e.g., multiple word/part-word repetitions). Phenomena such as these occurring within the original speech sample, could arguably be identified by some listeners more easily in contrast to other stretches of more fluent speech, as the nature of the delexicalised samples being syllabic schwa-like tones would not be too dissimilar from the rhythmic patterning of the original sample.

Same-speaker pairs vs. different-speaker pairs

As was apparent in the previous section of the experiment, there were some tasks in which listeners performed better than others. Tasks which contained same-speaker pairs generally resulted in a higher proportion of correct responses being obtained by the listeners (a trend which was also evidenced in Section Three of the Pilot Study). It was also shown that tasks which contained same-speaker pairings would also see listeners present as more confident in their responses – that is, they would more readily utilise the extremes of the 9-point Likert scale (i.e., (1) or (2)) to indicate that the delexicalised pair of samples sounded ‘very similar’ in relation the rhythmic characteristics they possessed. This was in contrast to correct speaker identification assessments that were made for different-speaker pairs where listeners would be more conservative in their use of the Likert scale and not use the extremes of the ‘very different’ end of the scale (i.e., (8) or (9)). One reason for this could be that those

same-speaker pairs which listeners rated as being very similar did in fact exhibit marked idiosyncratic behaviour with regards to the rhythmic attributes of the delexicalised samples. As there was no qualitative feedback obtained from this section, it is not possible to determine what resulted in listeners being more confident in rating same-speaker pairs as sounding 'very similar'.

A search of the literature reveals very little in the way of aiding with a direct explanation for this apparent 'same-speaker preference'. It is, however, possible to more broadly link a body of psychology research to this finding – that being that research has shown that it takes less time to determine that two stimuli are the same than to determine that two stimuli are different. Although the stimuli within such experiments have been wide-ranging, from alphanumeric stimuli to non-linguistic visual patterns, there have been some studies which have found the same results with words (Fraisee, 1970; Johnson, 1975; Smith, 1967). Although this body of research is concerned with the time it takes to make a judgement call on the (dis)similarity of stimuli, a factor which was not the focus of the present experiment, the fact that respondents take less time in determining that two stimuli are the same would indicate that making such a judgement call is therefore 'easier' than the alternative (i.e., determining that two stimuli are different). If this is the case, then it could also be presumed that respondents would therefore also be more confident when making their 'same-stimuli' decision. Although the findings of this psychology research do not directly map onto the present experiment, it is possible that there could be some link between the two, in that respondents are predisposed to being 'better' at determining whether patterns, be it in terms of speech rhythm patterns or otherwise, are similar to one another more so than if they are different from each other.

Forensic experts vs. non-experts

This section of the experiment also took a closer look at whether there was a significant difference in performance between the expert listeners with expertise in forensic phonetics (FCW and FPR groups; 13 listeners) and the non-expert listeners (13 listeners). Results showed that there was a significant difference in performance

between the forensic phoneticians and the non-expert listeners in terms of the percentage of correct speaker identification assessments, with the forensic phoneticians being the better performing – a finding in fitting with previous forensic research (Bartle & Dellwo, 2015; Hollien, 2002).

Although it was shown that the expert listeners performed better as a group overall that the non-expert listeners, there were some notable outliers from both groups of listeners. For the forensic phoneticians, there were two individuals who performed comparatively poorly to the rest of the group and for the non-experts there was one outlier at each end of the performance scale – an 'underachiever' and an 'overachiever'. With regards to those individuals who performed poorly, it is not apparent as to why this was the case. Controlling for listener motivation in taking part in the experiment (other than the monetary incentive) was not factored into the pre-experiment questionnaires, and it could therefore be possible that some listeners simply didn't try as hard as other listeners. On the other hand, it could also be the case that some listeners are simply just not as proficient at these types of listening tasks as others. In the case of the underachieving forensic phoneticians, it could also be possible that, although they would have likely faced similar speaker-discrimination tasks previously, they would look to go about their decision making in a more analytical way (e.g., with more rigorous analysis not afforded by the present experimental design). In accounting for the 'overachiever' in the non-expert group, it could be that this individual is especially good at this type of listening task and is able to discern and make accurate similarity ratings when presented with delexicalised speech samples. However, in assessing this individual's performance in the previous section of the experiment (Section Two), their performance was actually the worst of all the expert and non-expert speakers who participated. It is therefore more probable that this individual exercised some noteworthy guesswork when rating the similarity of the speech samples in this section (or they simply had a torrid time with the tasks in Section Two).

Finally, it was also possible to investigate whether the forensic phoneticians (FCW and FPR groups) were more or less confident than the non-expert listeners when making their speaker identification assessments. Although 'confidence' was not

accounted for explicitly, the 9-point Likert scale which listeners used to either rank pairs of samples as very similar (1) or very different (9) can be transposed onto degrees of confidence – that is, a rating of (1) or (2) would be (very) certain of a same-speaker pairing; (3) and (4) would be (semi-)sure of a same-speaker pairing; (9) and (8) would be (very) certain of a different-speaker pairing; and (7) and (6) would be (semi-)sure or a different-speaker pairing (a rating of (5) would indicate that the listener could not discern whether the two samples sounded similar or different from one another and would therefore be interpreted as an incorrect response).

It would seem a logical supposition that the forensic phoneticians would be more confident when rating the speech samples given that their experience and training. However, it is this very experience and training which accounts for their awareness as to the variability of speech in general and would alert them to fact that the samples that they are dealing with only contain limited speaker information (e.g., being devoid of voice quality). Moreover, within FVC casework, it would be extremely unlikely that a forensic expert would ever make a categorical decision with regards to speaker identity (French et al., 2007). Factors such as these could therefore see the expert listeners be more conservative when making their (dis)similarity ratings. The results showed that when accuracy was not a factor it was indeed the non-experts who presented as more confident when rating the speech samples both with regards to determining that pairs were very similar (1) - (2) or very different (9) - (8) and that the forensic phoneticians were more cautious in their judgements.

When the results were analysed with regards to ratings that were correct versus ratings that were incorrect, some further interesting trends emerged. In relation to correct ratings, the forensic phoneticians remained, overall, more cautious than the non-experts when correctly identifying that samples were same-speaker or different-speaker pairs. Where the non-experts made greater use of the (1) on the scale when expressing that pairs of samples sounded very similar, the forensic phoneticians made greater use of the (2) and (3) on the scale. When expressing (correctly) that pairs of samples with different-speaker pairs, the forensic phoneticians were somewhat more cautious and opted to predominantly use (7) and (6) on the scale in comparison to the non-experts who favoured ratings of (8) and (7). For the incorrect ratings, the forensic

phoneticians were again more cautious predominantly using (3) for (incorrect) similar ratings and (7) for incorrect different ratings. This is in comparison to the non-experts who used the extremes of each end of the scale to a much greater extent – that is, (1) for (incorrect) similar speaker ratings and (9) and (8) for (incorrect) different speaker ratings. This indicates that when listeners recorded incorrect responses, the non-expert listeners were incorrect to a greater extent than the forensic phoneticians.

5.5. Chapter summary

This chapter has reported the results from two speech rhythm perception experiments, the first of these being the Pilot Study and the second being the Main Experiment – a modified and extended version of the former. Results from the Pilot Study showed that naïve listeners were able to make *some* correct speaker identification assessments, however, as the level of challenge increased through the different sections of the experiment, the performance of these non-expert listeners declined. When required to give qualitative feedback for Section Two of the experiment to explain why they had selected a given sample, these listeners commented on a number of different speech features as being relevant in making their decisions. As could be expected, these features were only referenced in a general sense – that is, absent of any fine-grained specific detail and without the use of specific terminology. In the final section of the Pilot Study, when tasks required listeners to rate the (dis)similarity of pairs of delexicalised speech samples, listeners performed better in tasks which contained same-speaker pairs.

The Main Experiment followed the same format as the Pilot Study, however, both expert listeners and non-expert listeners participated. Overall results from across all sections of the experiment showed that expert listeners performed better than non-expert listeners, and those expert listeners with expertise in forensic phonetics performed better than those who did not. Results from the first section of the experiment saw both expert and non-expert listeners record high proportions of correct responses. The high success rate in identifying the correct speech samples in this section was attributed to the fact that it was possible for listeners to employ a strategy

of making direct comparisons (i.e., 'matching up' specific patterns) between the original (natural) sample and the correct delexicalised speech sample.

In Section Two of the experiment, the proportion of correct responses across all groups of listeners declined as the level of challenge involved in the tasks increased. Here, listeners were no longer able to directly 'match up' specific patterns when arriving at a decision as to which delexicalised sample contained the same speaker as the original sample. The qualitative feedback obtained from the expert listeners in this section was far more detailed than that obtained from the Pilot Study. The expert listeners made reference to a variety of speech rhythm characteristics and explained the combinations of features which they relied on in specific tasks when making their speaker identification assessments. The qualitative feedback generated here will be used in developing meaningful descriptors of speech rhythm which will feed into the development of a perceptual rhythm framework for forensic speech analysis (see Chapter 6). The exact weighting which specific features had when listeners made correct identification assessments was not clearcut, however, what is clear from the qualitative feedback is that listeners used different strategies when discriminating between speakers, and that some of these strategies were more effective than others.

In the final section of the experiment, listeners were required to rate the (dis)similarity of pairs of delexicalised speech samples. This section of the experiment proved the most challenging overall for all of the listener groups. One reason proposed for this was that listeners had no access to an original (natural) speech sample from which to use as a starting point when assessing specific rhythmic characteristics. That is, that it was easier for listeners to discern rhythmic attributes from a natural speech sample and then map these onto a delexicalised sample, rather than having to decipher rhythmic patterns from two delexicalised samples. One trend that reemerged from the Pilot Study here was that listeners performed better in tasks which contained same-speaker pairs as opposed to different-speaker pairs. Arriving at an explanation as to why this was the case was not straightforward given the lack of any directly relevant prior research in this area. Instead, this finding was broadly linked to a pool of psychology research which demonstrated that respondents are more confident in

determining when two stimuli are the same rather than when two stimuli are different and that this could be an innate predisposition.

This final section of the experiment also took a closer look at whether there was a significant difference between the forensic experts and the non-experts in their ability to discriminate between speakers, with the results showing that the forensic experts were indeed significantly better than the non-experts. This result is something which would be expected given the experience and training the forensic phoneticians would have received and the fact that they would have likely encountered speaker discrimination tasks similar to this one before. That is, it is likely that the forensic experts went about the tasks in a specific analytical way, drawing upon their experience, and employing specific strategies when making their assessments, whereas the non-expert listeners would be using an essentially non-analytical approach.

Finally, it was shown that the forensic phoneticians, despite being significantly better than the non-experts at the tasks in this section, were more cautious in their responses. This finding is explained by the fact that these experienced forensic experts are more aware of factors such as within-speaker variation, and that, within real FVC casework scenario, an expert would rarely, if ever, make a categorical decision with regards to speaker identity.

The overall results from the experiment suggest that listeners are, to varying degrees, able to discriminate between speakers based on speech rhythm and that the development of a perceptual rhythm framework for forensic speech analysis could be a useful tool for forensic practitioners within forensic voice comparison cases.

CHAPTER 6

A Perceptual Rhythm Framework for Forensic Speech Analysis

6.1. Introduction

The production experiments carried out in the present thesis (Chapter 3 and Chapter 4) yielded results with varying degrees of speaker-discriminatory potential. The results obtained from the content-mismatched, spontaneous speech data (Chapter 3) were, on the whole, relatively weak. Overall, it was shown that the acoustic methods used to capture speech rhythm patterns were too sensitive to the variation that spontaneous, content-mismatched speech contains. The chapter which followed (Chapter 4), which used the same acoustic methods to analyse the rhythmic characteristics of a set of, so-called, "frequently occurring speech units", produced more promising results. That is, in comparison to the results from Chapter 3, the speaker-discriminatory potential observed when assessing these speech units was much improved.

Considering the results of Chapter 3 and Chapter 4 together, it was concluded that attempting to measure speech rhythm for forensic purposes (i.e., for speaker discrimination purposes when forensically realistic speech data is being used) using acoustic methods is for the most part untenable. That is, the acoustic complexity of speech rhythm (i.e., the numerous speech parameters involved in its makeup and the and the interrelations between these parameters) was deemed too susceptible to the

type of speech data encountered in forensic scenarios (e.g., content mismatch, degradations in recording quality, etc.). Furthermore, it is highly suspected that these acoustic methods may fail to capture some rhythmic nuances, and it is plausible that perception could be used as a mechanism to draw out further relevant rhythmic details. The results obtained from these production experiments, however, have served to highlight the necessity for forensic phonetics research to create and evaluate methodologies that are specifically tailored to the analytical tasks encountered by forensic practitioners.

Within FVC casework, voice quality is one speech feature which is analysed almost exclusively by perceptual means. It is highly unlikely that a forensic analyst will conduct an acoustic analysis of voice quality characteristics (e.g., measuring spectral tilt and additive noise parameters) owing to how sensitive these parameters are to any sort of degradation in recording quality. Furthermore, the perceptual analysis of voice quality requires the analyst to make both componential observations and Gestalt observations. When carrying out their perceptual assessment of voice quality, forensic practitioners will often make use of a recognised methodological approach such as the Vocal Profile Analysis scheme (Laver, 1980). This scheme has been refined through forensic research efforts that have served to improve its effectiveness in FVC cases (e.g., San Segundo et al., 2019; San Segundo & Mompean, 2017; San Segundo & Skarnitzl, 2021). Taking inspiration from such research, along with acknowledging the limitations of assessing speech rhythm through acoustic means, it was decided that the present thesis should direct its focus to strengthening the auditory analytical potential of rhythm as a speech analysis feature.

As such, perception experiments were carried out in Chapter 5 which sought to assess the extent to which listeners (both expert and non-expert) could differentiate between speakers when presented with just the rhythmic attributes of speech. One of the main outcomes of the perception experiments from the previous chapter was the vast amount of qualitative feedback obtained from both expert and non-expert listeners. Whil the previous chapter highlighted the rhythmic features which listeners were supposedly making reference to when making their speaker identification assessments, the present chapter looks to use this qualitative feedback as the basis for

the development of meaningful descriptors of speech rhythm – these could then feed into a perceptual rhythm framework for forensic speech analysis. The development of such a framework could be a useful tool for forensic practitioners given that within the auditory-phonetic and acoustic approach to forensic voice comparison there is currently no structured framework analysts can use to effectively account for speakers' speech rhythm patterns.

As a means of illustrating why a speech rhythm framework for the purposes of forensic speech analysis could be desirable to the forensic expert within FVC cases, the present chapter begins by providing an overview of some of the frameworks which currently exist. These are frameworks/methodologies which have been specifically designed (or modified) for their implementation and application within the forensic domain.

Providing a description as to how and why these frameworks have been both developed and tested will therefore serve as a useful starting point for the proposition of a perceptual rhythm framework for forensic speech analysis.

6.2. Existing frameworks

The following subsections describe a number of existing frameworks which have been developed and tested for application within forensic speech analysis procedures. The frameworks which will be discussed are as follows:

- **VPA:** Vocal Profile Analysis (Laver, 1980)

Also, two modified variants:

  - SVPA – Simplified Vocal Profile Analysis (San Segundo & Mompéan, 2017)
  - MVPA – Modified Vocal Profile Analysis (San Segundo et al., 2019)

- **TOFFA:** Taxonomy of Fluency Features for Forensic Analysis (McDougall & Duckworth, 2017)

Also, one modified variant:

- TOFFAMo (Carroll, 2019a, 2019b, 2019c)

- **PASS:** Phonetic Assessment of Spoofed Speech (Lee et al., 2023)

6.2.1. VPA (Vocal Profile Analysis scheme)

The VPA (Laver, 1980) is one of the most widely used methodological frameworks for the componential assessment of voice quality (henceforth VQ). VQ can be defined as the combination of long-term, quasi-permanent laryngeal and supralaryngeal features and their associated perceptual effects, with this definition holding that each of the speech organs has influence over a speaker's VQ. Within forensic forensics, particularly within FVC, VQ is a feature which is frequently analysed (Nolan, 2005) given its important role in speaker identification (Laver, 1980). Indeed, in an international survey on FVC practices, 94% of respondents stated that they examine VQ, with 61% of those doing so using a recognised scheme such as the VPA (Gold & French, 2011). In fact, within the forensic domain, it is more likely a practitioner will make use of a modified variant of the VPA for reasons which will be explained below.

The VPA scheme was developed by John Laver and colleagues (1980) for use within the clinical setting. Its purpose was to allow clinicians to obtain a comprehensive overview of the characteristics of a voice and, more specifically, to provide a means of characterising different forms dysphonia, quantifying their severity, and providing a basis for planning and monitoring therapy. As such, the original VPA scheme is renowned for its comprehensiveness and exhaustiveness with regards to the physiological detail it can capture. This is exemplified in one of the most common versions of the framework (Beck, 2007) in which there are a total of 36 settings (i.e., VQ features) which can be assessed: 25 describe vocal tract (supralaryngeal) features, 7 describe phonation features, and 4 describe overall muscular (laryngeal and vocal tract) tension features. This version of the VPA also includes some extra features

relating to prosody and temporal structure. Figure 6.1 below shows the original version of the VPA scheme (extra features excluded).

Vocal Profile Analysis (VPA)

	First Pass		Second Pass						
				Moderate			Extreme		
	Neutral	Non-Neutral	Setting	1	2	3	4	5	6
A. Vocal tract features									
1.Labial			Lip rounding/protrusion						
			Lip spreading						
			Labiodentalization						
			Extensive range						
			Minimized range						
2. Mandibular			Close jaw						
			Open jaw						
			Protruded jaw						
			Extensive range						
			Minimized range						
3. Lingual tip/blade			Advanced tip/blade						
			Retracted tip/blade						
4. Lingual body			Fronted tongue body						
			Backed tongue body						
			Raised tongue body						
			Lowered tongue body						
			Extensive range						
			Minimized range						
5. Pharyngeal			Pharyngeal constriction						
			Pharyngeal expansion						
6. Velopharyngeal			Audible nasal escape						
			Nasal						
			Denasal						
7. Larynx height			Raised larynx						
			Lowered larynx						
B. Overall muscular tension									
8. Vocal tract tension			Tense vocal tract						
			Lax vocal tract						
9. Laryngeal tension			Tense larynx						
			Lax larynx						
C. Phonation features									

		Present		Scalar Degree					
				Moderate			Extreme		
	Setting	Neutral	Non-Neutral	1	2	3	4	5	6
10. Voicing type	Voice								
	Falsetto								
	Creak								
	Creaky								
11. Laryngeal frication	Whisper								
	Whispery								
12. Laryngeal irregularity	Harsh								
	Tremor								

Figure 6.1. The VPA scheme adapted from Beck (2007). Shaded cells mean that the corresponding setting does not admit the specified degree(s) or label.

As shown in Figure 6.1, the majority of features are gradable into six different scalar degrees, meaning that there is a great deal of scope within the framework which the VQ analyst can utilise. Having such comprehensive scope with which to assess a speaker's voice may very well be beneficial within the clinical setting, however this version of the VPA scheme has been labelled as being "too complex" (McGlashan & Fourcin, 2008, p. 2175) with Webb at al. (2004, p.429) highlighting that "its greater scope is at the expense of reliability". Having a framework which is reliable (e.g., one in which a high degree of inter- and intra-rater agreement can be established) is of the utmost importance within the forensic sphere, and therefore the original version of the VPA scheme has been subjected to modification for its application within forensic phonetics. These modifications have come in the form of simplifications to the scheme, with there being two VPA variants specifically oriented towards forensic application.

The earliest of these variants is the Simplified Vocal Profile Analysis scheme (henceforth the SVPA) which was developed by San Segundo and Mompéan (2017) and implements a reduction in the number of settings and uses binary judgments rather than scalar degrees. The simplifications implemented are based on issues related to the perceptual assessment of VQ using the VPA and include the following:

(1) the highly multidimensional nature of VQ and the subsequent difficulties in isolating specific dimensions;

(2) raters having different definitions of a voice feature and a different understanding of the labels which should be assigned to a feature;

(3) although the analysis of pathological voice may require a complex framework, such a framework might be superfluous when examining non-pathological voice;

(4) the perceptual assessment of VQ is cognitively demanding and therefore a more refined framework may reduce this demand on raters.

The SVPA scheme was developed as a means of alleviating these issues, creating a framework in which intra- and inter-rater agreement was optimised and from which a distance measure of speaker similarity could be obtained. As such, San Segundo and Mompéan implemented the following modifications:

(1) reduction from 36 settings to 22;

(2) 10 major "setting groups" with 22 possible settings within those groups, that is, two articulatory strategies as possible deviations from neutrality;

(3) no scalar degrees; use of a binary (neutral/non-neutral) rating for each setting group;

(4) no marking of intermittent settings;

(5) possibility of including holistic descriptions regarding the settings being rated or any other VQ aspects.

(2017, pp. 14-15)

These modifications result in a markedly simplified framework which is shown in Figure 6.2 below.

A. Featural *(tick the appropriate box)*				
		Numerical Labels for One Neutral (N) and Two Non-Neutral Configurations		
Major Setting Groups	Settings	-1	0	+1
Vocal tract settings	Labial	Spreading	N	Rounding
	Mandibular	Close	N	Open
	Apical	Retracted	N	Advanced
	Dorsal	Backed and lowered	N	Fronted and raised
	Velopharyngeal	Denasal	N	Nasal
	Pharyngeal	Constricted	N	Expanded
	Laryngeal height	Lowered	N	Raised
Overall muscular tension	Vocal tract tension	Lax	N	Tense
	Laryngeal tension	Lax	N	Tense
Phonation	Voice type	Whisper/Breathy	N	Creaky/Harsh
B. Holistic				
(fill with qualitative input; comments, etc)				

Figure 6.2. The Simplified Vocal Profile Analysis framework (SVPA). Adapted from San Segundo and Mompéan (2017).

In their study, San Segundo and Mompéan explain the processes undertaken in the development and testing of the simplified framework which included two experienced phoneticians independently rating 24 speakers using the SVPA on two different occasions separated by a week (as a means of establishing intra-rater reliability). Prior to undertaking these assessments, the two phoneticians engaged in a calibration exercise in which they listened to a subset of voices together as a means of establishing agreeable definitions of the different settings as well as a mutual understanding of possible deviations from the neutral settings.

The results from their study showed that the SVPA scheme enabled high levels of intra-rater agreement and considerably good levels of inter-rater agreement to be obtained. Despite the positive results achieved, the authors highlight how improvements to the inter-rater agreements could be made by increasing the number

of training sessions between the analysts, including perceptual anchors within the study, as well as developing clearer definitions of the neutral baseline for the speaker population under evaluation. With regards to the second aim of the study, that being to establish whether a distance measure of speaker similarity could be derived through implementation of the SVPA, this is something which the framework permitted successfully. The experimental design of using twin pair speakers for the study subsequently allowed for the simple matching coefficients to be compared across twin pairs and non-twin pairs, with the expectation that twin pairs would have more similar VQ features. That is, despite the simplifications made, the SVPA seemingly preserved the most relevant settings from the original VPA, indicated by the higher matching coefficients attributed to the most similar speakers. San Segundo and Mompéan point towards the advantages of using an index of speaker similarity for VQ assessments within the forensic domain and further highlight that the SVPA could make using such an index more widespread within the field.

In discussing the limitations of the SVPA framework, the authors acknowledge that raters having to make a compulsory binary choice for each of the VQ features may not always be the most appropriate, in particular where a combination of settings is possible (e.g., a harsh-whispery voice type). In order to compensate for the strictness imposed by the forced binary choice for each of the VQ settings, San Segundo and Mompéan point out the addition within the SVPA framework for qualitative comments to be made pertaining to any holistic observations. With regards to further testing of the SVPA framework, the developers advocate for future studies to assess its potential for characterising VQ in additional languages, as well as checking the proposed settings against instrumental acoustic measures to assess the extent of any correlation between perceptual and acoustic assessments.

A later study, carried out by San Segundo et al. (2019), also sought to make modifications to the original VPA scheme in the form of simplifications. This modified version of the VPA (henceforth MVPA) was again based on Beck's (2007) version and was developed in part at JP French Associates, a forensic speech and acoustics laboratory in the UK, and further modified by San Segundo and colleagues for the purposes of their study. The modifications made to the MVPA were

implemented following an initial pilot study (a blind perceptual assessment) in which an initial ten speakers were independently rated by three trained phoneticians all of whom held experience of using the VPA scheme in forensically-orientated research. A calibration meeting was then held in which the raters' results were compared, problematic perceptual labels were discussed, and differences in analytic strategy were identified. The reader is directed to San Segundo et al. (2019, pp. 363-366) for detailed explanation and examples pertaining to the practical issues which emerged from the calibration session which factored into the final version of the MVPA. The main differences between the MVPA and the original VPA described by Beck (2007) is the reduction of settings and the reduction of the scalar degrees used. The modifications made are summarised below:

(1) Removal of *protruded jaw* and *audible nasal escape* settings.

(2) Merging of *fronted tongue body* and *raised tongue body*; *backed tongue body* and *lowered tongue body*; *creak* and *creaky*; *whisper* and *whispery*.

(3) Removal of the 'extra features' provided in a supplementary page of Beck (2007) which pertain to prosodic features and temporal organisation features.

(4) Reduction of the number of scalar degrees permitted from six down to three with these being defined as SLIGHT (1), MARKED (2) and EXTREME (3).

As well as the above simplifications to the original scheme, a section for 'notes' was also added to allow for initial holistic impressions of the voices to be made by raters. The authors comment on the usefulness of this addition during the calibration meetings, as there were occasions in which raters had marked the same perception impression, however had then conceptualised it differently according to the set of original VPA (Beck, 2007) pre-determined labels. The modified VPA framework is shown below in Figure 6.3.

	FIRST PASS		SECOND PASS				Notes
	Neutral	Non-Neutral	SETTING	Slight 1	Mrkd. 2	Extrm. 3	
A. VOCAL TRACT FEATURES							
Labial			Lip rounding/protrusion				
			Lip spreading				
			Labiodentalisation				
			Extensive labial range				
			Minimised labial range				
Mandibular			Close jaw				
			Open jaw				
			Extensive mandibular range				
			Minimised mandibular range				
Lingual tip/blade			Advanced tongue tip/blade				
			Retracted tongue tip/blade				
Lingual body			Fronted/raised tongue body				
			Backed/lowered tongue body				
			Extensive lingual range				
			Minimised lingual range				
Pharynx			Pharyngeal constriction				
			Pharyngeal expansion				
Velopharyngeal			Nasal				
			Denasal				
Larynx height			Raised larynx				
			Lowered larynx				
B. OVERALL MUSCULAR TENSION							
Vocal tract tension			Tense vocal tract				
			Lax vocal tract				
Laryngeal tension			Tense larynx				
			Lax larynx				
C. PHONATION FEATURES							
			Present		Scalar Degree		
	SETTING		Neutral	Non-neutral	Slight 1	Mrkd. 2	Extrm. 3
Voicing type	Falsetto						
	Creaky						
	Whispery						
	Breathy						
	Murmur						
	Harsh						
	Tremor						

Figure 6.3. The modified Vocal Profile Analysis framework (MVPA). Adapted from San Segundo et al. (2019).

As a means of testing the reliability of the MVPA, following the first calibration meeting, an additional 89 speakers were independently rated by the three phoneticians, resulting in a total of 99 speakers being assessed (including the 10 speakers from the initial pilot experiment who were subsequently reassessed). The main aims of the study were therefore to assess the reliability of the framework by examining the levels of inter-rater agreement obtained across the three analysts, and to evaluate the extent to which the settings of the MPVA are independent from one another (a factor which carries forensic implications as analyses carried out within FVC casework should not

rely upon correlated features as doing so could lead to the over-weighting of evidence).

In order to assess inter-rater agreement, San Segundo et al. conducted a number of different tests. In the first instance they assessed percentage agreement both with regards to absolute agreement and within one scalar degree. They found that results improved for most settings when measured in one scalar degree, particularly for seven out of the 32 settings. Overall, they found that agreement was 'very good' (> 70%) but that this was highly dependent upon the setting being assessed. Also worthy of note here is the finding that where a setting is more frequent (i.e., the more in which raters made observations on a given setting), this actually resulted in lower inter-rater agreement for that particular setting.

In terms of the intra-rater agreement results, San Segundo et al. report promising results although they acknowledge that these are only preliminary results which only considered a subgroup of ten speakers. They found that within-rater agreement ranged between 93% and 96% when all settings were considered. More specifically, results which considered only the settings used more frequently across the corpus as a whole (more than 60%) were found to be 'very good', with percentage agreement ranging between 73% and 87%.

When discussing their inter-rater agreement findings, San Segundo et al. highlight the finding that seven of the 32 settings showed marked improvement in percentage agreement if measured within one scalar degree (as opposed to absolute agreement). They point out that these seven settings in fact occupied much of the raters' discussion during the calibration meetings, indicating the apparent issues involved in their definition, labelling or perceptual salience, with this being offered as reasoning as to why the agreement reached for them was not as high as with other settings.

Acknowledging that it is not advisable to compare certain inter-rater values across different studies (see Uebersax, 1987), San Segundo and colleagues nevertheless compare their findings to Webb et al.'s (2004) results which also assessed inter-rater agreement in the context of the VPA protocol. San Segundo et al. found their results to be markedly better than those reported by Webb et al. In making comparisons to an additional study which reported inter-rater agreement for VPA settings, San Segundo

and colleagues turn to Beck's (2005) study which reported percentage agreement results across two raters. Beck et al. sound that for a number of settings, inter-rater agreement was no higher than 50%, meaning that, again, San Segundo et al.'s results were far superior. Finally, they also point to a study which assessed pathological voices using the VPA framework (Wirz & Mackenzie Beck, 1995), where the inter-rater agreement results obtained were only 'modest' at best. In drawing attention to this variability in inter-rater agreement results across different studies, San Segundo et al. highlight the need for a theoretical framework to explain such variation (Kreiman et al. 1993; Kreiman & Gerratt, 2011).

In offering explanations for the comparably high levels of inter-rater agreement in their study, the authors highlight that each of the raters possessed their own strengths and weaknesses in their assessment of different settings and that they were conscious of these. They stress the importance of adopting a team approach as a route to overcoming errors and alleviating individual biases, with the calibration sessions held throughout the study being a key factor in the promising level of inter-rater agreement achieved. They do concede, however, that, although in line with previous studies with a similar focus, the relatively small sample size of the study may have been influential in the inter-rater agreement results they obtained.

With regards to the study's second aim of evaluating the extent to which the MVPA settings were independent from one another, they found that, although some settings were more correlated than others (e.g., *raised larynx* and *tense larynx*), none of the correlations were strong enough to merit any settings being merged into a single setting – that is, each individual setting was shown to provide useful, specific information for speaker characterisation.

Given the promising results obtained for each of the aims of the study, it could be presumed that San Segundo and colleagues would go on to advocate for the use of the MVPA framework by forensic phoneticians. Conversely however, although the study was conducted as part of a larger, forensically-oriented research project, the authors refrain from making any such assertion due to the data on which the MVPA was developed lacking the diversity found in typical forensic recordings. Instead, the authors choose to draw attention to the methodological procedures which they

followed – the calibration meetings and the important discussions held regarding practical and methodological issues – as a means of providing 'an example for those wishing to make adaptations of the VPA for forensic applications' (San Segundo et al., 2019, p. 354). That is, they highlight how their study can serve as a methodological example of how a forensically-orientated framework can be adapted for different purposes. The reader is directed to San Segundo et al. (2019, pp. 370-372) for their summary of the main points discussed pertaining to the practical and methodological issues they encountered in developing their framework.

One point that shall be mentioned here given its direct relevance to the present work is that, when discussing certain aspects which the VPA failed to account for when assessing the speakers in their data (e.g., phenomena such as *audible oral escape* or, alternatively, *inadequate breath control*), the authors discuss how there is no label for any rhythmic aspects evidenced by speakers. Although usually considered separately from VQ, San Segundo and colleagues nevertheless comment on how some rhythmic features were saliant in some of the speakers, resulting in raters adding impressionistic, holistic comments descriptions such as 'lively', 'active' or 'monotonous' to their assessments.

In relation to the development of the PARFA framework (see Section 6.3), the review of the VPA scheme and its two modified variants presented above is useful in a number of ways. Firstly, the structure and layout of the PARFA framework are largely based upon the VPA framework (and modified variants). Specifically, the initial draft of the PARFA framework (see Figure 6.12) took inspiration from San Segundo et al.'s (2019) MVPA in that rhythmic features could be marked using a rating of (1) slight, (2) marked, or (3) extreme. Following consultation with a forensic practitioner, this initial design was modified, however, as having this scalar rating was deemed to be superfluous (see Section 6.3.3 for further discussion). The modification made with regards to removing the scalar aspect of the framework has resulted in the PARFA framework adapting a structure more closely resembling San Segundo and Mompéan's (2017) SVPA framework. That is, the PARFA framework's current structure provides raters with a binary option when assessing a specific rhythmic feature (e.g., whether a specific feature is 'absent' or 'present'). Further influence is

also derived from the structure of both the SVPA and MVPA in that the PARFA framework offers a dedicated section for raters to comment on their 'holistic assessment of speech rhythm', whilst also providing a 'notes' section for each of the rhythmic features (see Section 6.3.1 for discussion on the structure of the PARFA framework).

Secondly, the review of the VPA and its modified variants has highlighted the ways in which frameworks designed for use within the forensic domain can be tested in order to determine what degree of inter and intra-rater agreement can be established. The use of pilot studies, statistical testing, calibration meetings and follow up discussion groups are all highlighted by Segundo and colleagues as being the most important element of their work, suggesting that the procedures they followed can act as an exemplar to others who wish to undertake the development of a framework. Indeed, the present study follows a number of these procedures in the testing stages of the PARFA framework (see Section 6.3.5.).

Thirdly, the VPA and its variants refer to characteristics that have a mechanistic basis (e.g., modes of vocal fold vibration) that are known to vary between speakers. The PARFA framework also includes rhythmic features which have a mechanistic basis, albeit to a lesser degree than the VPA. One such feature for which observations can be made relates to *amplitude* patterns. As discussed in Chapter 2, Section 2.5, previous research has demonstrated a relationship between the size of the mouth aperture and the amplitude produced. Specifically, a larger mouth opening correlates with increased amplitude, whereas a smaller opening is related to reduced amplitude. Furthermore, research examining subglottic and pulmonic air pressure, both of which are fundamentally associated with speech amplitude, has revealed considerable variability amongst speakers. Aside from *amplitude*, there are other features for which observations can be marked within the PARFA framework which are also mechanistic in their nature such a speaker's *syllabic organisation*. Observations in relation to *syllabic organisation* include how speakers distribute syllables whilst speaking, whether syllables are subject to prolongations, and the prosodic patterning attributed to syllable delivery. Features such as these will be influenced by the distinct ways in which speakers engage their articulatory mechanisms, along with specific anatomical

traits that vary from one speaker to another. Indeed, previous speech rhythm research has indicated that individual articulatory patterns are the most credible explanation for explaining between-speaker differences. (e.g., Dellwo et al., 2015; Leemann et al., 2014). There are, however, other elements of the PARFA framework which assess speech rhythm from a more behavioural perspective as opposed to a mechanistic basis. Such features include assessing a speaker's *pausing behaviour* and *disfluency behaviour*, both of which also hold the potential to tap into speaker-specific patterns, with prior research demonstrating how speakers can use these behaviours in idiosyncratic ways (e.g., Kolly et al., 2015; McDougall & Duckworth 2017, 2018). Overall, having some features within the PARFA framework which have a mechanistic basis similar to that of the VPA, and some features which are more behavioural in nature, aims to provide a balanced format for which rhythm can be assessed in a structure and comprehensive manner.

6.2.2. TOFFA (Taxonomy of Fluency Features for Forensic Analysis)

The TOFFA framework was published by McDougall and Duckworth (2017) as a means of a providing a systematic way for quantifying variation in disfluency phenomena for forensic purposes. Where previously disfluency behaviour was only described at an impressionistic level in FVC cases, TOFFA seeks to offer a more objective approach to the analysis of disfluencies through a clearly defined methodology, enabling precise quantification and replication of findings, whilst allowing the analyst to capture features that are not necessarily perceptually salient. However, despite this being the aim of TOFFA, it should be acknowledged that actually achieving replicable disfluency analyses within forensic casework is by no means straightforward owing to the degree of subjective judgement involved in identifying and categorising different disfluency types, as well as the nature of the recordings which analysts will be dealing with (e.g., discrepancies between the speech samples being compared in terms of recording quality, speaking style, speaking topic, interlocuter, etc.).

Motivation for analysing disfluencies stems from the notion that speech disfluencies possess promising potential for consideration within FVC casework. Speech

disfluencies, that is, phenomena such as repetitions, prolongations, self-interruptions, filled pauses and silent pauses, are realised in the temporal domain. This separates speech disfluencies from other speech features routinely analysed within FVC cases (e.g., formant frequencies), which are analysed through spectral information. Therefore, whilst information carried by formant frequencies can be affected by degraded recordings (e.g., reduced bandwidth of telephone transmissions, background noise, etc.), the information transmitted through speakers' disfluency behaviour will be more robust to these challenging recording conditions (given that the speech remains intelligible).

Another reason why the analysis of speakers' disfluency patterns is appealing to forensic casework arises from the notion that disfluencies fall within unconscious phenomena which manifest spontaneously within everyday, unmonitored speech. Thus, it is supposed that such phenomena will be difficult to deliberately and consistently disguise as speakers (and listeners) are generally unaware that they are occurring (Finlayson & Corley, 2012).

The composition of the framework was informed by previous research on both normally-fluent speakers (e.g., Shriberg, 2001) and the speech of people who stutter (e.g., Wingate, 1964; Van Riper, 1973), as well as observations made in relation to their dataset (the DyViS database (Nolan et al., 2009)). The framework adopts a general definition of a 'fluency disruption' as: any phenomenon originated by the speaker which changes the flow of the speaker's utterance. The structure of TOFFA is outlined below in Table 6.1.

Table 6.1. The TOFFA framework showing the categories and subcategories of disfluency features. Adapted from McDougall et al. (2019).

Main category	Subcategories and examples
Silent Pauses	- 'grammatical' **[pg]** - 'other' **[po]**
Filled Pauses	- *er* **[er]** - *erm* **[erm]** - others, e.g., *ah* **[fpo]**
Repetitions	- part-word **[pwr]** *on the road I park my car th-there's* - whole word **[wrep]** *but she- she's also* - phrase **[prep]** *on your-on your left there's a reservoir* - multiple (i.e., more than 2 iterations) **[mrep]** *a hairdresser at the- at the- at the- at the*
Prolongations	(duration ≥ 200 msecs) - vocalic, e.g., vowel, nasal, lateral **[prov]** - fricative **[prof]** - plosive closure duration or affricate closure or release duration **[prop]**
Interruptions	(speaker interrupts self and discontinues the utterance, or continues with a modification) - phrase **[pint]** *pighty road which- and then then you …* - word **[wint]** *I th- I probably recognise like the bar lady*

In their publication of TOFFA, McDougall and Duckworth presented a study in which they tested the framework by assessing the individual variation in disfluencies for a group of 20 male speakers from the DyViS database with the aim of determining the range of usage of disfluency phenomena and the extent to which the disfluency profiles were speaker-specific. In order to create TOFFA profiles for the speakers, they first transcribed the speech data orthographically in a TextGrid (within Praat) and then annotated the disfluency features (using the square-bracketed codes). The transcriptions and annotations were then transferred to a spreadsheet along with a record of the number of phonetic syllables per utterance. From this they could then calculate the number of occurrences of each disfluency feature per 100 syllables for each speaker.

Given the degree of subjective judgement involved in identifying and categorising the disfluency types analysed, McDougall and Duckworth carried out an inter-analyst

consistency study. Both McDougall and Duckworth, along with a third analyst, undertook training together to become familiar with the criteria for identifying each disfluency, as well as using the coding system and the method for counting syllables. The three analysts reanalysed a subset of five speakers and the consistency of disfluency feature measurements across analysts was evaluated. Following this, the analysts held a subsequent meeting in which they discussed their experiences of using the categorisation system and jointly decided on revised criteria for the identification of features which had proved ambiguous or problematic. The results from the consistency study showed that for some disfluency types the correlations between the analysts were high (e.g., filled pauses had correlation rates ranging between $r = 0.84$ and 0.91), whereas for other disfluency types the levels of correlation were much lower (e.g., prolongations: **[prov]**: $r = 0.30–0.63$, **[prof]**: $r = 0.08–0.41$, **[prop]**: $r = 0.50–0.53$). When all disfluency types were considered together, this yielded high levels of correlation amongst all three pairs of analysts at $r = 0.88–0.93$, confirming that fluency feature analysis if all the features are considered exhibits a reasonable level of inter-analyst consistency.

In terms of the results from the main study, McDougall and Duckworth found that filled and unfilled pauses occur most frequently in the sample with repetitions, prolongations and interruptions occurring commonly, but less frequently. There was extensive between-speaker variation in disfluency behaviour in the disfluency profile used by each individual as well as the extent to which each feature was used. They employed discriminant analyses in order to assess the speaker-specificity of each of the disfluency types, with classification rates ranging from 5.7–11.3% (chance level = 5% as there were 20 speakers). The best performing individual features were the filled pauses **[er]** (10.3%) and **[erm]** (11.3%), with the full set of features for each speaker producing a classification rate of 14.4%. A combined analysis which accounted for the 7 best performing disfluency types produced a markedly higher classification rate of 29.4%, demonstrating the importance of considering speakers' fluency profiles for characterising differences among speakers. Given their overall finding that speakers demonstrated extensive speaker-specific differences in their fluency profiles both in terms of the types of disfluency features they employed and their rate of occurrence, they conclude by stating that, where relevant, disfluency

analysis using the TOFFA framework could be a useful tool for the forensic practitioner within FVC casework.

Before the TOFFA framework was published in the aforementioned study, McDougall, Duckworth and Hudson (2015) had in fact already used the then unnamed taxonomy to investigate whether patterns of disfluency differed across two different accents – Standard South British English and York English. Using the same methods as described above, they analysed the disfluency behaviour of 20 male speakers from each accent group and found that the overall frequency of disfluencies across accent groups was similar, but that there were some differences for certain disfluency subcategories that showed group-specificity.

With the above two studies illustrating the forensic potential of the TOFFA framework, McDougall and Duckworth (2018) sought to further test its efficacy in a follow up study which examined the extent to which individuals' disfluency behaviour was preserved over two different conversational styles. The examined the disfluency patterns of 20 male speakers of South Standard British English from the DyViS corpus in two different forensically relevant tasks – a mock police interview and a telephone conversation with an 'accomplice'.

Their results showed that disfluency features displayed speaker-specific variation in both interview and telephone speaking styles, and that this speaker-specific information has a degree of within-speaker consistency across the two styles. Correlations between the speakers' rates of disfluency in the two styles were found both for speakers' overall rates of disfluency **[all]** and for many of the separate categories of disfluency feature examined, in particular, filled pauses, silent pauses, and repetitions. There were certain features with a low overall occurrence for most speakers that tended to show fewer clear patterns of correlation, as well as some that were also infrequently occurring, but showed within-speaker consistency (due to some speakers not using the feature at all and others being relatively consistent in their small amounts of usage).

Similar to their earlier (2017) study, McDougall and Duckworth carried out discriminant analyses to assess the levels of speaker-specificity presented by a speakers' disfluency profiles as well as the speaker-specificity of individual

disfluency features. They found that, for both speaking styles, all of the disfluency features exhibited some degree of speaker-specificity, with certain features bearing larger amounts of speaker-discriminating information than others. The best-performing disfluency category in telephone style was the filled pauses **[erm]** (18%) which was also equal-best alongside **[prof]** and **[prop]** in interview style (11.3%). In addition, the overall disfluency metric **[all]** yielded encouraging levels of correct classification: 14.4% in interview style and 13.5% in telephone style. The combined analyses (of the seven best performing features) returned classification rates of classification rates 29.4% for interview style and 35.5% for telephone style, highlighting that disfluency features work in concert to convey individual differences between speakers.

McDougall and Duckworth therefore advocate that in pursuing sources of speaker-distinguishing information in disfluency, it is essential that speakers' *disfluency profiles* are examined. In concluding their findings, they suggest that the patterns of speaker-specificity and consistency across styles highlight the potential of the analysis of disfluency profiles to contribute in forensic voice comparison cases where recordings involve different speech styles. Disfluency features provide a useful source of information about a behavioural aspect of a speaker's performance to complement other phonetic features typically analysed in forensic cases.

It has been shown above how a framework with forensic implications can be tested through academic, laboratory-based research – but how does such a framework get applied within real FVC casework? In the case of TOFFA, it has been employed by analysts in FVC cases since 2015 (McDougall et al., 2018b).

McDougall et al. (2019) provide examples of three FVC cases in which the TOFFA framework has been applied, and where the findings contributed to the conclusions of the forensic reports. In this paper, the authors also highlight that the development and subsequent application of the framework was only made possible through laboratory-based research, their own casework investigations and regular discussion with other forensic phoneticians. These comments would appear to suggest that, although experimental research is a useful starting point in the testing of a forensic methodology, receiving input from forensic professionals who carry out casework is

of equal, if not greater, importance. This practitioner-based feedback is something which the PARFA framework has been fortunate enough to receive throughout its development (see Section 6.3).

Indeed, the paper also details slight modifications that were made to the original framework in order to make its application to casework more efficient such as collapsing the two silent pause variants (**[pg]** and **[po]**) into one unfilled pause label (**[ufp]**), collapsing the two variants for consonant prolongations (**[prof]** and **[prop]**) into one label (**[proc]**), and using a time-based approach rather than a syllable-based approach when quantifying the disfluency phenomena. Such modifications were evidently the result of discussions being held amongst those involved in utilising the framework.

This notion of engaging forensic phoneticians in the development of a framework which could have potential application within casework is something which the present author sought to achieve through the experimental design of the perception experiments in the previous chapter. This is something which other 'framework developers' also evidently acknowledge (e.g., the development of the VPA framework (see Section 6.2.1) and the development of the PASS framework (see Section 6.2.3)).

Although TOFFA has shown its merit with regards to its application within real FVC cases, this framework, as is often the case with frameworks or methodologies of this nature, has been subjected to modification. Modifications were made to the original TOFFA framework in Carroll (2019c) as a means of attempting to increase the speaker-discriminatory capacity of the original by introducing a more fine-grained classification to some disfluency phenomena. The TOFFAMo framework is summarised below in order to highlight the procedures involved in the development of a new framework, or rather in the *further* development of an existing framework.

The TOFFAMo framework (Taxonomy of Fluency Features for Forensic Analysis Modified) maintained the original five disfluency categories from the original, however implemented additional durational detail to some of the disfluency subcategories. Table 6.2 provides a summary of the TOFFAMo framework.

Table 6.2. The TOFFAMo framework showing the categories and subcategories of disfluency features.

Main category	Subcategories and examples
Silent Pauses	A silence ≥200 msecs within a single speaker's turn, including instances of 'breath pauses' (audible inhalation or exhalation) and 'clicks'. **[sp1]** a silent pause of ≥200 msecs **[sp2]** a silent pause of ≥500 msecs **[sp3]** a silent pause of ≥800 msecs
Filled Pauses	**[er1]** vowel alone e.g. *er* <300 msecs **[er2]** vowel alone e.g. *er* ≥300 msecs **[erm1]** vowel plus nasal e.g., *erm* <300 msecs **[erm2]** vowel plus nasal e.g., *erm* ≥300 msecs **[fpo]** any other sound which is not a central vowel e.g., */m:/, /a:/* ≥300 msecs
Repetitions	- part-word **[pwr]** *on the road I park my car th-there's* - whole word **[wrep]** *but she- she's also* - phrase **[prep]** *on your-on your left there's a reservoir* - multiple (i.e., more than 2 iterations) **[mrep]** *a hairdresser at the- at the- at the- at the*
Prolongations	(duration ≥ 300 msecs) - vocalic, e.g., vowel, nasal, lateral **[prov]** - fricative **[prof]** - plosive closure duration or affricate closure or release duration **[prop]**
Interruptions	(speaker interrupts self and discontinues the utterance, or continues with a modification) - phrase **[pint]** *pighty road which- and then then you …* - word **[wint]** *I th- I probably recognise like the bar lady*

The durational modifications implemented correspond to the silent pause category, filled pause category, and the prolongations category. The modifications made were motivated in the first instance by the nature of the speech data analysed in Carroll (2019a). The spontaneous conversational data in this earlier disfluency-focussed study meant that there was frequently noticeable variation in the duration of silent pauses, filled pauses and prolongations exhibited by individuals, leading Carroll to speculate that these durational patterns may be idiosyncratic (see Carroll (2019a, 2019b, 2019c) for further explanation as to the modifications implemented).

Having implemented these changes, Carroll (2019c) sought to test the efficacy of the TOFFAMo framework in a study which looked to build on the work of McDougall and Duckworth (2018) in assessing the consistency of speakers' disfluency behaviour across two different speaking styles. Using the TOFFAMo framework, Carroll analysed the disfluency patterns in the speech of 20 adult speakers engaged within an ethnographic interview are compared these with the disfluency patterns of the same 20 speakers involved in spontaneous conversation. Results showed considerable individual variation with regards to both overall rates of disfluency as well as the types of disfluency features used by speakers across both interview and conversational styles. Furthermore, speakers exhibited relatively consistent within-speaker patterns in disfluency behaviour across the two styles both with regards to overall rates of disfluency and the rates of occurrence of the individual features used. Similar to the results obtained by McDougall and Duckworth (2017, 2018), discriminant analyses of speakers' disfluency profiles revealed encouraging levels of speaker-distinguishing information for both interactional styles, with disfluency phenomena being shown to bear speaker-specific information.

In relation to the modifications introduced by the TOFFAMo framework, the additional durational detail afforded presented as being useful in helping to distinguish between speakers, with three of TOFFAMo duration-based pauses being amongst the best performing when subjected to the discriminant analyses. It is therefore suggested that these duration-based subcategories could potentially add probative value within certain real FVC cases. Indeed, having shared the TOFFAMo framework and the findings of the Carroll (2019c) study with the developers of the original TOFFA framework, it is understood that future research will implement durational subcategories as a means of further testing their probative potential (K. McDougall, personal communication, 10 July, 2023). The potential importance of accounting for durational information is echoed in the PARFA framework as it provides raters with the option to mark observations in relation to duration for a number of different features. Given that the design of PARFA is in part influenced by the qualitative feedback obtained from expert listeners (phoneticians/forensic phoneticians), this reinforces the idea that durational detail is a highly perceptible characteristic which listeners make use of when assessing speaker's speech rhythm patterns.

In relation to the development of the PARFA framework (see Section 6.3), both the TOFFA and TOFFAMo frameworks provide useful models from which influence can be drawn. Firstly, the development of the original TOFFA framework was motivated by there being no structured way in which disfluency phenomena could be measured for forensic purposes. That is, prior to TOFFA, disfluency behaviour would have only been described at an impressionistic level. At present, this too is the current state of affairs with regards to accounting for speech rhythm behaviour within forensic casework. Forensic research which has served to test the potential of the TOFFA framework has ultimately led to modifications being made to The TOFFA framework (e.g., TOFFAMo by the present author; but also by forensic practitioners (see McDougall et al. (2019)) in order to optimise its applicability within the forensic domain. The combination of thorough testing through research and modification to enhance applicability have ultimately led to the TOFFA framework being applied in real-life forensic casework (see McDougall et al. 2019). The review of the TOFFA framework, along with the modified variant TOFFAMo, has been provided here to demonstrate how an issue within the area of forensic casework can be remedied by the introduction of a structured framework – such is the purpose of the PARFA framework introduced in the present chapter.

6.2.3. PASS (Phonetic Assessment of Spoofed Speech)

The PASS framework is a framework developed by Lee et al. (2023) for detecting the presence of voice spoofing (fake speech) artefacts in speech recordings. Voice spoofing is the reconstruction of a target individual's speech which could be achieved by disguise, replay (i.e., replaying a previous recording of the target speaker), voice conversion or synthesis. In relation to the latter two methods, voice spoofing is an ever-increasing threat as criminals exploit rapid research advancements in machine learning models that can be used to create increasingly realistic sounding voice clones. Indeed, spoofed speech is being used as a tool to carry out various criminal activities such as impersonation (e.g., Brewster, 2021; Stupp, 2019), propagating fake news (e.g., Caldwell et al., 2020), and bypassing biometric authentication systems (e.g., Mirsky et al., 2022). In order to combat the criminal use of spoofed speech, there has

been some research by the anti-spoofing and countermeasures research community focussed on how speaker recognition technologies withstand spoofing attacks (e.g., Delgado et al., 2021; Todisco et al., 2019; Yamagishi et al., 2021).

In contrast to the upturn in research focussed on how automatic methods combat spoofed speech, there has been little research conducted on the ability of human listeners to detect spoofed speech. Initiating the quest to beset this lacuna in the research, Terblanche et al. (2021) conducted a study which assessed listeners' abilities to detect spoofed and genuine speech in samples of different quality (clean audio, background noise, mobile telephone transmission and internet video call transmission). Their results showed that of the 165 human listeners who participated in the study, spoofed speech samples were correctly detected 56% of the time, with one spoofing method being particularly successful in creating samples that were wrongly attested as being genuine human speech. The level of expertise of the listeners in this study is not reported, however, given that the experiment was conducted online and there was no targeted approach for specific listeners, it can be assumed that the majority were non-expert listeners.

A more recent study carried out by Mai et al. (2023) assessed the ability of 529 listeners to detect genuine and spoofed speech and reported that listeners' detection capabilities were unreliable as they only correctly identified spoofed speech 73% of the time. Again, it is not overtly apparent the level of expertise these listeners possessed, however, given that they were recruited online and selected on the basis that they were either fluent in English or Mandarin, it can be assumed that they were non-expert listeners.

In order to determine whether an expert listener possessed greater capacity to identify spoofed speech, Kirchhübel and Brown (2022) assessed the performance of a highly experienced full-time forensic speech practitioner in evaluating 300 samples which contained either spoofed speech or genuine human speech. The spoofed samples were derived from four different spoofing methods, allowing for the expert's performance to be compared to the results of previous research. Results showed that three of the spoofing methods proved no problem for the expert listener, whose performance

greatly exceeded the performance of the automatic systems reported in previous research (e.g., Hsu et al., 2017; Schroeder et al., 2011; Toda et al., 2005).

One of the spoofing methods, however, was problematic for the expert listener, with 26% of the samples being evaluated as genuine human samples, and the remaining 74% also presenting as being ambiguous as to whether they were spoofed samples or not. The spoofing method in question here was a type of *Text-to-Speech synthesis* system and is the same method identified in Terblanche et al.'s (2021) study as being the most successful in deceiving the non-expert listeners, a factor which prompted its inclusion in Kirchhübel and Brown's study. Kirchhübel and Brown present observations based on the qualitative notes of the expert listener in their study focussed on the samples derived from this spoofing method. Auditory observations showed the spoofed samples to resemble genuine human speech in many ways, including the following:

- lip smacks
- clicks
- audible breathing
- connected speech processes
- ejective release of plosive sounds
- sibilant sounds with a whistled quality
- speaker-idiosyncratic voice quality phonation types.

In addition to the spoofed samples bearing many of the characteristics of natural speech when assessed auditorily, acoustic observations of the spectrograms supported these auditory observations with features such as the ejective-release of velar plosives, breathiness, creaky voice and denasality all being evidenced. Given the apparent success of this one particular synthesis method in its ability to create spoofed samples which are able to deceive both human and machine, and the increasing possibility that

spoofed speech samples could feature within forensic casework, the question arises as to what measures can be put in place to ensure that the forensic speech analyst is best prepared for such an eventuality.

This is the question which Lee and colleagues have sought to answer with the development of the PASS framework. Making use of existing automatic methods, they have sought to complementarily provide phonetically-oriented insights on spoofing detection, combining expert and automatic tools for holistic countermeasures. The PASS framework has been developed to formalise the systematic and language-independent artefacts of voice spoofing, whilst also striving to present a unified vocabulary for describing spoofing artefacts in speech data. Development of the framework was in part inspired by the Vocal Profile Analysis scheme (see Section 6.2.2) and can be seen as being analogous to the VPA scheme in that PASS examines the naturalness ('human-likeness' or lack thereof) and artefacts of potential spoofs. Lee and colleagues also highlight that PASS is intended to complement the current automatic methods of spoof detection given that certain linguistic aspects of the signal can escape automatic methods. Therefore, PASS is intended to serve as a supplementary countermeasure tool in the fight against the threat of spoofed speech.

The PASS framework and its categories of spoofing artefacts were initially based on comparisons made between genuine and synthetic speech samples. Lee and colleagues closely analysed 36 samples of synthetic speech alongside their genuine counterparts using an auditory-phonetic and acoustic method to investigate patterns of spoofing artefacts. The initial impressionistic assessments of these samples corroborated the results obtained by Kirchhübel and Brown (2022) with regards to their evaluations of spoofed speech features evidenced, whilst also unearthing further insights. All auditory-phonetic and acoustic observations made on the comparisons between the spoofed samples and their genuine counterparts were then compiled, resulting in a set of candidate features. These features were then divided into 'auditory', 'visual', and 'acoustic-phonetic' categories in a triadic format, illustrated in Figure 6.4 below.

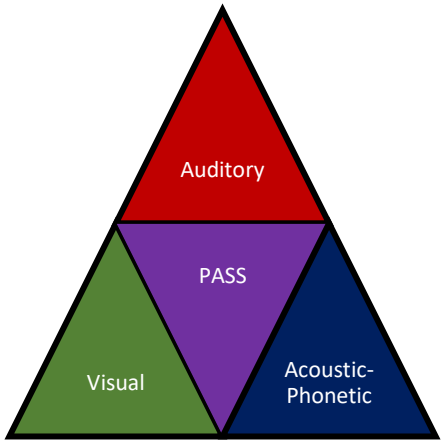


Figure 6.4. The PASS triadic method of validation (taken from Lee et al. (2023)).

Table 6.3 below shows the PASS framework in full.

Table 6.3. The Phonetic Assessment of Spoofed Speech framework (taken from Lee et al. (2023)).

	Description (Auditory)
TINNY QUALITY	An auditory label for 'hollow' or 'thin'-sounding audio.
CRACKLY QUALITY	An auditory label for 'bubbling' or 'crackling' sounds that occur constantly or frequently in the audio background.
MUFFLED QUALITY	An auditory label for the overall attenuation of segmental sounds, with dampening effects particularly pronounced for obstruent consonants.
RHYTHMIC QUALITY	An auditory label for the impression of an artificial rhythm, tempo, and metrical feet.
	Description (Visual)
FOGGING	A visual label for the 'smearing' or 'blurring' of otherwise distinctive structural features in the spectrogram for vowels and consonants.
FORMANT ATTENUATION	Refers to the loss of formant structure definition, particularly for vowels in the higher frequency regions.
PSEUDO-FORMANTS	Formant-like structures in the spectrogram that occur during the articulation of approximant consonants, which behave differently in spoofed audio depending on the specific segment.
CONCATENATEDNESS	Visible overly 'neat' segmental chunking and relative lack of dynamic between-segment features in the acoustic signal.
HYPERNEATNESS	Overly 'neat' linear predictive coding (LPC) points and tracks for formants, with unusually minimal errors in the spectrogram.
	Description (Acoustic-Phonetic)
HYPERFLAT PROSODY	An auditorily perceptible and acoustically analysable property that may be described as an overly level or flat prosodic pattern that is characteristic of 'robotic' speech.
COARTICULATORY DEFICIT	The deficit of between-segment coarticulatory features, which can result in the speech sounding overly 'neat' due to the concatenation of cleanly spliced segment content.

Observing the three categories outlined in the PASS framework, the *auditory* labels can be seen to described perceptual 'qualities' (taking inspiration from the VPA framework), whilst *visual* labels refer to visibly atypical features evidenced in spectrograms and waveforms. Of particular relevance to the present thesis is the

perceptual label 'RHYTHMIC QUALITY' which is the label for 'the impression of an artificial rhythm, tempo, and metrical feet'. At present, no further information is provided by Lee and colleagues as to what determines a specific 'rhythmic quality' as sounding 'artificial', however what is pertinent is that this 'rhythmic quality' label is something which is assessed perceptually by the listener (i.e., a perceptual label). This would seemingly indicate acceptance and agreement with regards to the need for the analysis of speech rhythm within forensic casework and that this analysis is one which should be perceptual in its nature. The *Acoustic-phonetic* labels make reference to features that can be detected auditorily and acoustically, invoking linguistic-theoretic knowledge. Lee and colleagues also propose that the framework should be implemented in a system of phases as illustrated by Figure 6.5 below.

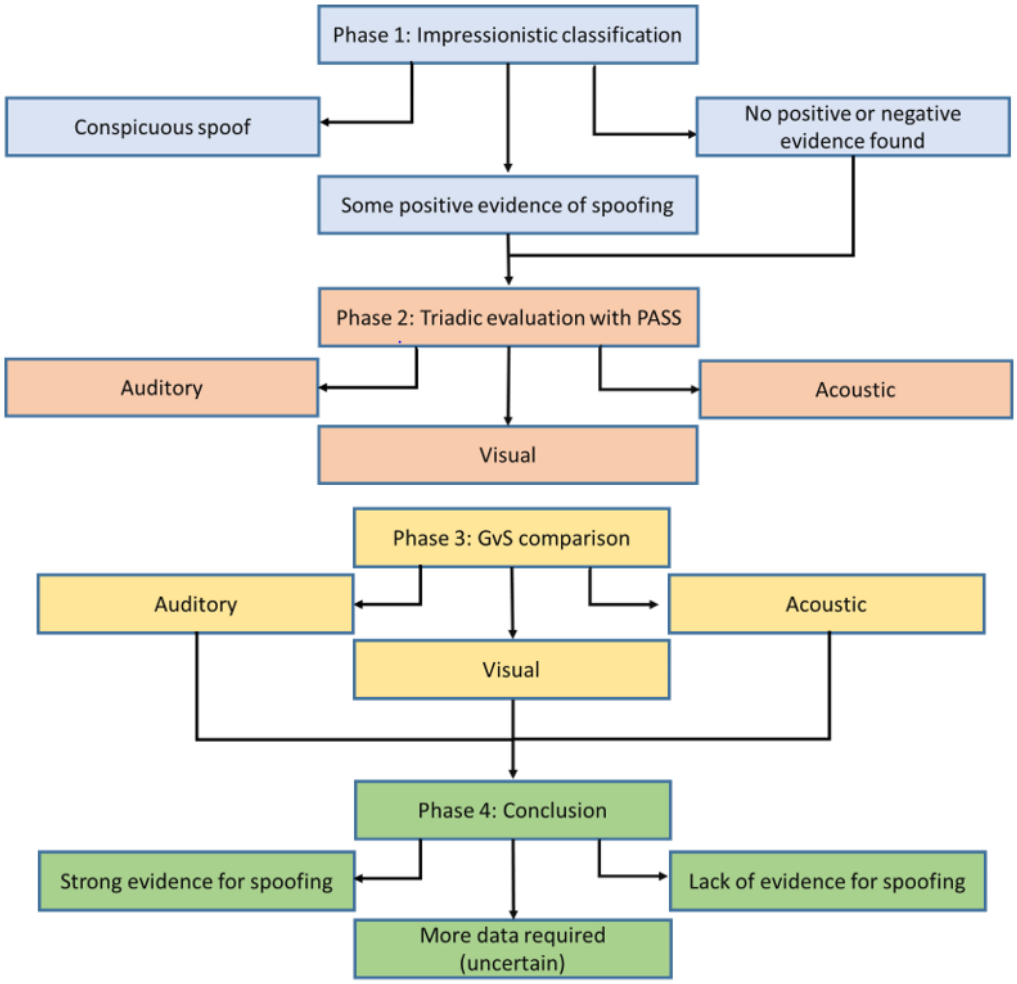


Figure 6.5. The PASS framework in phases (taken from Lee et al. (2023)).

Having the PASS framework in place and a series of phases through which suspected spoofed material should be filtered through, Lee and colleagues next sought to test the application of the PASS categories. They conducted a pilot study in which a phonetically trained listener applied the PASS framework to a blind test involving 10 samples of genuine and spoofed speech in a binary choice task in which the listener had to decide if each sample was genuine or spoofed. Results showed that the majority of judgments were correct (9 out of 10), and that FORMANT ATTENUATION and FOGGING were particularly effective in determining whether a sample was in fact a spoof. In accounting for the one incorrect judgement, in which the listener judged a spoofed sample to be genuine human speech, they explain that within Phase 2 of the phasing system (i.e., the triadic evaluation with PASS) the listener conducted auditory and visual analysis of the sample, however omitted the acoustic-phonetic step in want of assessing whether an expert listener could evade the time-consuming process of manual acoustic analyses. In addition, they also highlight that the expert listener completely omitted Phase 3, meaning that no comparison was made between the spoofed sample and a sample which was known to contain genuine human speech. Indeed, in a post-hoc analysis of the incorrectly evaluated speech sample, Lee and colleagues demonstrate how the implementation of Phase 3 (that is, comparing the sample to a genuine speech sample) could have been useful as visually comparing the spectrogram of the spoofed sample with the spectrogram of a genuine sample highlighted evidence of FORMANT ATTENUATION and PSEUDO-FORMANTS. In their discussion of the implementation of Phase 3, they also highlight how conducting an auditory comparison between the spoofed sample and a genuine sample could have been bolstered through the use of high-fidelity studio headphones, which when used in their post-hoc analysis revealed a slight degree of MUFFLED QUALITY.

Although the implementation of the PASS framework is in its very early stages, the development process and subsequent testing of the framework has demonstrated its potential as a practical aid for human experts to discriminate between spoofed speech and genuine human speech. Lee and colleagues advocate that the framework should now be subjected to further testing under other utterance and language contexts and also highlight the need to improve the inter-rater reliability and robustness of the

framework by having other trained (expert) listeners attempt to discriminate between spoofed and genuine speech samples using PASS. Given the ever-increasing threat which spoofed speech poses, not just to the forensic analyst but also the wider public (e.g., impersonation, propagating fake news, bypassing biometric authentication systems, etc.), the PASS framework (or more likely a condensed lay version of the framework) also carries the potential to benefit the wider public in providing an education on typical features of spoofing. This is one avenue which further testing of the framework might be geared towards (D. Lee, personal communication, 16 August, 2023).

In relation to the development of the PARFA framework (see Section 6.3), the PASS framework provides a useful example as to how and why a perceptual rhythm framework could be useful. Lee and colleagues identify that spoofed speech has the potential to be identified through an 'artificial' *rhythmic quality*, and that this *rhythmic quality* is a perceptual feature – that is, one which is detected by the human listener by auditory means. At present, the PASS framework does not provide any further elaboration as to what contributes towards the 'artificialness' of this *rhythmic quality* (the PASS framework is still currently under development). It stands to reason, however, that the detail offered in the PARFA framework (see Section 6.3) could be useful in helping to identify artificial speech rhythm properties. The PARFA framework could therefore be used alongside the PASS framework (where relevant), acting as a further (more detailed) means of spoofed speech detection following 'artificial' rhythmic characteristics being initial identified. Furthermore, it is likely that accounting for speech rhythm by perceptual means (whether for spoofed speech detection or for other reasons) is more favourable than depending on acoustic analyses (or automatic methods) given the complexity of the multiple features (and their interrelations) which make up speech rhythm. Finally, the review of the PASS framework provided above also helps to exemplify how a new methodological framework which stands to have implications and applications within the forensic domain is currently being tested and also how its development is influenced by previous forensically-motivated research.

6.3. Introducing the Perceptual Assessment of Rhythm for Forensic Analysis framework (PARFA)

The focus of the remainder of this chapter will be the proposition of a perceptual rhythm framework for forensic speech analysis. This framework has been titled the Perceptual Assessment of Rhythm for Forensic Analysis framework (henceforth PARFA). The framework draws heavily on the qualitative feedback obtained from the perception experiments presented in the previous chapter of this thesis, especially the qualitive feedback obtained from phoneticians, forensic phoneticians and forensic practitioners. The layout and structuring of the PARFA framework takes inspiration from the VPA scheme (Laver, 1980) and its subsequent forensically-orientated modifications (San Segundo and Mompéan, 2017; San Segundo et al., 2019). It is suspected that the current PARFA framework may undergo modifications/simplifications following initial testing/calibration sessions, nevertheless, to date, Figure 6.6 below shows the current version of the PARFA framework.

A. HOLISTIC ASSESSMENT OF SPEECH RHYTHM					
Conceptualisation	Notes			Suggested terminology	
Overall Rhythmic Feel				Active	Monotonous
Rhythmic Patterning				Balanced	Sporadic
Rhythmic Flow				Disjunct	Pulsing
Rhythmic Beat				Regular	Bouncy
				Disfluent	Fluent
				Lively	Unpredictable
	ATTRIBUTES	TICK TO INDICATE OBSERVATIONS			Notes
B. UTTERANCE-LEVEL FEATURES					
Pausing behaviour	Duration	Regularity		Variability	
	Distribution	Regularity		Variability	
	Frequency	Few		Many	
	Interactions	Absent		Present	
Intonation phrases	Duration	Regularity		Variability	
	Intonation	Regularity		Variability	
	Amplitude	Regularity		Variability	
	Opening cues	Absent		Present	
	Closing cues	Absent		Present	
C. WITHIN-PHRASE LEVEL FEATURES					
Filled pauses	Duration	Regularity		Variability	
	Distribution	Regularity		Variability	
	Frequency	Few		Many	
	Prosody	Neutral		Non-Neutral	
	Type	Regularity		Variability	
	Interactions	Absent		Present	
Disfluency behaviour	Frequency	Few		Many	
	Distribution	Regularity		Variability	
	Complexity	Absent		Present	
	Type	Regularity		Variability	
	Interactions	Absent		Present	
Syllabic organisation	Distribution	Regularity		Variability	
	Prolongations	Absent		Present	
	Prosody	Neutral		Non-Neutral	
	Rhythmic feel	Staccato		Legato	
D. OPENINGS AND CLOSINGS					
Prosody	Pitch	Neutral		Non-Neutral	
	Amplitude	Neutral		Non-Neutral	
	Prolongations	Absent		Present	
Common occurrences	Disfluencies	Absent		Present	
	Discourse markers	Absent		Present	

Figure 6.6. The PARFA framework.

The above PARFA framework has evolved from an earlier version, with modifications being implemented following an initial consultation session with a forensic practitioner. This earlier version of the framework is presented in Section 6.3.3, in which the reasons for the modifications are explained. The following subsections provide elaboration on the structure and layout of the PARFA framework along with detailed examples and advice for the analyst to follow when undertaking a PARFA analysis. It is also worth highlighting at this juncture that the PARFA framework is not intended to be a mere 'tick-box exercise'. That is, a meaningful PARFA analysis does not require the analyst to mark observations for each and every feature and their associated attributes. Rather, observations should be made as and when they are relevant to a particular analysis (see Section 6.3.4.3 for further discussion on marking observations).

The remainder of this section is structured as follows. First, in Section 6.3.1, the structure of the framework is described and explained. Following this, Section 6.3.2 provides detailed explanation and guidance with regards to how each section of the PARFA framework should be completed. In Section 6.3.3, the initial draft version of the framework is presented in order to provide elaboration regarding the modifications that were made, and which feature in the final version presented above (Figure 6.6). Lastly, Section 6.3.4 provides additional notes on the PARFA framework in relation to when the framework should be used, advice on the listening challenges posed by the analysis, guidance on marking observations when using the framework, as well as information regarding the proposed initial testing of the framework.

6.3.1. Framework structure

The PARFA framework is arranged into four sections (A, B, C and D), each concerned with rhythmic attributes at different levels of speech organisation. The framework begins with a wide scope which becomes narrower as the analyst progresses through the framework. Adopting a wide-to-narrow approach was taken on the basis that the perception of an individual's speech rhythm will generally be conceptualised as numerous interrelated rhythmic attributes combining over a stretch of speech. Therefore, prompting the analyst to initially conduct a holistic assessment in which

they consider the entirety of a given speech sample seemed the most appropriate modus operandi. In this initial section of the framework (Section A), on the left-hand side, the analyst is provided with terms to focus their conceptualisation of the speaker's speech rhythm, such that they ask themselves questions such as – what is this speaker's *overall rhythmic flow* like? What is the *rhythmic feel* of this speaker's speech? On the right-hand side of this section of the framework, the analyst is provided with some suggested terminology to use to describe the speaker's rhythmic patterning. In the centre of this section is the space provided for the analyst's descriptive notes pertaining to their holistic assessment of the speaker's speech rhythm. Figure 6.7 provides an example of this section of the framework.

A. HOLISTIC ASSESSMENT OF SPEECH RHYTHM			
Conceptualisation	Notes	Suggested terminology	
Overall Rhythmic Feel			
		Active	Monotonous
Rhythmic Patterning		Balanced	Sporadic
		Disjunct	Pulsing
Rhythmic Flow		Regular	Bouncy
		Disfluent	Fluent
Rhythmic Beat		Lively	Unpredictable

Figure 6.7. Exemplar of Section A of the PARFA framework form.

Following on from the holistic assessment, the analyst then progresses to make observations relating to more specific *utterance-level features* and their associated attributes (Section B). The *utterance-level features* in the PARFA framework are divided into two main categories (or features) – attributes associated with the speaker's PAUSING BEHAVIOUR or attributes associated with the speaker's INTONATION PHRASES. These two *utterance-level features* therefore can be seen as two contributing levels which make up a given speaker's speech turn (of which there may be many within a given speech sample). These two main categories in this section are located on the right-hand side of the framework. Figure 6.8 provides an example of this section of the framework. See Section 6.3.2 for detailed explanations pertaining to the labels/terminology used in the columns of the framework.

	ATTRIBUTES	TICK TO INDICATE OBSERVATIONS			Notes
B. UTTERANCE-LEVEL FEATURES					
Pausing behaviour	Duration	Regularity ✓		Variability	
	Distribution	Regularity		Variability ✓	
	Frequency	Few ✓		Many	
	Interactions	Absent		Present ✓	Silent pauses often followed by filled pause
Intonation phrases	Duration	Regularity		Variability ✓	
	Intonation	Regularity		Variability	
	Amplitude	Regularity		Variability	
	Opening cues	Absent		Present ✓	Often begin with filled pause – generally of raised pitch
	Closing cues	Absent		Present	

Figure 6.8. Exemplar of Section B of the PARFA framework form.

The next stage of the PARFA framework further narrows the analytical scope down to the level *within-phrase-level features* (Section C). The *within-phrase-level features* are divided into three main categories (or features) with these being attributes pertaining to the speaker's use of FILLED PAUSES, their DISFLUENCY BEHAVIOUR, and the SYLLABIC ORGANISATION of their speech. These three *within-phrase-level features* therefore can be seen as being the constitutes of a specific phrase uttered by the speaker (of which there may be many within a given speech turn). Figure 6.9 provides an example of this section of the framework form.

C. WITHIN-PHRASE LEVEL FEATURES					Notes
Filled pauses	Duration	Regularity		Variability	
	Distribution	Regularity		Variability	
	Frequency	Few		Many	
				✓	
	Prosody	Neutral		Non-Neutral	Often raised pitch
				✓	
	Type	Regularity		Variability	Predominantly 'er'
		✓			
	Interactions	Absent		Present	w/ silent pauses
				✓	
Disfluency behaviour	Frequency	Few		Many	
				✓	
	Distribution	Regularity		Variability	
				✓	
	Complexity	Absent		Present	
				✓	
	Type	Regularity		Variability	Predominantly part-word reps
		✓			
	Interactions	Absent		Present	
Syllabic organisation	Distribution	Regularity		Variability	Disjunct feel
				✓	
	Prolongations	Absent		Present	
	Prosody	Neutral		Non-Neutral	Relatively monotone
		✓			
	Rhythmic feel	Staccato		Legato	

Figure 6.9. Exemplar of Section C of the PARFA framework form.

The final stage of the PARFA framework allows the analyst to mark observations concerned with *openings and closings* (Section D). These *openings and closing* relate to the openings and closings of specific phrases uttered by the speaker (i.e., at the *within-phrase-level*). The features of the *openings and closings* are divided into two main categories with these attributes relating to the speaker's use of PROSODY and any COMMON OCCURRENCES associated with the speaker's openings and closings of phrases. (At *utterance-level*, openings and closings have their own sub-category in which to mark their presence or absence (*opening cues* and *closing cues)* – these openings and closing therefore correspond to the openings and closings at the

beginning and end of a speaker's speech turn.) Figure 6.10 provides an example of this section of the framework.

D. OPENINGS AND CLOSINGS					Notes
Prosody	Pitch	Neutral		Non-Neutral	Often raised pitch
				✓	
	Amplitude	Neutral		Non-Neutral	
	Prolongations	Absent		Present	Opening 'er'
				✓	
Common occurrences	Disfluencies	Absent		Present	
	Discourse markers	Absent		Present	'you know' / 'yeah' frequently used to close
				✓	

Figure 6.10. Exemplar of Section D of the PARFA framework form.

For sections B, C and D of the PARFA framework, the main category labels for each section are located on the left-hand side of the form (e.g., PAUSING BEHAVIOUR). To the right of the category labels, in a separate column, the ATTRIBUTES associated with each main category are provided (e.g., DISTRIBUTION). These ATTRIBUTES vary for each main category label (although there are some ATTRIBUTES which relate to numerous main category labels (e.g., DURATION)), with some main categories having more associated ATTRIBUTES than others. To the right of the ATTRIBUTES column is where the analyst marks their observation with regards to the specific ATTRIBUTE. Here the analyst. For every ATTRIBUTE, the analyst is able to, where relevant (see Section 6.3.4.3 for discussion relating to marking observations), make a binary decision between two opposing observations. The analyst is instructed to 'TICK TO INDICATE OBSERVATIONS' in the relevant box underneath one of the two options.

The two opposing options vary depending on the specific ATTRIBUTE being observed, however these two opposing options can be summarised as follows:

- REGULARITY vs. VARIABILITY
- FEW vs. MANY
- ABSENT vs. PRESENT
- NEUTRAL vs. NON-NEUTRAL
- STACCATO vs. LEGATO

To the far right-hand side of the PARFA form, next to where the analyst has indicated (ticked) the choice they have made for a specific attribute, there is a final column in which the analyst can make notes (where relevant) pertaining to that specific observation.

6.3.2. Completing the PARFA framework

The following subsections provide detailed guidance for the analyst with regards to making observations for each section of the PARFA form. Figure 6.11 below provides an illustration as to how the analyst should navigate through the different levels of the PARFA framework.

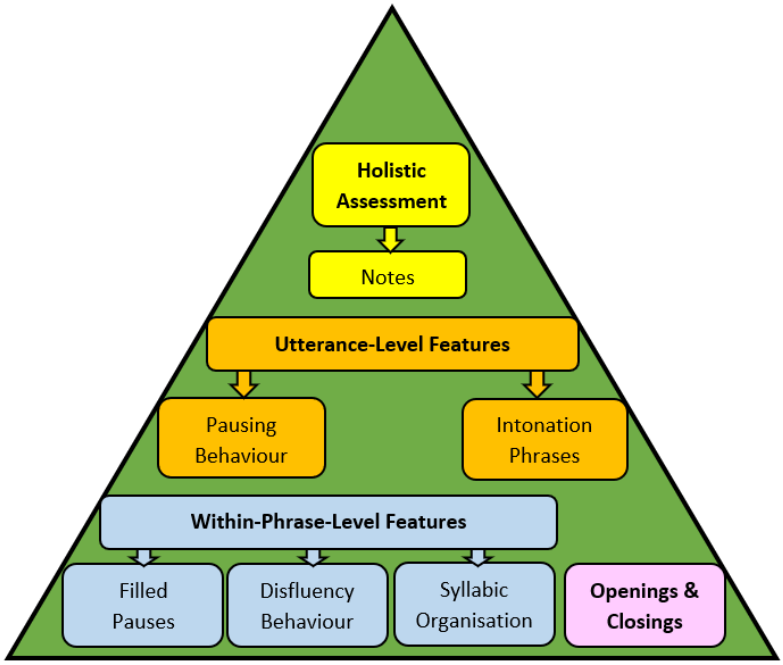


Figure 6.11. The different levels of the PARFA framework.

6.3.2.1. Holistic assessment of speech rhythm

The first stage of the assessment process is a holistic assessment of the speaker's speech rhythm. At this initial stage of the process, the assessor will take into account the entirety of the speaker's speech sample, or as much as deemed appropriate to account for the speaker's rhythmic patterns. This will likely involve the assessor listening to the sample numerous times. During this listening process, the assessor will make descriptive notes which conceptualise their holistic impression of the speaker's speech rhythm. As a means of assisting with how assessors should undertake this holistic assessment, the framework provides some suggested terms which the assessor should consider in conceptualising the perceived speech rhythm:

(1) Overall rhythmic feel

(2) Rhythmic patterning

(3) Rhythmic flow

(4) Rhythmic beat

The framework also provides some suggested terminology which the assessor might find useful when providing their descriptive assessment. These are terms which featured in the qualitative feedback from the perception experiment. These include:

(1) Balanced

(2) Sporadic

(3) Disjunct

(4) Pulsing

(5) Regular

(6) Bouncy

(7) Disfluent

(8) Fluent

(9) Unpredictable

(10) Monotonous

This set of terminological labels is by no means exhaustive, and assessors may wish to use their own descriptive labels when making their assessments.

6.3.2.2. Utterance-level features

The assessment of utterance-level features relates to the entirety of the speaking turn of the speaker being assessed. This speaking turn may be punctuated by pauses and disfluency phenomena and may consistent many or few individual intonation phrases. If there is an interlocuter featured within the speech sample, then the assessed speaker's turn (utterance) should be deemed to be complete when any such interlocuter takes up their turn.

Pausing behaviour (silent pauses)

Pausing behaviour is an important part of the PARFA framework. This feature, however, may be deemed by some to not be germane to the assessment of speech rhythm – rather, pausing is a separate component altogether. Its inclusion within the PARFA framework is based on the qualitative feedback from the perception experiments in the previous chapter where speakers' pausing behaviour was frequently reported by listeners as being one of the main features which aided in their identification of speakers (see Chapter 5, Section 5.4.3.2). In addition to this, there has been previous forensic phonetics research which has indicated that pausing behaviour can be useful as a speaker discriminant (e.g., Kolly et al., 2015; McDougall & Duckworth, 2017, 2018).

Duration

Observations made regarding the duration of silent pauses pertain to there being durational *regularity* or *variability* in these events. Observations of durational *regularity* would involve the speaker's within-turn silent pauses being relatively consistent durationally. For example, the speaker may primarily use pauses of a relatively short/long duration between intonation phrases, or the speaker may exhibit a *consistent* alternating pattern between longer and shorter pauses. Observations of durational *variability* would involve there being no consistent pattern with regards to the duration of the speaker's silent pause behaviour. For example, the speaker may alternate *sporadically* between longer pauses and shorter pauses with these pauses primarily being perceptually different durationally (e.g., pauses which are relatively short in duration, although being relatively short, would not perceptually be of notably similar durations).

Distribution

Observations made regarding the distribution of silent pauses pertain to there being distributional *regularity* or *variability* in these events. Observations of distributional *regularity* would involve the speaker's within-turn silent pauses being distributed relatively consistently throughout the utterance. For example, the speaker may primarily use pauses at regular intervals throughout an utterance, or they may exhibit a pattern of pausing towards the start and/or end of a given utterance. Observations of distributional *variability* would involve there being no consistent pattern with regards to where the speaker places their pauses within an utterance. For example, the speaker may alternate *sporadically* between using pauses at the beginning/middle/end of utterances.

Frequency

Observations made regarding the frequency of silent pauses pertain to the quantity of pauses used which can be judged as there being *many* or *few*. With regards to the

assessor judging what constitutes *many* versus what constitutes *few*, they may wish to consider the ratio of pauses in comparison to speech (e.g., how many pauses are being produced in comparison to how many within-utterance phrases being uttered by the speaker).

Interactions

Observations made regarding there being interactions (or cooccurrences) with silent pauses pertain to these either being *present* or *absent*. Observations of interactions being *present* would involve there being other phenomena which consistently cooccur alongside the speaker's use of silent pauses. For example, these may be instances of silent pauses always being followed by some form of disfluency phenomena such as a filled pause or a false start, or perhaps silent pauses are consistently preceded by a type of discourse marker or specific prosodic event such as rising intonation or a drop in intensity. Observations of interactions being *absent* would involve there being no consistent pattern of any other phenomena cooccurring with the speaker's use of silent pauses.

Intonation phrases

Duration

Observations made regarding the duration of intonation phrases pertain to there being durational *regularity* or *variability* in these events. Observations of durational *regularity* would involve the speaker's within-turn intonation phrases being relatively consistent durationally. For example, the speaker may speak for around five seconds, pause, speak for another similar durational period (i.e., 5 seconds), pause, and then conclude their turn with another durationally similar intonation phrase. Observations of durational *variability* would involve there being no consistent pattern with regards to the duration of the speaker's intonation phrases. For example, the speaker may alternate *sporadically* between longer and short intonation phrases or perhaps us

shorter intonation phrases at the beginning and end of a speaking turn with a longer stretches of speech mid-turn.

Intonation

Observations made regarding the intonation of intonation phrases pertain to there being *regularity* or *variability* in relation to the intonational patterning. Observations of intonational *regularity* would involve the speaker's within-turn intonation phrases being relatively consistent in terms of the intonation patterns exhibited. For example, the speaker's intonation may be relatively monotonous throughout an entire turn with no noteworthy variations or inflections with regards to pitch, or a speaker may exhibit a consistent intonational pattern such as always beginning a turn with a raised pitch or by placing an intonational inflections in a consistent manner within their intonation phrases. Observations of intonational *variability* would involve there being no consistent pattern with regards to the intonation patterns of the speaker's intonation phrases. For example, the speaker may alternate *sporadically* between using varying pitch patterns across intonation phrases or switch sporadically from being fairly monotonous to being more animated in terms of their variations and fluctuations in pitch.

Amplitude

Observations made regarding the amplitude of intonation phrases pertain to there being *regularity* or *variability* in relation to the amplitude patterns. Observations of amplitude *regularity* would involve the speaker's within-turn intonation phrases being relatively consistent in terms of the amplitude patterns exhibited. For example, the speaker's amplitude may be relatively stable throughout an entire turn with no noteworthy variations or inflections with regards to amplitude, or a speaker may exhibit a consistent amplitude pattern such as always beginning a turn with a raised amplitude or by decreasing their amplitude in a consistent manner towards the end of their intonation phrases. Observations of amplitude *variability* would involve there being no consistent pattern with regards to the amplitude patterns of the speaker's

intonation phrases. For example, the speaker may alternate *sporadically* between using varying amplitude patterns across intonation phrases or switch sporadically from being fairly quiet to speaking at a greater amplitude.

Opening cues

Observations made regarding there being opening cues within intonation phrases pertain to these either being *present* or *absent*. Observations of opening cues being *present* would involve there being specific phenomena which consistently occur at the beginning of a speaker's intonation phrase. For example, these may be instances of a speaker always commencing intonation phrases using a specific discourse marker or perhaps having a specific prosodic inflection at the start of a speaking turn (e.g., raised pitch). Observations of opening cues being *absent* would involve there being no consistent pattern of any specific phenomena occurring at the start of the speaker's intonation phrases.

Closing cues

Observations made regarding there being closing cues within intonation phrases pertain to these either being *present* or *absent*. Observations of closing cues being *present* would involve there being specific phenomena which consistently occur at the end of a speaker's intonation phrase. For example, these may be instances of a speaker always finishing intonation phrases using a specific discourse marker or perhaps having a specific prosodic inflection at the end of a speaking turn (e.g., raised pitch). Observations of closing cues being *absent* would involve there being no consistent pattern of any specific phenomena occurring at the end of the speaker's intonation phrases.

6.3.2.3. Within-phrase-level features

The assessment of within-phrase-level features relates to the phenomena which occur within the individual (intonation) phrases of the speaker who is being assessed. These

features may be individual speech units (e.g., individual phrase-final syllables) or clusters of speech units (e.g., complex disfluency phenomena) within a given intonation phrase.

Filled pauses

Duration

See 'Pausing behaviour (silent pauses)' in Section 6.3.2.2 above. The description provided there for *duration* is transferable to *duration* for filled pauses here.

Distribution

See 'Pausing behaviour (silent pauses)' in Section 6.3.2.2 above. The description provided there for *distribution* is transferable to *distribution* for filled pauses here.

Frequency

See 'Pausing behaviour (silent pauses)' in Section 6.3.2.2 above. The description provided there for *frequency* is transferable to *frequency* for filled pauses here.

Prosody

Observations made regarding the prosody of filled pauses pertain to there being *regularity* or *variability* in relation to the prosodic patterning of filled pauses. Observations of prosodic *regularity* would involve the speaker's filled pauses being relatively consistent in terms of the prosodic patterns exhibited. For example, the speaker's filled pauses may be *consistently* relatively monotonous with no noteworthy variations or inflections with regards to pitch or amplitude, or a speaker may exhibit a consistent prosodic pattern whereby their filled pauses are inflected with a raised pitch or by a lower amplitude. Observations of prosodic *variability* would involve there being no consistent pattern with regards to the prosodic patterning of filled

pauses. For example, the speaker may alternate *sporadically* between using varying pitch patterns across different filled pauses or switch sporadically from having fairly monotonous filled pauses to exhibiting filled pauses inflected with variations and fluctuations in pitch and/or amplitude.

Type

Observations made regarding the type of filled pauses pertain to there being *regularity* or *variability* in relation to the type of filled pauses used. Observations of *regularity* would involve the speaker's filled pauses all being either of the type *er* (i.e., a schwa-like tone with no final nasal portion) or of the type *erm* (i.e., a schwa-like tone with a final nasal portion). Observations of *variability* would involve there being no consistent pattern with regards to the speaker's use of the two different types of filled pause. For example, the speaker may alternate *sporadically* between using both *er* and *erm*.

Interactions

See 'Pausing behaviour (silent pauses)' in Section 6.3.2.2 above. The description provided there for *interactions* is transferable to *interactions* for filled pauses here.

Disfluency behaviour

Frequency

See 'Pausing behaviour (silent pauses)' in Section 6.3.2.2 above. The description provided there for *frequency* is transferable to *frequency* for disfluency behaviour here.

Distribution

See 'Pausing behaviour (silent pauses)' in Section 6.3.2.2 above. The description provided there for *distribution* is transferable to *distribution* for disfluency behaviour here.

Complexity

Observations made regarding the complexity of disfluency behaviour pertain to complex disfluent events either being *present* or *absent*. Observations of complex disfluency phenomena being *present* would involve the speaker using different types of disfluency features in combination with one another. For example, the speaker may exhibit multiple false starts, or a combination of repeated words/part-words perhaps with the addition of filled pauses or other disfluency phenomena. Observations of complex disfluency behaviour being *absent* would involve the speaker not exhibiting combinations of disfluencies features alongside one another. For example, the speaker may still exhibit false starts, however these events would be independent of any other disfluency phenomena and would not combine to form a complex disfluent event.

Type

Observations made regarding the type of filled pauses pertain to there being *regularity* or *variability* in relation to the type of disfluency features used. Observations of *regularity* would involve the speaker primarily using only one type of disfluency feature. For example, the speaker may exhibit whole-word repetitions at the start of intonation phrases but seldom use any other type of disfluency feature. Observations of *variability* would involve there being no consistent pattern with regards to the speaker's use of different types of disfluency features. For example, the speaker may alternate *sporadically* between using part-word repetitions, whole word-repetitions or self-interruptions.

Interactions

See ‘Pausing behaviour (silent pauses)’ in Section 6.3.2.2 above. The description provided there for *interactions* is transferable to *interactions* for disfluency behaviour here.

Syllabic organisation

Distribution

Observations made regarding the distribution of syllables pertain to there being distributional *regularity* or *variability*. Observations of distributional *regularity* would involve syllables being distributed relatively consistently throughout the speaker’s intonation phrases. For example, the speaker may primarily space (distribute) syllables at even intervals throughout an intonation phrase, or they may exhibit a pattern of spacing a cluster of syllables more tightly (i.e., exhibiting a quicker articulation rate) at the start of a phrase before syllables become more widely spaced (distributed) towards the middle and end of the intonation phrase. Observations of distributional *variability* would involve there being no consistent pattern with regards the speaker’s syllabic organisation. For example, the speaker may alternate *sporadically* between clustering syllables together tightly to spacing syllables at a greater distance apart at varying parts of an intonation phrase.

Prolongations

Observations made regarding there being syllabic prolongations pertain to these either being *present* or *absent*. Observations of prolongations being *present* would involve the speaker *consistently* prolonging syllables within intonation phrases. These syllabic prolongations may occur at any point during an intonation phrase. A syllable would be deemed to be prolonged if it stands out as being markedly durationally longer than the speaker’s usual syllable length. Observations of prolongations being *absent* would involve there being no consistent pattern of the speaker prolonging syllables in any marked or consistent fashion.

Prosody

Observations made regarding the prosody of syllables pertains to there being *regularity* or *variability* in relation to the prosodic patterning of syllables. Observations of prosodic *regularity* would involve the speaker being relatively consistent in terms of the prosodic patterns exhibited across clusters of syllables within an intonation phrase. For example, the speaker may be *consistently* relatively monotonous with no noteworthy variations or inflections with regards to pitch or amplitude across syllables, or a speaker may exhibit a consistent prosodic pattern whereby specific syllables are inflected with a raised pitch or by a lower amplitude. Observations of prosodic *variability* would involve there being no consistent pattern with regards to the prosodic patterning of syllables. For example, the speaker may alternate *sporadically* between using varying pitch patterns across different syllables or switch sporadically from having fairly monotonous syllables to exhibiting syllables inflected with variations and fluctuations in pitch and/or amplitude.

Rhythmic feel

Observations made regarding the 'rhythmic feel' of a speaker's syllabic organisation pertains to this being either *staccato*-like or *legato*-like in nature. Observations of a *staccato* rhythmic feel would involve syllables being delivered in a short/and or sharp manner creating a pulsing, punctuated feel. Observations of a *legato* rhythmic feel would involve syllables being delivered in a smooth and/or connected manner creating a more free-flowing rhythmic feel.

6.3.2.4. Openings and closings

The assessment of openings and closings relates specifically to any phenomena which occurs at the beginning or end of the individual intonation phrases of the speaker being assessed. Given the finding that a number of different prosodic and linguistic phenomena typically occur at the openings and closings of intonation phrases, and that such phenomena may have an influence on a speaker's perceived rhythmic

patterning, the PARFA framework affords an individual analytic section to these phrase locations.

Prosody

Pitch

Observations made regarding the pitch of openings and/or closings pertain to this being either *neutral* or *non-neutral*. Observations of pitch being *neutral* would involve the openings and/or closings of a speaker's intonation phrases showing no marked difference from the rest of the intonation phrase. Observations of the pitch being *non-neutral* would involve the openings and/or closings of a speaker's intonation phrases showing a marked variation from the rest of the intonation phrase. For example, this may be that the speaker tends to end phrases with a raised pitch or initiates phrases with a lowered pitch compared to the rest of the intonation phrase.

Amplitude

Observations made regarding the amplitude of openings and/or closings pertain to this being either *neutral* or *non-neutral*. Observations of amplitude being *neutral* would involve the openings and/or closings of a speaker's intonation phrases showing no marked difference from the rest of the intonation phrase. Observations of amplitude being *non-neutral* would involve the openings and/or closings of a speaker's intonation phrases showing a marked variation from the rest of the intonation phrase. For example, this may be that the speaker tends to end phrases with lowered amplitude or initiates phrases with greater amplitude compared to the rest of the intonation phrase.

Prolongations

Observations made regarding prolongations attributed with the openings and/or closings pertain to these being either *present* or *absent*. Observations of prolongations

being *present* would involve the openings and/or closings of a speaker's intonation phrases exhibiting marked syllabic lengthening. A syllable would be deemed to be prolonged if it stands out as being markedly durationally longer than the speaker's usual syllable length. Observations of prolongations being *absent* would involve there being no consistent pattern of syllabic lengthening.

Common occurrences

Disfluencies

Observations made regarding the cooccurrence of disfluency phenomena with openings and/or closings pertain to these being either *present* or *non-absent*. Observations of disfluency phenomena being *present* would involve the openings and/or closings of a speaker's intonation phrases *consistently* being punctuated with disfluency features. Observations of disfluency phenomena being *absent* would involve there being no consistent pattern of disfluency features cooccurring at the beginning or end of a speaker's intonation phrases.

Discourse markers

Observations made regarding the cooccurrence of discourse markers with openings and/or closings pertain to these being either *present* or *non-absent*. Observations of discourse markers being *present* would involve the openings and/or closings of a speaker's intonation phrases *consistently* being marked by the use of a specific type of discourse marker. For example, the speaker may consistently begin intonation phrases with speech units such as '*yeah*' or '*well*', or end intonation phrases with speech units such as '*you know'*. Observations of discourse markers being *absent* would involve there being no consistent pattern of discourse markers cooccurring at the beginning or end of a speaker's intonation phrases.

6.3.3. Initial PARFA framework design and modifications

Figure 6.12 below shows the initial draft PARFA framework.

		SCALAR RATING			Notes
	ATTRIBUTES	Slight	Mrkd.	Extrm.	
		1	2	3	
A. UTTERANCE LEVEL FEATURES					
Pausing behaviour	Durational regularity				
	Durational variability				
	Distributional regularity				
	Distributional variability				
	Frequency				
	Co-occurrences / cueing preferences				
Intonation phrases	Durational regularity				
	Durational variability				
	Intonational regularity				
	Intonational variability				
	Amplitude regularity				
	Amplitude variability				
	Opening cues				
	Closing cues				
B. PHRASE LEVEL FEATURES					
Filled pauses	Durational regularity				
	Durational variability				
	Distributional regularity				
	Distributional variability				
	Frequency				
	Co-occurrences / cueing preferences				
	Type variation / regularity				
	Prosodic inflection				
Disfluency behaviour	Frequency				
	Distributional regularity				
	Distributional variability				
	Co-occurrences / cueing preferences				
	Type variation / regularity				
	Complex vs. simple				
Syllabic delivery	Pace regularity				
	Pace variability				
	Durational regularity				
	Duration variability				
	Prolongations				
	Rhythmic 'feel' – staccato vs. legato				
	Accenting behaviour – prosodic influence				
C. OPENINGS / CLOSINGS					
Prosody	Pitch modulation (e.g., HRT)				
	Amplitude modulation (e.g., final fall)				
	Syllabic prolongations				
Common occurrences	Disfluency phenomena				
	Discourse markers				
D. HOLISTIC PERCEPTION OF SPEECH RHYTHM					
Rhythmic Feel	Regular rhythmic patterning / structure				
	Disjunct rhythmic patterning / structure				
	Balanced rhythmic flow				
	Sporadic / unpredictable flow				
	Regular / pulsing rhythmic beat				
	Energetic / lively / bouncy rhythm				
	Fluent / poised / confident delivery				
	Agitated / hurried / panicked rhythm				

Figure 6.12. Initial draft version of the PARFA framework.

The initial draft version of the PARFA framework incorporated a scalar rating for each of the attributes – slight (1), marked (2) and extreme (3). This scalar rating mimicked that of the modified VPA framework (San Segundo et al., 2019), however, following an initial consultation meeting with a forensic practitioner, it was determined that

these scalar ratings were superfluous for the rhythmic attributes observed and that their inclusion would likely lead to inconsistent results being obtained both within-analyst and between-analyst. As such, the scalar rating system was removed from the current version of the framework. It was also decided following this consultation that, owing to the nature of speech rhythm as being the combinations of a number of different speech characteristics which manifest over a period of time, that Section D, the *holistic assessment of speech rhythm*, would be better placed at the start of the assessment form, thus adopting a wide-scope to narrow-scope approach.

6.3.4. Additional notes on the PARFA framework

The following information is offered as general guidance which may be useful to analysts when completing the PARFA form.

6.3.4.1. When to use the PARFA framework

It is the intention that the PARFA framework should be applied within FVC casework in which two (or more) samples of speech are to be compared to one another. However, it is obvious that the framework will not be relevant (or compatible) for all FVC tasks. For example, in order to complete the initial section of the framework form, the analyst must be able to generate a general holistic description of the speaker's speech rhythm patterns. This may not be possible in some cases where perhaps the speech samples are too short in duration or where the speaker is only providing short answer responses or statements (e.g., a 'no comment' police interview). Similarly, there could be speech evidence where the speaker is continuously interrupted/disrupted by an interlocutor (e.g., a heated altercation), thus not allowing the speaker to establish any kind of 'rhythmic flow'. Other cases in which the application of PARFA may not be feasible include if there is a considerable mismatch between the samples being analysed. For example, if one speech sample consisted of read speech and the other was spontaneous, free-flowing speech it could be expected that the speaker(s) involved would adopt different rhythmic behaviours (e.g., Fraser & Mora, 2023; Kim & Jang, 2009). A further example of where mismatch

could be problematic for the application of PARFA would be if the speaker(s) involved were in highly contrastive emotional states. For example, this might involve a voicemail message in which the speaker is making a threat using very raised vocal effort throughout, in comparison to an early morning police interview in which the suspect is clearly subdued and speaking with very reduced vocal effort throughout. It should be mentioned, however, that instances such as these where the mismatch between samples is substantial, that many other speech features which would usually be analysed by a forensic practitioner (e.g., voice quality, vowel formants, pitch, etc.) are likely to also be rendered unsuitable for analysis (with this potentially resulting in the forensic voice comparison not being possible).

It is suspected that there may be some circumstances in which the PARFA framework may be particularly useful. For example, in the context of FVC case, the quality of both the questioned speech material and the known speech sample may be of poor technical quality, meaning the analysis of some speech characteristics is compromised. Depending on the degree of degradation, it is possible that taking acoustic measurements in terms of vowel formant measurements and acoustic pitch measurements is not possible. It may also be the case that the perceptual assessment of voice quality is not possible owing to the poor audio quality or perhaps another compromising factor such a speaker’s proximity to the recording device. However, in such cases, it is possible that the rhythm patterns of a speaker's speech could still be perceptually recognised, even if the lexical content is unclear or unintelligible. In situations where a forensic analyst is tasked with transcribing audio of poor quality featuring multiple speakers, the perceptual evaluation of speech rhythm could prove beneficial in attributing specific utterances to individual speakers. Therefore, even when the lexical content is ambiguous and other speech features (e.g., voice quality) are influenced by degradation, the discernible properties of speech rhythm could still provide useful insights for the forensic analyst, with the application of the PARFA framework further bolstering this perceptual analysis.

6.3.4.2. Listening

Given that the PARFA framework is a perceptual analysis tool, it is the analyst's listening skills which will determine the relative usefulness of the form's completion. The listening skills required to undertake PARFA analysis are somewhat different to those used in other kinds of perceptual analysis such as segmental phonetics and Vocal Profile Analysis. For example, segmental analysis relies on perceptual isolation of the features that differentiate each segment from its neighbouring sounds, where in Vocal Profile Analysis the task is to identify which features are common to all (or most) of the segments in a speech sample. For PARFA analysis, the analyst is making observations relating to the speaker's speech rhythm patterns, which will inevitably involve deciphering how a number of interrelating prosodic characteristics are combining to create an overall picture or *feel* of the speaker's rhythmic patterning. The PARFA analysis may be best thought of as being a two-stage analytical process which requires two different perceptual listening strategies. The first requires the ability to take a holistic approach to the listening task and form a general impression of the speaker's speech rhythm. This strategy will involve the listener ignoring the linguistic message and instead focus on the likely numerous speech features which combine across the entirety of the speech sample, and which ultimately give an impression of the speaker's speech rhythm. This strategy feeds into Section A of PARFA.

The second strategy involves the listener focussing on more specific and directed characteristics of the speaker's speech, thus narrowing their initial holistic assessment down in order to determine at which levels of speech and which features and attributes are those contributing towards their initial general impression. For some sections of PARFA, for example, Section C – the *within-phrase-level features* – this analysis relates more closely to a typical segmental analysis to some extent. For example, when making observations pertaining to SYLLABIC ORGANISATION, the analyst will be in essence focussing on individual (and relevant) syllables and their specific prosodic makeup. This therefore requires the auditory ability to isolate segments from the stream of continuous speech and hold them to memory long enough to analyse their perceptual attributes. This strategy feeds into Sections B, C and D of the PARFA form.

6.3.4.3. Marking observations

The PARFA framework form is not intended to be simply a 'tick the box' exercise. That is, a meaningful PARFA analysis does not require the analyst to mark observations for each and every feature and their associated attributes. Rather, following completion of the holistic assessment (Section A), the analyst will conduct more detailed and focussed listening to the most relevant parts of the speech sample in which different rhythmic features are present and in which they are deemed to be worthy of 'observation' (i.e., marking one of the binary choices). It should therefore be the case that if the analyst was posed the question as to why they have described the speaker's speech rhythm the way they have in the holistic assessment, they could then point towards the most relevant features where they have made observations in Sections B, C and D as being contributing factors which lead to the holistic description of the speaker's *overall rhythmic feel/rhythmic flow*.

On the far right-hand side of the PARFA form, is the column in which the analyst may choose to make any specific notes in relation to the observation that they have recorded. Adding notes in this column is not mandatory but may be useful in some circumstances for the analyst to elaborate on their perceived observation.

When completing the PARFA form, there may be some occasions in which there is overlap between marking one observation and marking another observation. Such instances may occur both within a specific section or across different sections of the form. This is best illustrated with an example. Figure 6.13 below shows an example of overlapping observations within the same section of the form.

	ATTRIBUTES	TICK TO INDICATE OBSERVATIONS			Notes
B. UTTERANCE-LEVEL FEATURES					
Pausing behaviour	Duration	Regularity ✓		Variability	
	Distribution	Regularity		Variability ✓	
	Frequency	Few ✓		Many	
	Interactions	Absent		Present ✓	Silent pauses often followed by filled pause
Intonation phrases	Duration	Regularity		Variability ✓	
	Intonation	Regularity		Variability	
	Amplitude	Regularity		Variability	
	Opening cues	Absent		Present ✓	Often begin with filled pause – generally of raised pitch
	Closing cues	Absent		Present	

Figure 6.13. Section B of the PARFA form showing within-section observational overlap.

As Figure 6.13 shows, the two marked observations highlighted within the red borders convey a degree of overlap in their nature. It is therefore important that the analyst is aware of this and doesn't 'double count' this particular feature when making comparisons with an associated speech sample. That is, care should be taken to ensure that any instances of observational overlap do not lead to the over-weighting of evidence.

6.3.5. Testing the PARFA framework

As has been shown above in relation to the TOFFA framework and the VPA framework variants (i.e., MVPA and SVPA), an important element following the initial development of a framework is testing its reliability in terms of agreement

between analysts and within analysts when using the framework. That is, ideally, a given framework would exhibit high degrees of inter- and intra-rater agreement. Realistically, perfect agreement is unachievable, but it is important to have an understanding of the levels of agreement that can be reached for a given framework. In order to determine whether the PARFA framework lends itself to such desirable rater agreement, initial testing of the framework would incorporate a 'testing session' in which a group of expert analysts (forensic phoneticians/practitioners) would be asked to rate a set of samples using the PARFA framework. In the first instance, this rating would be restricted to just the first section of the framework, that is, the holistic assessment section with the initial aim here being to look for consistency between the experts with regards to the impressionistic notes made and whether or not the suggested terminology is being utilised in a consistent manner. Initially, this would take the form of a closed-set scenario whereby there would be one speech sample provided per proposed descriptive label. For example, four speech samples from the WYRED corpus (voicemail task) would be selected by the present author, all of which would have something marked about their rhythmic patterning which would lend themselves to being labelled as either 'disjunct', 'bouncy', 'monotonous' or 'balanced'. Each of the participants in the initial testing session would be tasked with listening to the four speech samples and assigning each of the aforementioned descriptive labels to the 'correct' speech sample. It is, of course, important to point out that regardless as to whether participants all assigned the same labels to the same speech samples (i.e., indicating promising levels of inter-rater agreement) that this testing does not amount to ground-truth testing owing to the fact that such testing is not available for the perceptual assessment of speech rhythm. Rather, any inter-rater agreement or 'correct' responses should only ever be indicative of participants/raters having the same perception of a speaker's speech rhythm patterns and selecting the same label to describe these patterns.

At this stage of the testing procedure, given the setup of the closed-set scenario, it might also be possible to establish as to whether or not 'cardinal rhythm types' could be incorporated into the framework. Having a set of cardinal rhythm types could prove useful as these could act as reference points in a similar way to the VPA which provides description in relation to, for example, a *neutral* setting for various voice

quality features (e.g., *neutral* phonation features, *neutral* larynx height, *neutral* lingual tip/blade, etc.; see Beck (2007)). It is anticipated that the incorporation of cardinal rhythm types, if plausible, would likely be born out of the discussions held between the forensic experts who participated in this initial testing stage described above. One possible way in which this might be achieved is through expert collaboration with regards to detailing what were the key factors which contributed towards attributing each of the speech samples to the each of the four rhythm descriptors (i.e., disjunct', 'bouncy', 'monotonous' or 'balanced'). For example, for the sample that *should* have been 'correctly' attributed to the 'bouncy' label, could collaboration between analysts lead to an agreed-upon description for this 'bouncy' speech rhythm type? If so, this could ultimately result in a 'bouncy' cardinal rhythm type being established which could be used as a reference point and/or point of comparison for similar/contrastive speech rhythm types. Although establishing cardinal rhythm types through these means could potentially be possible, it is important to point out that, just because agreed-upon descriptions have been developed, the conceptualisation of these descriptions and how they are subsequently interpretated will be ultimately be subjective to a greater degree than, for example, the cardinal voice quality types/features which the VPA describes. This is owing to the fact that, for the VPA, a number of the descriptions provided are directly related to the physiology of certain articulators and are therefore more objective in their nature. For example, the *neutral* setting for the labial category is described as 'where the long-term average lip posture is as it would be for a schwa vowel, i.e. the lips are neither spread, nor rounded, nor protruded' (Beck, 2007: 6). It is clear, for the most part at least, that any such descriptions pertaining to cardinal speech rhythm types will unlikely be grounded in such direct anatomical/physiological terminology. Overall, at the present time, where testing has yet to carried out, it is somewhat difficult to provide a definitive approach regarding whether the inclusion of 'cardinal rhythm types' could be possible for the PARFA framework. Nevertheless, this is an area which holds promise for potential development and one which could assist in the application (and accessibility) of the PARFA framework.

Following on from this initial closed-set scenario testing, further testing would incorporate more samples being provided to participants/raters to allow for an open-

set scenario that reduces the likelihood of between-analyst consistencies just by chance and/or elimination. These samples would be from a range of different speakers from the WYRED corpus and would contain a range of different speech rhythm styles (selected by the present author). This open-set testing would also include the entirety of the PARFA framework form being available to the participants/raters (as opposed to just the initial holistic assessment section in the closed-set scenario testing). The participants/raters would be asked to use the PARFA framework form to assess the speech samples on two different occasions (two rounds) with a time lapse of one week. This 'two round' setup would be implemented in order to assess intra-rater reliability. Prior to undertaking their analyses, the raters will have completed the closed-set scenario analysis and will have subsequently carried out a group listening session in which they could discuss their experience of the closed-set task along with the strategies they adopted when listening to the samples and the observations that they each made. Following completion of both rounds of the open-set task, the degree of intra- and inter-rater agreement could be assessed, with any statistical testing only being relevant to Sections B, C and D of the PARFA form (i.e., where raters may indicate a binary choice observation).

On completion of these initial testing stages, any modifications that are deemed necessary in light of the results, as well as raters' feedback, could then be implemented. As a final testing protocol, the framework would then be trialled in a mock FVC case. In this mock scenario, two forensic practitioners would be served two (or more) speech samples in which the speech rhythm of the suspect and unknown speaker are a noteworthy feature. That is, where it would be expected that a forensic expert would likely comment on aspects of the speakers' speech rhythm patterns. The analysts would be instructed to complete a full comparison of the speech samples (i.e., not just analysis of the speakers' speech rhythm) and produce a final report as they would in a real-life case. This would allow for observations to be made with regards to the experts' use of the PARFA framework along with how observations marked on the framework factor into a final report. Feedback from the practitioners involved would then be collected with respect to the processes they implemented in using the framework, factoring their observations into their final report, as well as any specific challenges encountered in assessing the speakers' speech rhythm. This feedback

would then be considered in light of making any final modifications to the PARFA framework.

6.4. Chapter summary

The aim of the present chapter has been to introduce a new perceptual framework which can be used to assess speaker's speech rhythm patterns within the forensic domain. This framework is titled the Perceptual Assessment of Rhythm for Forensic Analysis (PARFA). In order to situate the proposed framework within the forensic context, the opening of the chapter presented and discussed a number of existing forensically orientated analytical frameworks. The purposes of these frameworks ranged from the analysis of disfluency phenomena to the assessment of spoofed speech, to the analysis of voice quality. One aspect which was intentionally highlighted when discussing these frameworks was the manner in which they were designed and tested (and also, on occasion, modified). The reasoning for drawing specific attention to these elements was in order to illustrate the steps that have been taken in the design and modification of the PARFA framework.

The design and layout of the framework has taken some inspiration from the VPA framework and its modified variants, whereas the content of the framework has been derived from the qualitative feedback obtained from expert listeners from the perceptual experiment featured in the previous chapter. Detailed guidance as well as some generalised information has been provided alongside the proposition of the framework to support the analyst.

Although the framework has yet to undergo its own testing, the testing/calibration methodologies implemented in the previously discussed frameworks would serve as a suitable model to follow for the PARFA framework (as discussed in Section 6.3.5). It is therefore evident that the next critical step towards the implementation and application of the PARFA framework is for these initial testing/calibration sessions to take place. Once this initial testing has been completed, it should be evident as to whether further modifications to the framework need to be made in order to make the

framework as practical as possible for the forensic analyst to make use of within FVC casework.

CHAPTER 7

Discussion and Conclusions

In this final chapter, the thesis is brought to its conclusion. In doing so, a summary of the thesis is provided, and further discussion is offered in relation to the opportunities for future research.

7.1. Thesis summary

Chapter 1 situated the research within the field of forensic speech science by first providing a summary of forensic voice comparison (FVC). Following this initial introduction to FVC, attention was focussed on the discrepancies between forensic phonetics research and FVC practice. Here it was shown that there were a number of ways in which forensic research could be more targeted towards FVC practice, with speech rhythm being shown to be one feature which could benefit from development.

Another important consideration that was discussed was ensuring that the methodologies and analytical procedures used in forensic research were orientated towards the types of analysis tasks found within FVC casework. The development of methodologies, particularly those which could support the forensic analyst's perceptual expertise, were suggested as being particularly favourable in aiding within FVC casework.

A review of the speech rhythm literature was presented in Chapter 2. To provide context for the present thesis, previous forensically-motivated speech rhythm research was foregrounded as well as speech rhythm research which has accounted for the three

main parameters – intensity, f_0 and duration – the production experiments in this thesis focussed on. Because this thesis has entertained the idea of analysing the speech rhythm of frequently occurring speech units in Chapter 4, previous literature on these units was also provided.

The purpose of Chapter 3 was to bring three different parameters of speech rhythm (intensity, f_0 and durational characteristics) to spontaneous speech in a single study. Statistical analysis in the form of linear discriminant analysis was carried out on the data to determine which measures carried the most speaker discriminatory potential. Results showed that the discriminatory power for all of the parameters was relatively weak overall, with classification rates only marginally surpassing chance level.

Overall, it was shown that the rhythm measurements (and metrics) applied do not transfer over well to the spontaneous speech condition. Attempting to measure speech rhythm using acoustic methods involves making comparisons across syllables that are different with respect to their phonetic content, level of stress, whole-utterance factors, etc.; all of which will contribute to the variables we are aiming to use to capture speech rhythm. In essence, these rhythm measures were shown to be too sensitive to the variation that spontaneous, content-mismatched speech contains. Although analysing these rhythmic parameters using laboratory-based, controlled speech material might allow for speaker discriminatory potential to be evidenced, little is to be gained from using these methods with speech material representative of that found in forensic casework. The experiments carried out in this chapter therefore helped to emphasise the need for forensic phonetics research to develop and test methodologies which are geared more towards the types of analysis tasks faced by forensic practitioners. Making use of the spontaneous speech data afforded by forensic databases was also shown to be a key consideration for future forensic research.

Because of the low performance results witnessed in Chapter 3, Chapter 4 aimed to apply some of the speech rhythm analysis techniques to the units of speech that could be expected to be most controlled between spontaneous speech samples. Chapter 4 therefore focused on four types of, so-called, “frequently occurring speech units”. These units were analysed in terms of their rhythmic characteristics, again measuring intensity, f_0 and duration characteristics. The measurements obtained were

subsequently normalised against the spontaneous speech utterance data presented in Chapter 3, allowing for the rhythmic characteristics of these units to be captured relative to the speakers' spontaneous speech patterns. Results showed that these speech units have substantially more speaker discriminatory power than the spontaneous utterances analysed in Chapter 3. Amongst the three rhythmic parameters analysed, intensity proved to be the most effective at distinguishing between speakers, followed by f_0 and then duration.

In Chapter 5, perception experiments were carried out to determine to what extent listeners (expert and non-expert) were able to discriminate between speakers predominantly based on features of speech rhythm. Speech samples were subjected to delexicalisation which foregrounded the rhythmic characteristics.

Results showed that expert listeners outperformed non-expert listeners, with expert listeners who had expertise in forensic phonetics generally performing better than those who did not. Listeners were also required to provide qualitative feedback in which they were asked to explain why they had selected specific delexicalised samples. This qualitative feedback therefore helped to elicit what features the listeners were tapping into when making their speaker identification assessments. Listeners reported a number of key characteristics which they were using when making their (correct) speaker identification assessments, with these lending to the development of meaningful descriptors of speech rhythm. These descriptors would subsequently be used to feed into the development of a perceptual rhythm framework for forensic speech analysis.

In Chapter 6, a new perceptual framework for the assessment of speech rhythm within the forensic context was introduced. This framework was titled the Perceptual Assessment of Rhythm for Forensic Analysis (PARFA). To contextualise the proposed framework within the forensic context, the initial section of the chapter examined several pre-existing forensic analytical frameworks. The ways in which these frameworks were designed and tested was made a focal point as a means of providing justification for the steps that were taken in the design and modification of the PARFA framework. In addition to the framework proposal, detailed guidance and general information were included to assist the analyst with regards to its application in

assessing spontaneous speech samples. Detail was also provided regarding how the framework should be initially tested. The development of the PARFA framework has therefore aimed to provide a structured way in which forensic practitioners can analyse spontaneous speech rhythm patterns within the auditory-phonetic and acoustic approach to FVC.

7.2. Opportunities for future research

7.2.1. Developing production experiments

The production experiments presented in Chapter 3 and Chapter 4 could be developed and extended in a number of forensically relevant ways. Firstly, owing to the affordances of the WYRED corpus, cross-style comparisons could be investigated. The speech material used for the experiments in Chapter 3 and Chapter 4 were derived from the mock police interview data (Task 1) of the WYRED corpus. Task 2 and Task 4 of WYRED are also of forensic relevance in that they feature 'suspect speakers' in a telephone conversation with an 'accomplice' (Task 2) and also leaving an incriminating voicemail message (Task 4). As Tasks 1, 2 and 4 all relate to the same fictious crime, there are a number of phrase iterations which occur across all three tasks (e.g., addresses, building names, specific directions/descriptions, etc.). Using the methods employed in Chapter 3 (i.e., using a contour-approach and a variability-approach) the rhythm patterns (i.e., measurements of intensity, f_0 and duration) associated with these specific phrases could be compared across the three different styles. The cross-style comparison of specific phrases would highlight the extent of within-speaker variation for the parameters under analysis in a content-matched scenario. Determining whether low with-speaker variability (a valued commodity for the forensic analyst) is evidenced within a content-matched context would be a useful starting point from which further cross-style comparisons could be launched.

Similarly, for Task 2 and Task 4, cross-channel comparisons could be made. For both of these tasks, there are two different audio qualities available from the WYRED corpus: studio quality and mobile telephone transmission quality. The experiments presented in Chapter 3 and Chapter 4 could be replicated using the data of either Task

2 or Task 4 with cross-channel comparisons being made between the rhythm measurements. This would allow for the assessment of how robust these measures are to the technical effects of mobile phone transmission – a key consideration within FVC practice.

7.2.2. Developing perception studies

The perception experiments presented in Chapter 5 demonstrated that listeners are able to make varying degrees of 'correct' speaker identification assessments, with levels of success dependent on factors such as participants being either expert or naïve listeners, specific expertise (of expert listeners), and the nature of the discrimination tasks (which varied over three sections of the experiments). Listeners reported making use of a number of key features relating to speakers' speech rhythm patterns when making their identification assessments. Of these features, speakers' pausing behaviour (silent pauses), use of filled pauses, and use of other disfluency features (e.g., word/part-word/phrase repetitions) were amongst the most frequently mentioned. Suggestions as to why these features were predominant in the qualitative feedback relate to the nature of the stimuli. The speech samples which featured in the perception experiments were taken from voicemail messages meaning there was no interlocuter and speakers were essentially performing a form of monologue (see Section 5.4.3.2 for further discussion). It would be of interest to see whether these features continue to predominate listeners' observations within different speaking styles, or whether other rhythmic features move to the forefront when, for example, an interlocuter is introduced.

As well as assessing which features listeners are tuning into when within different speaking styles, future research could look to determine the extent to which it is possible to make speaker identification assessments across different speaking styles. The present research did initially intend to incorporate tasks of this nature into the main perception experiment (what would have been Section Four), however initial testing of the Main Experiment revealed that having this fourth section would make the (already lengthy) experiment far too long and taxing to expect participants to not become fatigued (and may also have deterred participation). The setup for the unused

fourth section was similar to Section Two: a 30-second original sample from the opening of a voicemail message was provided along with two delexicalised samples, one of which contained the same speaker as the original sample. The delexicalised samples here were created from Task 2 of the WYRED corpus which featured speakers in a telephone conversation with an 'accomplice'. The telephone calls contained a female speaker (the 'accomplice') and male speaker, with the male speaker's voice being delexicalised. Both samples contained a net total of approximately 30 seconds of the males' delexicalised speech in alignment with the original (non-delexicalised) voicemail sample. As was the case in Section Two of the experiment, listeners would have therefore been faced with a binary decision between two samples. They would have also been asked to provide qualitative feedback as to why they made the decision that they did. Further variations of this task would have involved the original speech (previously non-delexicalised) sample also being delexicalised, meaning that listeners would have no access to any linguistic or voice quality information to use as potential guiding reference material. The results obtained from research along these lines would provide insight into the potential usefulness of comparing speakers' speech rhythm patterns when there is a mismatch between speaking style (as is predominantly the situation within FVC casework).

The delexicalised speech samples used throughout the perception experiments were created manually within Praat following the procedure outlined in Section 5.2.3. As acknowledged prior to describing the delexicalisation procedure, a number of different delexicalisation methods were initially tested to assess whether they could serve the purpose of the perception experiments. However, these methods were deemed problematic for reasons such as rhythmic characteristics being lost (e.g., misrepresentation of syllable lengths), and being too 'distracting' in their makeup (e.g., spectrally rotated speech samples). Nevertheless, it could be of interest to compare these signal manipulation methods with the procedure used presently to assess whether a specific delexicalisation method leads to 'better' listener performance. Obtaining qualitative feedback here would also be of interest to assess whether listeners' assessments are based on different rhythmic features depending on the delexicalisation method used.

Using the delexicalisation method applied in the present research, syllables were represented by schwa-like tones which foregrounded the rhythmic characteristics of the original samples. Given the length of the samples required for the tasks in the experiments (i.e., in being a relatively lengthy 30 seconds to allow for rhythmic patterns to be discerned), and the length of the experiments as a whole, the nature of the delexicalised stimuli (i.e., schwa-like tones) could have potentially exacerbated any listener fatigue in completing the experiments. That is, having delexicalised stimuli which were more 'speech-like' in their composition, and therefore potentially more auditorily 'pleasant', could have been preferable (at least for the listeners). Future research could look to develop and apply further delexicalisation methods which remain as speech-like as possible whilst also facilitating testing with regards to which features are most important for making ('correct') speaker identification assessments.

7.2.3. Links between production and perception experiments

The production and perception experiments carried out in this thesis used the same 20 speakers throughout. Where the spontaneous utterances of all 20 speakers were analysed in Chapter 3, some of these speakers were omitted from the analyses of the frequently occurring speech units in Chapter 4 due to not producing enough instances of the units. With regards to the perception experiment, again, not all of the speakers featured in all three sections (for the Pilot Study and Section One of the Main Experiment, this was a result of there being a limited number of tasks to include all of the speakers). Within the perception experiments, there was no set criteria with regards to which speakers would be the 'correct' choice in the binary decision tasks (i.e., which speakers featured as the original sample; sections one and two), nor was there criteria regarding which speakers would feature as same-speaker or different-speaker pairings (Section Three). In want of establishing any potential links between the production experiments and the perception experiments, for example, in determining whether speakers who exhibited the highest degrees of speaker-specificity in the production experiments were also aligned with 'correct' identification assessments in the perception experiments, future work would look to factor this into the

experimental design of the experiments. If there were 'standout' speakers in the production experiments, then these speakers could feature more prevalently in the perception experiments to determine whether their speech rhythm patterns also standout perceptually.

In addition, and as an extension of the suggestion above, further work could be carried out with regards to those speakers who were more easily distinguished by their spontaneous speech rhythm patterns or frequently occurring speech units. Such work could be perceptually based with qualitative observations being made in relation to a variety of aspects of the speakers' speech rhythm patterns. For example, if a speaker was distinguished comparatively well in terms of the intensity characteristics of their filled pauses, is this something which is salient to listeners when making perceptual assessments of their spontaneous speech. Establishing whether there is agreement between production and perception experiments within the context of speaker discrimination could lend support to certain features potentially being considered for acoustic analysis in terms of specific parameters (e.g., intensity or f_0).

7.2.4. New applications for the PARFA framework

In the penultimate section of Chapter 6 (Section 6.3.5), the proposed methodology for testing the PARFA framework was outlined. Following the initial testing stages, any subsequent modifications, and trialling the framework in a mock FVC case, further applications of the framework could then be considered. One such potential application is the use of PARFA in detecting AI-generated (or 'spoofed') speech. This could be within the context of a FVC task or for more generalised spoofed speech detection exercises. In Chapter 6 (Section 6.2.3), Lee et al.'s (2023) PASS framework for detecting the presence of voice spoofing artefacts in speech recordings was described. Lee et al.'s framework is divided into three categories: *auditory*, *visual*, and *acoustic-phonetic*. Under the auditory category, Lee et al. identify a label of *rhythmic quality* which they describe as an 'auditory label for the impression of an artificial rhythm, tempo, and metrical feet'. The acknowledgment here that detecting spoofed speech from a rhythmic perspective is primarily a perceptual (auditory) task provides an initial basis for using the PARFA framework for detecting spoofed speech. In their

framework (which is currently still in its development), Lee and colleagues provide no additional detail at present as to what constitutes a specific *rhythmic quality* as sounding 'artificial'. However, it is proposed here that the affordances of the PARFA framework could serve as a useful tool for identifying any such 'artificial' speech rhythm characteristics. As the threat of voice spoofing continues to grow, owing to the continued advancements in machine learning models, it is essential that the anti-spoofing and countermeasures research community have as many tools as possible at its disposal to combat this threat. At present, there has been little research carried out which has focussed on detecting spoofed speech from the perspective of speech rhythm. A search of the literature reveals that any efforts along these lines have been based predominantly within the automatic spoofing detection field. One such example is Lu et al.'s (2023) study which looked at exploiting the flaws of rhythm information that are inherent within text-to-speech-generated speech to increase the reliability of spoofing detection systems. Results from this study showed that the method Lu and colleagues developed to introduce rhythm artifacts into spoofed utterances significantly improved the detection of text-to-speech-generated speech in their dataset. Although it is encouraging to see such research focus on speech rhythm specifically, it remains that such work is founded on automatic methods – methods which may not be readily accessible or practically feasible under certain circumstances (e.g., within forensic casework). In addition, as was shown in the work of Kirchhübel and Brown (2022; discussed in Chapter 6, Section 6.2.3), the performance of an expert listener seems to surpass that of automatic methods. Therefore, adopting an auditory approach for the detection of spoofed speech is likely most preferable. This is arguably even more pertinent for the analysis of speech rhythm given its comparative complexity in terms of the multiple acoustic features (and interrelations of these) which feed into its makeup. Finally, returning to Lee et al.'s PASS framework, they also identify a label of *hyperflat prosody* under their acoustic-phonetic category. They describe this property as 'auditorily perceptible and acoustically analysable property that may be described as an overly level or flat prosodic pattern that is characteristic of 'robotic' speech'. Again, the affordances of the PARFA framework would allow any such *hyperflat prosody* evidenced in speech samples to be analysed in a structured and rigorous manner.

7.3. Conclusion

This thesis sought to develop the way in which speech rhythm is analysed within the forensic domain. In the first instance, production experiments were carried out which tested the tenability of assessing speech rhythm through acoustic means. The results obtained demonstrated that measuring spontaneous speech rhythm patterns using the parameters and methods employed herein yields little in the way of speaker discriminatory potential. However, measuring the rhythmic properties of specific speech units which occur frequently within spontaneous speech (relative to speakers' spontaneous speech patterns) was shown to be a potentially promising way in which we can start to measure, at least to some degree, spontaneous speech patterns.

In consideration of the overall results from the production experiments, it was determined that measuring speech rhythm through acoustic approaches, when applied to forensically realistic speech data, is generally not viable. The complexity of speech rhythm in terms of its acoustic properties is too sensitive to the specific types of speech data encountered in forensic contexts. Additionally, it is also suspected that these acoustic methods might overlook certain rhythmic details. As such, focus was shifted towards analysing speech rhythm via perceptual methods, with the notion that perception serves as a more effective means of eliciting additional relevant rhythmic features. Such is the case with voice quality being analysed perceptually within FVC casework, the present thesis subsequently adopted the stance that the analysis of speech rhythm should be dealt with in a similar way. As voice quality is analysed through the use of a recognised methodological framework, the decision was made to develop a new perceptual framework for the assessment of speech rhythm for use within the forensic domain.

In order to accomplish this aim, perception experiments were conducted which reported on the ability of listeners to distinguish between speakers based on delexicalised speech samples which foregrounded rhythmic attributes. Expert listeners performed better than non-expert listeners in the discrimination tasks, with expert listeners who possessed expertise in forensic phonetics being the overall

highest performing group. Qualitative feedback obtained from the listeners suggested that there were a number of standout features which were drawn on to make speaker identification assessments. This qualitative feedback was used as the basis for developing speech rhythm descriptors which were used in the development of the PARFA framework discussed below.

Finally, the thesis marked the introduction of a new perceptual framework proposed for application within the forensic domain. The Perceptual Assessment of Rhythm for Forensic Analysis (PARFA) framework is presented with the purpose of providing a structured approach for the analysis of spontaneous speech rhythm patterns, a commodity which has been absent within the auditory-phonetic and acoustic approach to forensic voice analysis. It is hoped that, following initial testing, this framework could be of use to forensic speech practitioners within forensic voice comparison casework.

Bibliography

Abercrombie, D. (1967). *Elements of general phonetics*. University Press.

Adami, A. G., Mihaescu, R., Reynolds, D. A., & Godfrey, J. J. (2003). Modelling prosodic dynamics for speaker recognition. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing 2003*, *4*, IV-788–791.

Adams, C. (1979). *English speech rhythm and the foreign learner*. Mouton.

Alku, P., Bäckström, T., & Vilkman, E. (2002). Normalized amplitude quotient for parametrization of the glottal flow. *The Journal of the Acoustical Society of America*, *112*(2), 701–710.

Allen, G. D. (1972). The Location of Rhythmic Stress Beats in English: An Experimental Study I. *Language and Speech*, *15*(1), 72–100.

Arvaniti, A. (2009). Rhythm, Timing and the Timing of Rhythm. *Phonetica*, *66*(1–2), 46–63.

Arvaniti, A. (2012). The usefulness of metrics in the quantification of speech rhythm. *Journal of Phonetics*, *40*(3), 351–373.

Arvaniti, A., & Ross, T. (2010). Rhythm classes and speech perception. *Proceedings of Speech Prosody 2010*, Chicago, 11-14 May 2010.

Asadi, H., Nourbakhsh, M., He, L., Pellegrino, E., & Dellwo, V. (2018). Between-speaker rhythmic variability is not dependent on language rhythm, as evidence from Persia reveals. *International Journal of Speech Language and the Law*, *25*(2), 151–174.

Attorresi, L., Salvi, D., Borrelli, C., Bestagini, P., & Tubaro, S. (2022). *Combining Automatic Speaker Verification and Prosody Analysis for Synthetic Speech Detection* (arXiv:2210.17222). arXiv.

Audhkhasi, Kandhway, K., Deshmukh, O. D., & Verma, A. (2009). Formant-based technique for automatic filled pause detection in spontaneous spoken English, *International Conference on Acoustics, Speech and Signal Processing*, 2009, 4857-4860.

Auer, P. (1993). *Is a rhythm-based typology possible? A study of the role of prosody in phonological typology*. Doctoral Thesis. University of Freiburg.

Barry, W., Andreeva, B., & Koreman, J. (2009). Do Rhythm Measures Reflect Perceived Rhythm? *Phonetica*, *66*(1–2), 78–94.

Bartkova, K., Gac, D. L., Charlet, D., & Jouvet, D. (2002). Prosodic parameter for speaker identification. *7th International Conference on Spoken Language Processing (ICSLP 2002)*, 1197–1200.

Bartle, A., & Dellwo, V. (2015). Auditory speaker discrimination by forensic phoneticians and naive listeners in voiced and whispered speech. *International Journal of Speech Language and the Law*, *22*(2), 229–248.

Beck, J. (2005). Perceptual analysis of voice quality: the place of Vocal Profile Analysis. In *A Figure of Speech: a Festchrift for John Laver*, pp. 285-322, London.

Beck, J. (2007). *Vocal Profile Analysis Scheme: A User's Manual*, Queen Margaret University College-QMUC, Speech Science Research Centre.

Beckman, M. E., & Pierrehumbert, J. (1986). Intonational structure in Japanese and English. *Phonology Yearbook, 3*, 255-309.

Benus, S., Enos, F., Hirschberg J. & Shriberg, E. (2006). Pauses in deceptive speech. In *Speech Prosody. Volume 18*, pages 2–5.

Benus, S., Gravano, A. & Hirschberg, J. (2007). The prosody of backchannels in American English. In: *Proceedings of the 16th International Conference of Phonetic Sciences*, 1065–1068.

Bertinetto, P. M. (1980). The perception of stress by Italian speakers. *Journal of Phonetics*, 8, 385-395.

Bertinetto, P. M., & Bertini, C. (2008). On modeling the rhythm of natural languages. *Speech Prosody 2008*, 427–430.

Betz, S., Eklund, R. & Wagner, P. (2017). Prolongation in German. Proceedings of DiSS 2017, 18–19 August 2017, Royal Institute of Technology, Stockholm Sweden.

Birkholz, P., Kroger, B. J., & Neuschaefer-Rube, C. (2011). Model-Based Reproduction of Articulatory Trajectories for Consonant–Vowel Sequences. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(5), 1422–1433.

Blankenship, J., & Kay, C. (1964). Hesitation Phenomena in English Speech: A Study in Distribution. *WORD, 20*(3), 360–372.

Blesser, B. (1972). Speech Perception Under Conditions of Spectral Transformation: I. Phonetic Characteristics. *Journal of Speech and Hearing Research*, *15*(1), 5–41.

Bloch, B. (1950). Studies in Colloquial Japanese IV Phonemics. *Language*, *26*(1), 86.

Boersma, P., (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proc. Inst. Phonetic Sci.* (Univ. Amsterdam) 17, 97–110

Boersma, P. & Weenink, D. (2020). Praat: Doing Phonetics by Computer (version 5.3.65). http://www.praat.org.

Bolinger, D. (1965). Pitch accent and sentence rhythm. In I. Abe & T. Kanekiyo (Eds.), *Forms of English: Accent, morpheme, order* (pp. 139–180). Harvard University Press.

Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., & Brennan, S. E. (2001). Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender. *Language and Speech*, *44*(2), 123–147.

Braber, N., Smith, H., Wright, D., Hardy, A., & Robson, J. (2023). Assessing the Specificity and Accuracy of Accent Judgments by Lay Listeners. *Language and Speech*, *66*(2), 267–290.

Bradshaw, L., Hughes, V., & Chodroff, E. (2020). Investigating the Forensic Applications of Global and Local Temporal Representations of Speech for Dialect Discrimination. *Speech Prosody 2020*, 635–639.

Brander, D. (2014). Phonetic characteristics of hesitation vowels in Swiss German and their use for forensic phonetic speaker identification. Poster presented at the annual conference of the *International Association for Forensic Phonetics and Acoustics*. Zürich, Switzerland.

Braun, A. (1995). Fundamental frequency – How speaker-specific is it? BEIPHOL 64, *Studies in Forensic Phonetics*, 9–23.

Braun, A., Elsässer, N., & Willems, L. (2023). Disfluencies Revisited—Are They Speaker-Specific? *Languages*, *8*(3), 155.

Brennan, S. E., & Schober, M. F. (2001). How Listeners Compensate for Disfluencies in Spontaneous Speech. *Journal of Memory and Language*, *44*(2), 274–296.

Brewster, T. (2021). Fraudsters Cloned Company Director's Voice In $35 Million Bank Heist, Police Find. Available from: https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/?sh=7dfbccf67559.

Butterworth, B. (1980). Evidence from pauses in speech. In: Butterworth, B. (Ed.), *Language Production Volume 1: Speech and Talk*. Academic Press, London, pp. 155–176.

Caldwell, M., Andrews, J. T. A., Tanay, T., & Griffin, L. D. (2020). AI-enabled future crime. *Crime Science*, *9*(1), 14.

Carmo, S., Rehder, M. I. B. C., Almeida, L. N., Villegas, C., Dantas, C. R. V., Vasconcelos, D., & Andrade, E. (2023). Forensic analysis of auditorily similar voices. *Revista CEFAC*, *25*(2), e4022.

Carroll, L. A. (2019a). *Profiling Disfluency: A Forensic Analysis of Disfluency Behaviour in Spontaneous Speech*. Unpublished manuscript. Lancaster University. Lancaster, UK.

Carroll, L. A. (2019b). *A critical analysis of the taxonomy of fluency features for forensic analysis (TOFFA): a classification system for the analysis of disfluency features in forensic speaker comparison cases*. Unpublished manuscript. Lancaster University. Lancaster, UK.

Carroll, L. A. (2019c). *A forensic phonetic investigation of disfluency behaviour across two interactional styles: applying a modified analytical framework TOFFAMo*. Unpublished MA thesis. Lancaster University. Lancaster, UK.

Chan, R. K. W. (2023). Evidential value of voice quality acoustics in forensic voice comparison. *Forensic Science International*, *348*, 111725.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The Natural Statistics of Audiovisual Speech. *PLoS Computational Biology*, *5*(7).

Cichocki, W., Selouani, S.-A., Perreault, Y. (2014). Measuring rhythm in dialects of New Brunswick French: Is there a role for intensity? *Canadian Acoustics* 42, 90-91.

Clark, H. H. & Fox Tree, J. E. (2002). Using uh and um in spontaneous speech. *Cognition*, 84: 73-111.

Corley, M., MacGregor, L. J., & Donaldson, D. I. (2007). It's the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition*, *105*(3), 658–668.

Crystal, D. (1985). *A dictionary of linguistics and phonetics* (2nd ed.). B. Blackwell in association with A. Deutsch.

Cumming, R. E. (2011). The Language- Specific Interdependence of Tonal and Durational Cues in Perceived Rhythmicality. *Phonetica*, *68*(1–2), 1–25.

Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics*, *11*(1), 51–62.

Dauer, R. (1987). Phonetic and Phonological Components of Language Rhythm. Paper presented at the *11th International Congress of Phonetic Sciences*, Tallinn, Estonia, 1st-7th August 1987.

De Boer, M. M., & Heeren, W. F. L. (2020). Cross-linguistic filled pause realization: The acoustics of *uh* and *um* in native Dutch and non-native English. *The Journal of the Acoustical Society of America*, *148*(6), 3612–3622.

De Boer, M. M., Quené, H., & Heeren, W. F. L. (2022). Long-term within-speaker consistency of filled pauses in native and non-native speech. *JASA Express Letters*, *2*(3), 035201.

De Nil, L. F., & Abbs, J. H. (1991). Influence of speaking rate on the upper lip, lower lip, and jaw peak velocity sequencing during bilabial closing movements. *The Journal of the Acoustical Society of America*, *89*(2), 845–849.

Delgado, H., Evans, N., Kinnunen, T., Lee, K. A., Liu, X., Nautsch, A., Patino, J., Sahidullah, M., Todisco, M., Wang, X., & Yamagishi, J. (2021). *ASVspoof 2021: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan*. arXiv:2109.00535.

Dellwo, V. (2006). *Rhythm and Speech Rate: A Variation Coefficient for deltaC*. In: Karnowski, P; Szigeti, I. Language and language-processing. Frankfurt/Main: Peter Lang, 231-241.

Dellwo, V. (2008). The role of speech rate in perceiving speech rhythm. *Speech Prosody 2008*, 375–378.

Dellwo, V. (2009). Choosing the right rate normalization methods for measurements of speech rhythm, in *Proceedings of AISV*, 13–32.

Dellwo, V. (2010). *Influences of speech rate on the acoustic correlates of speech rhythm: An experimental phonetic study based on acoustic and perceptual evidence*. PhD thesis, University of Bonn, Bonn.

Dellwo, V. (2015). What does voice and silence tell us about speaker identity? An introduction to temporal speaker individualities and their use for forensic speaker comparison. In G. M. Schnider, M. C. Janner and B. Élie (eds). Vox and Silentium 17–35. Bern: Peter Lang.

Dellwo, V. and Fourcin, A. (2013). Rhythmic characteristics of voice between and within languages. *Travaux Neuchâtelois de Linguistique 59*: 87–107.

Dellwo, V., Fourcin, A., & Abberton, E. (2007). Rhythmical classification of languages based on voice parameters. In J. Trouvain, & W. J. Barry (Eds.), *Proceedings of the XVI International Congress of Phonetic Sciences* (1129–1132). Saarbrucken: University of Saarland.

Dellwo, V., Huckvale, M., & Ashby, M. (2007). How Is Individuality Expressed in Voice? An Introduction to Speech Production and Description for Speaker Classification. In C. Müller (Ed.), *Speaker Classification I* (Vol. 4343, 1–20). Springer Berlin Heidelberg.

Dellwo, V. & Koreman, J. (2008). How speaker idiosyncratic is measurable speech rhythm? Paper presented at the *International Association for Forensic Phonetics and Acoustics Annual Conference*, Lausanne, 20–23 July 2008.

Dellwo, V., Leemann, A. and Kolly, M.-J. (2012). Speaker idiosyncratic rhythmic features in the speech signal. In *Proceedings of INTERSPEECH 2012*. 1582–1585. Portland, USA.

Dellwo, V., Leemann, A., & Kolly, M.-J. (2015). Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors. *The Journal of the Acoustical Society of America*, *137*(3), 1513–1528.

Dellwo, V., Steiner, I., Aschenberner, B., Dankovičova, J. & Wagner, P. (2004). The BonnTempo-Corpus and BonnTempo-Tools: a database for the study of speech rhythm and rate. In *Proceedings of the 8th ICSLP/INTERSPEECH 2004*. 777–780. Jeju Island, Korea.

Deterding, D. (2001). The measurement of rhythm: a comparison of Singapore and British English. *Journal of Phonetics, 29,* 217-230.

Donovan, A., & Darwin, C. J. (1979). *The Perceived Rhythm of Speech*. 9th International Congress of Phonetic Sciences, Copenhagen, Denmark.

Drummond, K., & Hopper, R. (1993). Back Channels Revisited: Acknowledgment Tokens and Speakership Incipiency. *Research on Language & Social Interaction*, *26*(2), 157–177.

Duarte, D., Galves, A., Garcia, N. L., & Maronna, R. (2001). *The statistical analysis of acoustic correlates of speech rhythm*. Paper presented at the Workshop on Rhythmic Patterns, Parameter Setting and Language Change, ZiF, University of Bielefeld.

Eklund, R. (2004). *Disfluency in Swedish human–human and human–machine travel booking dialogues*. Linköping University Studies in Science and Technology Dissertation No. 88.

Erdemir, A., Walden, T. A., Tilsen, S., Mefferd, A. S., & Jones, R. M. (2023). A Preliminary Study of Speech Rhythm Differences as Markers of Stuttering Persistence in Preschool-Age Children. *Journal of Speech, Language, and Hearing Research: JSLHR*, *66*(3), 931–950.

Fairclough, L., Brown, G. & Kirchhübel, C. (2023). Reviewing the performance of formants for forensic voice comparison: a meta-analysis of forensic speech science research. In: Radek Skarnitzl & Jan Volín (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences* (3834–3838). Guarant International.

Faure, G., Hirst, D. J., & Chafcouloff, M. (1980). Rhythm in English: Isochronism, pitch and perceived stress. In L. Waugh R. & C. H. van Schooneveld (Eds.), *The Melody of Language* (71–79). University Park Press.

Ferragne, E. (2008). *Étude phonétique des dialectes modernes de l'anglais des Îles Britanniques: vers l'identification automatique du dialect*. Unpublished doctoral thesis, Université Lyon 2, Lyon.

Finlayson, I. R., & Corley, M. (2012). Disfluency in dialogue: An intentional signal from the speaker? *Psychonomic Bulletin & Review*, *19*(5), 921–928.

Foulkes, P., & Barron, A. (2000). Telephone speaker recognition amongst members of a close social network. *International Journal of Speech, Language and the Law*, *7*(2), 180–198.

Foulkes, P., Carrol, G. and Hughes, S. (2004). Sociolinguistics and acoustic variability in filled pauses. Paper presented at the annual conference of the *International Association for Forensic Phonetics and Acoustics*. Helsinki, Finland.

Foulkes, P., & French, P. (2012). Forensic Speaker Comparison: A Linguistic–Acoustic Perspective. In L. M. Solan & P. M. Tiersma (Eds.), *The Oxford Handbook of Language and Law* (1st ed., pp. 558–572). Oxford University Press.

Fowler, C. A. (1979). "Perceptual centers" in speech production and perception. *Perception & Psychophysics*, *25*(5), 375–388.

Fox Tree, J. E. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. Journal of Memory and Language 34.709–38.

Fraisse, P. (1963). The psychology of time. New York: Harper & Row.

Fraisse, P. (1982). Rhythm and Tempo. In D. Deutsch (Ed.), *Psychology of Music* (pp. 149-180). New York and London: Academic Press.

Fraundorf, S. H., & Watson, D. G. (2011). The disfluent discourse: Effects of filled pauses on recall. *Journal of Memory and Language*, *65*(2), 161–175.

French, P., Harrison, P., Kirchhübel, C., Rhodes, R. & Wormald, J. (2017). From receipt of recordings to dispatch of report: Opening the blinds on lab practices. Paper presented at the *International Association for Forensic Phonetics and Acoustics conference*, Split, Croatia.

Fuchs, R. (2014). Towards a perceptual model of speech rhythm: Integrating the influence of f0 on perceived duration. *Interspeech 2014*, 1949–1953.

Fuchs, R. (2015). *Speech rhythm in varieties of english: Evidence from educated Indian English and British English*. Springer Berlin Heidelberg.

Fujii, S., & Wan, C. Y. (2014). The Role of Rhythm in Speech and Language Rehabilitation: The SEP Hypothesis. *Frontiers in Human Neuroscience*, *8*.

Gardner, R. (1998). Between Speaking and Listening: The Vocalisation of Understandings1. *Applied Linguistics*, *19*(2), 204–224.

Ghez, C. & Krakauer, J. (2000). The organization of movement. In: *Principles of neural science* (Kandel ER, Schwartz JH, Jessell TM, eds), pp 653–673. New York: McGraw-Hill.

Gibb-Reid, B., Foulkes, P. & Hughes, V. (2022). Exploring the phonetic variation of yeah and like. *York Papers In Linguistics* (YPL2). 1-27.

Gibbon, D., & Gut, U. (2001). *Measuring Speech Rhythm*. Paper presented at Eurospeech 2001, Aalborg, Denmark, 3rd-7th September 2001.

Gold, E. & French, P. (2011). International practices in forensic speaker comparison, *International Journal of Speech, Language and the Law* 18(2), 293-307.

Gold, E., & French, P. (2019). International practices in forensic speaker comparisons: second survey. *International Journal of Speech, Language and the Law*, 26(1), 1-20.

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. *Interspeech 2018*, 2748–2752.

Goldman-Eisler, F. (1961a). A Comparative Study of two Hesitation Phenomena. *Language and Speech*, *4*(1), 18–26.

Goldman-Eisler, F. (1961b). The Distribution of Pause Durations in Speech. *Language and Speech*, *4*(4), 232–237.

Goldman-Eisler, F. (1968). *Psycholinguistics: Experiments in Spontaneous Speech*. Academic Press, London.

Gósy, M. (2023). Occurrences and Durations of Filled Pauses in Relation to Words and Silent Pauses in Spontaneous Speech. *Languages*, *8*(1), 79.

Gósy, M., Bóna, J., Beke, A., & Horváth, V. (2014). Phonetic characteristics of filled pauses: The effects of speakers' age. *Proceedings of the 10th International Seminar on Speech Production (ISSP)* (150–153).

Grabe, E., & Low, E. L. (2002). Durational variability in speech and the Rhythm Class Hypothesis. In N. Warner & C. Gussenhoven (Eds.), *Papers in Laboratory Phonology VII* (pp. 515-543). Berlin: Mouton de Gruyter.

Green, T., Rosen, S., Faulkner, A., & Paterson, R. (2013). Adaptation to spectrally-rotated speech. *The Journal of the Acoustical Society of America*, *134*(2), 1369–1377.

Grivicic, T., & Nilep, C. (2004). When phonation matters: The use and function of yeah and creaky voice. *Colorado Research in Linguistics*, 17, 1–11.

Grosjean, F. (1980). Temporal variables within and between languages. In: Dechert, H. W., Raupach, M. (eds), *Towards a Cross-Linguistic Assessment of Speech Production*. Frankfurt: Lang, pp. 39–53.

Handel, S. (1993). *Listening: An introduction to the perception of auditory events* (1st edition). MIT Press.

Harrington, L., Rhodes, R. & Hughes, V. (2021). Style variability in disfluency analysis for forensic speaker comparison. *International Journal of Speech, Language and the Law*, *28*(1), 31–58.

Harsin, C. A. (1997). Perceptual-center modeling is affected by including acoustic rate-of-change modulations. *Perception & Psychophysics*, *59*(2), 243–251.

He, L. (2012). Syllabic intensity variations as quantification of speech rhythm: Evidence from both L1 and L2. In Q. Ma, H. Ding, & D. Hirst (Eds.), *Proceedings of the 6th International Conference on Speech Prosody* (466–469). Shanghai, China: Tongji University Press.

He, L. (2018). Development of speech rhythm in first language: The role of syllable intensity variability. *The Journal of the Acoustical Society of America*, *143*(6).

He, L., & Dellwo, V. (2016). The role of syllable intensity in between-speaker rhythmic variability. *International Journal of Speech Language and the Law*, *23*(2), 243–273.

He, L., & Dellwo, V. (2017). Between-speaker variability in temporal organizations of intensity contours. *The Journal of the Acoustical Society of America*, *141*(5).

He, L., & Dellwo, V. (2017). Amplitude envelope kinematics of speech: Parameter extraction and applications. *The Journal of the Acoustical Society of America*, 141, 3582.

Herment, S. (2012). *Is intensity a relevant criterion in the perception of spontaneous speech? The case of emphasis in English*. ⟨hal-01489663⟩.

Hoequist, C. E. (1983). The Perceptual Center and Rhythm Categories. *Language and Speech*, *26*(4), 367–376.

Hoequist, Jr., C. (1983). Syllable Duration in Stress-, Syllable- and Mora-Timed Languages. *Phonetica*, *40*(3), 203–237.

Hollien, H. (1990). *The Acoustics of Crime*. Springer US.

Holm, S. (2008). *Intonational and durational contributions to the perception of foreign-accented Norwegian. An experimental phonetic investigation*. PhD thesis, Norwegian University of Science and Technology, Trondheim.

Howell, P. (1984). An acoustic determinant of perceived and produced isochrony. Paper presented at the *10th International Congress of Phonetic Sciences* Utrecht, The Netherlands, 1st-6th August 1984.

Hsu, C.-C., Hwang, H.-T., Wu, Y.-C., Tsao, Y., & Wang, H.-M. (2017). Voice conversion from non-parallel corpora using variational autoencoder. In: *Signal and Information Processing Association Annual Summit and Conference (APSIPA),* 2016 Asia-Pacific, 1–6.

Hsu, W.-N., Zhang, Y., & Glass, J. (2017). Learning latent representations for speech generation and transformation, in *Proc. InterSpeech*, 2017, 1273–1277.

Hudson, T., de Jong, G., McDougall, K., Harrison, P. & Nolan, F. (2007). F0 Statistics for 100 Young Male Speakers of Standard Southern British English. Paper presented at the *16th International Congress of Phonetic Sciences*: ICPhS XVI, Saarbrücken, Germany, August 6–10; 1809–12.

Hughes, A., Trudgill, P., & Watt, D. (2013). *English accents & dialects* (Fifth edition). Routledge, Taylor & Francis Group.

Hughes, V. (2014). *The Definition of the Relevant Population and the Collection of Data for Likelihood Ratio-Based Forensic Voice Comparison*. PhD Thesis, University of York.

Hughes, V. (2017). Sample size and the multivariate kernel density likelihood ratio: How many speakers are enough? *Speech Communication*, *94*, 15–29.

Hughes, V., Cardoso, A., Foulkes, P., French, P., Gully, A., & Harrison, P. (2023). Speaker-specificity in speech production: The contribution of source and filter. *Journal of Phonetics*, *97*, 101224.

Hughes, V., Cardoso, A., French, P., Harrison, P., & Gully, A. (2019). 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., Wood, S., & Foulkes, P. (2016). Strength of forensic voice comparison evidence from the acoustics of filled pauses. *International Journal of Speech Language and the Law*, *23*(1), 99–132.

Illa, A., & Ghosh, P. K. (2020). The impact of speaking rate on acoustic-to-articulatory inversion. *Computer Speech & Language*, *59*, 75–90.

Ishihara, S. & Kinoshita, Y. (2010). Filler Words as a Speaker Classification Feature. *Proc. SST 2010*, 34-37.

Jassem, W., Hill, D. R., & Witten, I. H. (1984). Isochrony in English Speech: Its Statistical Validity and Linguistic Relevance. In D. Gibbon & H. Richter (Eds.), *Intonation, Accent and Rhythm* (pp. 203–225). DE GRUYTER.

Jefferson, G. (1984). Notes on a systematic deployment of the acknowledgement tokens “Yeah”; and “Mm Hm”; *Paper in Linguistics*, *17*(2), 197–216.

Jefferson, G. (2002). Is “no” an acknowledgment token? Comparing American and British uses of (+)/(−) tokens. *Journal of Pragmatics*, *34*(10–11), 1345–1383.

Jessen, M. (2018). Forensic voice comparison. In J. Visconti (Ed.), *Handbook of Communication in the Legal Sphere* (pp. 219–255). De Gruyter.

Jessen, M., Köster, O. & Gfroerer, S. (2005). Influence of vocal effort on average and variability of fundamental frequency. *Speech, Language and the Law* 12, 174–212.

Jessen, M., Koster, O., & Gfroerer, S. (2005). Influence of vocal effort on average and variability of fundamental frequency. *International Journal of Speech, Language and the Law*, *12*(2), 174–213.

Johnson, K. (2012). *Acoustic and auditory phonetics* (3rd ed.). Wiley-Blackwell.

Johnson, N. F. (1975). On the function of letters in word identification: Some data and a preliminary model. *Journal of Verbal Learning and Verbal Behavior*, *14*(1), 17–29.

Kakouros, S., Räsänen, O., & Alku, P. (2018). Comparison of spectral tilt measures for sentence prominence in speech—Effects of dimensionality and adverse noise conditions. *Speech Communication*, *103*, 11–26.

Kaushik, M., Trinkle, M., Hashemi-Sakhtsari, A. (2010). Automatic detection and removal of disfluencies from spontaneous speech. Proc. 13th Australasian Int. Conf. on Speech Science and Technology Melbourne, 98-101.

Kawahara, H., Masuda-Katsuse, I., & De Cheveigné, A. (1999). Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication*, *27*(3–4), 187–207.

Kawahara, H., & Zolfaghari, P. (2001). Systematic F0 glitches around nasal-vowel transitions. *7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, 2459–2462.

Kinoshita, Y., Ishihara, S., & Rose, P. (2009). Exploring the Discriminatory Potential of F0 Distribution Parameters in Traditional Forensic Speaker Recognition. *International Journal of Speech, Language and the Law*, *16*(1), 91–111.

Kirchhübel, C., & Brown, G. (2022). Spoofed speech from the perspective of a forensic phonetician. *Interspeech 2022*, 1308–1312.

Kisler, T., Reichel U. D., & Schiel, F. (2017). Multilingual processing of speech via web services, *Computer Speech & Language, Volume 45*, September 2017, 326–347.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *The Journal of the Acoustical Society of America*, *59*(5), 1208–1221.

Kochanski, G., Grabe, E., Coleman, J., & Rosner, B. (2005). Loudness predicts prominence: Fundamental frequency lends little. *The Journal of the Acoustical Society of America*, *118*(2), 1038–1054.

Kohler, K. J. (2009). Rhythm in Speech and Language. *Phonetica*, *66*(1–2), 29–45.

Kolly, M.-J., Boula De Mareüil, P., Leemann, A., & Dellwo, V. (2017). Listeners use temporal information to identify French- and English-accented speech. *Speech Communication*, *86*, 121–134.

Kolly, M.-J., & Dellwo, V. (2014). Cues to linguistic origin: The contribution of speech temporal information to foreign accent recognition. *Journal of Phonetics*, *42*, 12–23.

Kolly, M.-J., Leemann, A., & Dellwo, V. (2014). Foreign accent recognition based on temporal information contained in lowpass-filtered speech, in *Proceedings of Interspeech 2014*, September 14–18, Singapore, 2175–2179.

Köster, S. (2002). Acoustic-phonetic aspects of Lombard Speech for different text styles. *The Phonetician*, vol. 85, 9–16.

Kreiman, J., & Gerratt, B. R. (2011). Comparing Two Methods for Reducing Variability in Voice Quality Measurements. *Journal of Speech, Language, and Hearing Research*, *54*(3), 803–812.

Kreiman, J., Gerratt, B. R., & Ito, M. (2007). When and why listeners disagree in voice quality assessment tasks. *The Journal of the Acoustical Society of America*, *122*(4), 2354–2364.

Kreiman, J., Gerratt, B. R., Kempster, G. B., Erman, A., & Berke, G. S. (1993). Perceptual Evaluation of Voice Quality: Review, Tutorial, and a Framework for Future Research. *Journal of Speech, Language, and Hearing Research*, *36*(1), 21–40.

Krivokapić, J. (2014). Gestural coordination at prosodic boundaries and its role for prosodic structure and speech planning processes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1658), 20130397.

Künzel, H. J. (1989). How Well Does Average Fundamental Frequency Correlate with Speaker Height and Weight? *Phonetica*, *46*(1–3), 117–125.

Künzel, H. J. (2001). Beware of the 'telephone effect': the influence of telephone transmission on the measurement of formant frequencies. *Forensic Linguistics* 8, 80–99.

Large, E. W., & Jones, M. R. (1999). The dynamics of attending: How we track time varying events. *Psychological Review*, 106 (1), 119-159.

Laver, J. (1980). The Phonetic Description of Voice Quality. Cambridge: Cambridge University Press.

Laver, J. (1994). *Principles of Phonetics* (1st ed.). Cambridge University Press.

Lee, C. S., & Todd, N. P. M. (2004). Towards an auditory account of speech rhythm: Application of a model of the auditory 'primal sketch' to two multi-language corpora. *Cognition*, *93*(3), 225–254.

Lee, D., McDougall, K., Kelly, F. & Alexander, A. (2023). PASS (Phonetic Assessment of Spoofed Speech): Towards human-expert-based framework for spoofed speech detection. Paper presented at the International Association for Forensic Phonetics and Acoustics Annual Conference, Zurich, July 9-12, 2023.

Lee-Goldman, R. (2011). No as a discourse marker. *Journal of Pragmatics*, *43*(10), 2627–2649.

Leeman, A., Mixdorff, H., O'Reilly, M., Kolly, M.-J., & Dellwo, V. (2015). Speaker-individuality in Fujisaki model f0 features: Implications for forensic voice comparison. *International Journal of Speech Language and the Law*, *21*(2), 343–370.

Leemann, A., & Kolly, M.-J. (2015). Speaker-invariant suprasegmental temporal features in normal and disguised speech. *Speech Communication*, *75*, 97–122.

Leemann, A., Kolly, M.-J., & Dellwo, V. (2014). Speaker-individuality in suprasegmental temporal features: Implications for forensic voice comparison. *Forensic Science International*, *238*, 59–67.

Leemann, A., Kolly, M.-J., Nolan, F., & Li, Y. (2018). The role of segments and prosody in the identification of a speaker's dialect. *Journal of Phonetics*, *68*, 69–84.

Lehiste, I. (1970). *Suprasegmentals*. Cambridge, MA: MIT Press.

Lehiste, I. (1977). Isochrony reconsidered. *Journal of Phonetics 5*: 253–263.

Lester, R. A., Barkmeier-Kraemer, J., & Story, B. H. (2013). Physiologic and Acoustic Patterns of Essential Vocal Tremor. *Journal of Voice*, *27*(4), 422–432.

Lindh, J., & Eriksson, A. (2007). Robustness of long time measures of fundamental frequency. *Interspeech 2007*, 2025–2028.

Liu, S., & Takeda, K. (2021). Mora-timed, stress-timed, and syllable-timed rhythm classes: Clues in English speech production by bilingual speakers. *Acta Linguistica Academica*.

Llisterri, J., Machuca, M., de la Mota, C., Riera, M., & Rios, A. (2003). *The perception of lexical stress in Spanish*. Paper presented at the 15th International Congress of Phonetic Sciences, Barcelona, Spain, 3rd-9th August 2003.

Lloyd James, A. (n.d.). *Bible Readings*. Linguaphone.

Lloyd James, A. (1940). *Speech Signals in Telephony*. Sir Isaac Pitman & Sons.

Low, E. L. (1994). *Intonation Patterns in Singapore English*. Unpublished MPhil thesis, University of Cambridge, Cambridge.

Low, E. L. (1998). *Prosodic Prominence in Singapore English.* Unpublished doctoral thesis, University of Cambridge, Cambridge.

Lu, J., Zhang, Y., Wang, W., Shang, Z., & Zhang, P. (2023). *Enhancing Spoofing Speech Detection Using Rhythm Information* (arXiv:2310.12014). arXiv.

MacGregor, L. J., Corley, M., & Donaldson, D. I. (2010). Listening to the sound of silence: Disfluent silent pauses in speech have consequences for listeners. *Neuropsychologia*, *48*(14), 3982–3992.

Machado, C. L. (2021). A cross-linguistic study of between-speaker variability in intensity dynamics in L1 and L2 spontaneous speech. In C. Bernardasci, D. Dipino, D. Garassino, S. Negrinelli, E. Pellegrino, & S. Schmid (Eds.), *L'individualità del parlante nelle scienze fonetiche: Applicazioni tecnologiche e forensi* (Vol. 8, 157–174). Officinaventuno.

Mai, K. T., Bray, S., Davies, T., & Griffin, L. D. (2023). Warning: Humans cannot reliably detect speech deepfakes. *PLOS ONE*, *18*(8), e0285333.

Malisz, Z. (2013). *Speech Rhythm Variability in Polish and English: A Study of Interaction between Rhythmic Levels*. PhD dissertation, Adam Mickiwicz University.

McClelland, E. (2008). Voice recognition within a closed set of family members. Paper presented at the International Association for Forensic Phonetics and Acoustics Annual Conference, Lausanne, 20–23 July 2008.

McDougall, K. (2013). Assessing perceived voice similarity using Multidimensional Scaling for the construction of voice parades. *International Journal of Speech Language and the Law*, *20*(2), 163–172.

McDougall, K., & Duckworth, M. (2017). Profiling fluency: An analysis of individual variation in disfluencies in adult males. *Speech Communication*, *95*, 16–27.

McDougall, K. & Duckworth, M. (2018). Individual patterns of disfluency across speaking styles: a forensic phonetic investigation of Standard Southern British English. International Journal of Speech Language and the Law. 25. 205-230.

McDougall, K., Rhodes, R., Duckworth, M., French, J. P., Kirchhübel, C. and Wormald, J. (2019). Application of the 'TOFFA' framework to the analysis of disfluencies in forensic phonetic casework. In Calhoun, S., Escudero, P., Tabain, M. and Warren, P. (eds.) Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019 (pp. 731-725).

McGlashan, J., & Fourcin, A. (2008). Objective evaluation of the voice. In M. Gleeson (Ed.), *Scott-Brown's Otorhinolaryngology: Head and Neck Surgery 7Ed* (pp. 2170–2191). CRC Press.

Miller, M. (1984). On the perception of rhythm. *Journal of Phonetics*, *12*(1), 75–83.

Mirsky, Y., Demontis, A., Kotak, J., Shankar, R., Gelei, D., Yang, L., Zhang, X., Pintor, M., Lee, W., Elovici, Y., & Biggio, B. (2023). The Threat of Offensive AI to Organizations. *Computers & Security*, *124*, 103006.

Mok, P. (2009). On the syllable-timing of Cantonese and Beijing Mandarin. *Chinese Journal of Phonetics*, *2*, 148–154.

Morrill, R. J., Paukner, A., Ferrari, P. F., & Ghazanfar, A. A. (2012). Monkey lipsmacking develops like the human speech rhythm. *Developmental Science*, *15*(4), 557–568.

Morrison, G. S. (2014). Distinguishing between forensic science and forensic pseudoscience: Testing of validity and reliability, and approaches to forensic voice comparison. *Sci. Justice 54*, 245–256.

Morton, J., Marcus, S., & Frankish, C. (1976). Perceptual centers (P-centers). *Psychological Review*, *83*(5), 405–408.

Nespor, M. (1990). On the rhythm parameter in phonology. In I. M. Roca (Ed.), *Logical Issues in Language Acquisition* (pp. 157–176). De Gruyter.

Niebuhr, O. (2009). F_0-Based Rhythm Effects on the Perception of Local Syllable Prominence. *Phonetica*, *66*(1–2), 95–112.

Niebuhr, O., & Winkler, J. (2017). The Relative Cueing Power of F0 and Duration in German Prominence Perception. *Interspeech 2017*, 611–615.

Nolan, F. (1983). *The Phonetic Bases of Speaker Recognition*. Cambridge: CUP.

Nolan, F., & Jeon, H.-S. (2014). Speech rhythm: A metaphor? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1658), 20130396.

Nolan, F., McDougall, K., De Jong, G., & Hudson, T. (2009). The DyViS database: Style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech, Language and the Law*, *16*(1), 31–57.

Nolan, F., McDougall, K., & Hudson, T. (2011). Some acoustic correlates of perceived (dis)similarity between same-accent voices. In ICPhS, (pp. 1506–1509).

O'Connor, J. (1965). *The perception of time intervals* (Progress Report 2; pp. 11–15). Department of Phonetics, University College London.

Pallier, C. (2002). Computing discriminability and bias with the R software. Internet download from: www.pallier.org/ressources/aprime/aprime.pdf.

Pellegrino, E. (2019). The effect of healthy aging on within-speaker rhythmic variability: A case study on Noam Chomsky. *Loquens*, *6*(1), 060.

Pereira, A. S., Kavanagh, E., Hobaiter, C., Slocombe, K. E., & Lameira, A. R. (2020). Chimpanzee lip-smacks confirm primate continuity for speech-rhythm evolution. *Biology Letters*, *16*(5), 20200232.

Perrier, P. (2012). Gesture planning integrating knowledge of the motor plant's dynamics: A literature review for motor control and speech motor control, in Speech Planning and Dynamics, edited by S. Fuchs, M. Weirich, D. Pape, and P. Perrier. Peter Lang, Frankfurt, Germany, pp. 191–238.

Petyt, K. M. (1985). *'Dialect' and 'Accent' in Industrial West Yorkshire* (Vol. G6). John Benjamins Publishing Company.

Pike, K. (1945). *The Intonation of American English*. University of Michigan.

Plant, R. L., & Younger, R. M. (2000). The interrelationship of subglottic air pressure, fundamental frequency, and vocal intensity during speech. *Journal of Voice*, *14*(2), 170–177.

Pointon, G. E. (1980). Is Spanish really syllable-timed? *Journal of Phonetics*, *8*(3), 293–304.

Polyanskaya, L., Busà, M. G., & Ordin, M. (2020). Capturing Cross-linguistic Differences in Macro-rhythm: The Case of Italian and English. *Language and Speech*, *63*(2), 242–263.

Polyanskaya, L., Ordin, M., & Busa, M. G. (2017). Relative Salience of Speech Rhythm and Speech Rate on Perceived Foreign Accent in a Second Language. *Language and Speech*, *60*(3), 333–355.

Pompino-Marschall, B. (1989). On the psychoacoustic nature of the P-centre phenomenon. *Journal of Phonetics*, 17, 175-192.

Prieto, P., Vanrell, M., Astruc, L., Payne, E., & Post, B. (2012). Phonotactic and phrasal properties of speech rhythm. Evidence from Catalan, English, and Spanish. *Speech Communication. 54*, 681–702.

Quené, H., & Van Delft, L. E. (2010). Non-native durational patterns decrease speech intelligibility. *Speech Communication*, *52*(11–12), 911–918.

R Core Team (2019). R: A Language and Environment for Statistical Computing (version 3.1.1). R Foundation for Statistical Computing, Vienna.

Ramus, F. (2002). *Acoustic Correlates of Linguistic Rhythm: Perspectives*. Paper presented at Speech Prosody 2002, Aix-en-Provence, France, 11th-13th April 2002.

Ramus, F., & Mehler, J. (1999). Language identification with suprasegmental cues: A study based on speech resynthesis. *The Journal of the Acoustical Society of America*, *105*(1), 512–521.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, *73*(3), 265–292.

Ramus, F., Nespor, M., & Mehler, J. (2000). Correlates of linguistic rhythm in the speech signal. *Cognition*, *75*(1), AD3–AD30.

Remez, R. E., Fellowes, J. M., & Rubin, P. E. (1997). Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(3), 651–666.

Riazantseva, A. (2001). Second Language Proficiency and Pausing: A Study of Russian Speakers of English. *Studies in Second Language Acquisition*, *23*(4), 497–526.

Roach, P. (1982). On the distinction between 'stress-timed' and 'syllable-timed' languages. In D. Crystal (Ed.), *Linguistic Controversies: Essays in honour of F.R. Palmer* (pp. 73–79). Edward Arnold.

Robertson, B. & Vignaux, G. A. (1995). *Interpreting Evidence: Evaluating Forensic Science in the Courtroom*, John Wiley & Sons Ltd.

Rose, P. (1996). Between- and within-Speaker variation in the fundamental frequency of Cantonese citation tones. In P. Davies & F. Neville (Eds.), *Vocal Fold Physiology: Controlling Complexity and Chaos* (pp. 307–324). Singular Press.

Rose, P. (2002). *Forensic Speaker Identification*. Taylor & Francis, London.

Rose, P. & Morrison, G. S. (2009). A response to the UK position statement on forensic speaker comparison. *International Journal of Speech, Language and the Law 16*, 139–163.

Rose, R. & Watanabe, M. (2019). Crosslinguistic corpus study of silent and filled pauses: when do speakers use filled pauses to fill pauses? In Sasha Calhoun, Paola Escudero, Marija Tabain and Paul Warren (eds.) *Proceedings of the 19th International Congress of Phonetic Sciences*, Melbourne, Australia 2019 (pp. 2615-2619).

San Segundo, E., Foulkes, P., French, P., Harrison, P., Hughes, V., & Kavanagh, C. (2019). The use of the Vocal Profile Analysis for speaker characterization: Methodological proposals. *Journal of the International Phonetic Association*, *49*(3), 353–380.

San Segundo, E., & Mompean, J. A. (2017). A Simplified Vocal Profile Analysis Protocol for the Assessment of Voice Quality and Speaker Similarity. *Journal of Voice*, *31*(5), 644.e11-644.e27.

San Segundo, E., & Skarnitzl, R. (2021). A Computer-Based Tool for the Assessment of Voice Quality Through Visual Analogue Scales: VAS-Simplified Vocal Profile Analysis. *Journal of Voice*, *35*(3), 497.e9-497.e21.

Sautermeister, P., & Eklund, R. (1997). *Some Observations on the Influence of F0 and Duration to the Perception of Prominence by Swedish Listeners*. Paper presented at FONETIK, Sweden, 1997.

Schröder, M., Charfuelan, M., Pammi, S., & Steiner, I. (2011). Open source voice creation toolkit for the MARY TTS platform. *Interspeech 2011*, 3253–3256.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech Recognition with Primarily Temporal Cues. *Science*, *270*(5234), 303–304.

Shattuck-Hufnagel, S., & Turk, A. (1998). The domain of phrase-final lengthening in English. *The Journal of the Acoustical Society of America*, *103*(5_Supplement), 2889–2889.

Shen, Y., & Peterson, G. G. (1962). Isochronism in English. *Studies in Linguistics: University of Buffalo Occasional Papers*, *9*, 1-36.

Shriberg, E. (1994). *Preliminaries to a theory of speech disfluencies*. PhD thesis, University of California at Berkeley.

Shriberg, E. (1996). Disfluencies in Switchboard. Proceedings, *International Conference on Spoken Language Processing, Addendum*, 3–6 October 1996, Philadelphia, PA, 11–14.

Shriberg, E. (2001). To 'errrr' is human: ecology and acoustics of speech disfluencies. *Journal of the. International Phonetic Association.* 31, 153–169.

Shriberg, E. and Lickley, R. (1993). Intonation of clause-internal filled pauses. *Phonetica 50*.172–9.

Shue, Y.-L., P. Keating, P., Vicenik, C. & Yu, K. (2011). VoiceSauce: A program for voice analysis, *Proceedings of the ICPhS XVII*, 1846-1849.

Silipo, R. & Greenberg, S. (1999). "Automatic transcription of prosodic stress for spontaneous English discourse," in *Proceedings of the XIVth International Congress of Phonetic Sciences (ICPhS99),* pp. 2351–2354.

Silipo, R. & Greenberg, S. (2000). "Prosodic stress revisited: Reassessing the role of fundamental frequency," in *Proceedings of the NIST Speech Transcription Workshop*.

Skarnitzl, R., & Vaňková, J. (2017). Fundamental frequency statistics for male speakers of Common Czech. *AUC PHILOLOGICA*, *2017*(3), 7–17.

Smith, E. E. (1967). Effects of Familiarity on Stimulus Recognition and Categorization. *Journal of Experimental Psychology*, *74*(3), 324–332.

Sóskuthy, M. (2017). *Generalised additive mixed models for dynamic analysis in linguistics: A practical introduction*. arXiv:1703.05339.

Sóskuthy, M. (2021). Evaluating generalised additive mixed modelling strategies for dynamic speech analysis. Journal of Phonetics, 84, 101017.

Stouten, F., & Martens, J. P. (2003). A feature-based filled pause detection system for Dutch. *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721)*, 309–314.

Stupp, C. (2019). Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case. Available from: https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402.

Swerts, M. (1998). Filled pauses as markers of discourse structure. *Journal of Pragmatics*, *30*(4), 485–496.

Tajima, K., Port, R.F., & Dalby, J. (1997). Effects of temporal correction on intelligibility of foreign-accented English. *Journal of Phonetics*, 25, 1-24.

Terblanche, C., Harrison, P., & Gully, A. J. (2021). Human Spoofing Detection Performance on Degraded Speech. *Interspeech 2021*, 1738–1742.

Tilsen, S., & Arvaniti, A. (2013). Speech rhythm analysis with decomposition of the amplitude envelope: Characterizing rhythmic patterns within and across languages. *The Journal of the Acoustical Society of America*, *134*(1), 628–639.

Toda, T., Black, A. W., & Tokuda, K. (2005). Spectral Conversion Based on Maximum Likelihood Estimation Considering Global Variance of Converted Parameter. *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, *1*, 9–12.

Todisco, M., Wang, X., Vestman, V., Sahidullah, M., Delgado, H., Nautsch, A., Yamagishi, J., Evans, N., Kinnunen, T., & Lee, K. A. (2019). *ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection*.

Torgersen, E. N., & Szakay, A. (2012). An investigation of speech rhythm in London English. *Lingua*, *122*(7), 822–840.

Trask, R. L. (2006). *A dictionary of phonetics and phonology*. Routledge.

Traunmüller, H., & Eriksson, A. (2000). Acoustic effects of variation in vocal effort by men, women, and children. *The Journal of the Acoustical Society of America*, *107*(6), 3438–3451.

Trofimovich, P., & Baker, W. (2006). Learning Second Language Suprasegmentals: Effect of L2 Experience on Prosody and Fluency Characteristics of L2 Speech. *Studies in Second Language Acquisition*, *28*(01).

Trouvain, J., & Truong, K. P. (2012). Acoustic, Morphological, and Functional Aspects of "yeah/ja" in Dutch, English and German. *Proceedings of the Interdisciplinary Workshop on Feedback Behaviors*, 77–80.

Trubet͡skoĭ, N. S. (1958). *Principles of phonology*. University of California Press.

Truong, K. P., & Heylen, D. (2010). Disambiguating the functions of conversational sounds with prosody: The case of 'yeah'. *Interspeech 2010*, 2554–2557.

Tschäpe, N., Trouvain, J., Bauer, D. & Jessen, M. (2005). Idiosyncratic patterns of filled pauses. Paper presented at the annual conference of the *International Association for Forensic Phonetics and Acoustics*. Marrakesh, Morocco.

Turk, A., & Shattuck-Hufnagel, S. (2013). What is speech rhythm? A commentary on Arvaniti and Rodriquez, Krivokapić, and Goswami and Leong. *Laboratory Phonology*, *4*(1).

Uebersax, J. S. (1987). Diversity of decision-making models and the measurement of interrater agreement. *Psychological Bulletin*, *101*(1), 140–146.

Vaissière, J. (1991). Rhythm, accentuation and final lengthening in French. In J. Sundberg, L. Nord, & R. Carlson (Eds.), *Music, Language, Speech and Brain* (pp. 108–120). Macmillan Education UK.

Van Dommelen, W. A. (1987). The Contribution of Speech Rhythm and Pitch to Speaker Recognition. *Language and Speech*, *30*(4), 325–338.

Van Riper, C. (1973). *The Treatment of Stuttering*. Prentice-Hall, Englewood Cliffs, NJ.

Venables, W. N. & Ripley, B. D. (2002). *Modern Applied Statistics with S*, Fourth edition. Springer, New York.

Verkhodanova, V. & Shapranov, V. (2016). Experiments on Detection of Voiced Hesitations in Russian Spontaneous Speech. *Journal of Electrical and Computer Engineering*, vol. 2016.

Watt, D., & Tillotson, J. (2001). A spectrographic analysis of vowel fronting in Bradford English. *English World-Wide. A Journal of Varieties of English*, *22*(2), 269–303.

Webb, A. L., Carding, P. N., Deary, I. J., MacKenzie, K., Steen, N., & Wilson, J. A. (2004). The reliability of three perceptual evaluation scales for dysphonia. *European Archives of Oto-Rhino-Laryngology*, *261*(8).

Wenk, B. J., & Wioland, F. (1982). Is French really syllable-timed? *Journal of Phonetics*, *10*(2), 193–216.

Whalen, D. H., Chen, W.-R., Shadle, C. H., & Fulop, S. A. (2022). Formants are easy to measure; resonances, not so much: Lessons from Klatt (1986). *The Journal of the Acoustical Society of America*, *152*(2), 933–941.

White, L., & Mattys, S. L. (2007). Calibrating rhythm: First language and second language studies. *Journal of Phonetics*, *35*(4), 501–522.

Wieland, E. A., McAuley, J. D., Dilley, L. C., & Chang, S.-E. (2015). Evidence for a rhythm perception deficit in children who stutter. *Brain and Language*, *144*, 26–34.

Wiget, L. White, B. Schuppler, I. Grenon, O. Rauch, & S. L. Mattys, "How stable are acoustic metrics of contrastive speech rhythm?," J. Acoust. Soc. Am. 127, 1559–1569 (2010).https://doi.org/10.1121/1.3293004

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *The Journal of the Acoustical Society of America*, *91*(3), 1707–1717.

Wilson, J. V., & Leeper, H. A. (1992). Changes in laryngeal airway resistance in young adult men and women as a function of vocal sound pressure level and syllable context. *Journal of Voice*, *6*(3), 235–245.

Wingate, M. E. (1964). A Standard Definition of Stuttering. *Journal of Speech and Hearing Disorders*, *29*(4), 484–489.

Winters, S., & O'Brien, M. G. (2013). Perceived accentedness and intelligibility: The relative contributions of F0 and duration. *Speech Communication*, *55*(3), 486–507.

Wirz, S. & Mackenzie Beck, J. (1995). Assessment of voice quality: The vocal profile analysis scheme. In Sheila Wirz (ed.), *Perceptual approaches to communication disorders*, pp. 39–55. London: Whurr.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R*, Second Edition (CRC Press, Boca Raton, FL).

Woodrow, H. (1951). Time Perception. in S. S. Stevens (Ed.) *Handbook of Experimental Psychology*, New York, Wiley, pp. 1225–1226.

Wu, C.-H., & Yan, G.-L. (2004). Acoustic Feature Analysis and Discriminative Modeling of Filled Pauses for Spontaneous Speech Recognition. *The Journal of VLSI Signal Processing-Systems for Signal, Image, and Video Technology*, *36*(2/3), 91–104.

Xu, Y. (2013). ProsodyPro — A tool for large-scale systematic prosody analysis. In *Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013)* (pp. 7–10). Aix-en-Provence, France.

Yamagishi, J., Wang, X., Todisco, M., Sahidullah, M., Patino, J., Nautsch, A., Liu, X., Lee, K. A., Kinnunen, T., Evans, N., & Delgado, H. (2021). ASVspoof 2021: Accelerating Progress in Spoofed and Deepfake Speech Detection. *Proceedings of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, pp. 47–54, September 2021.

Zdravković, S., Jovičić, S. (2020). The importance of speech pauses for psychotherapeutic and forensic observations. *International scientific conference "Archibald Reiss days"*, Belgrade, Vol. 10, 513-524.

Zhang, P., Chen, Y., Li, M., Zhao, H., Zhang, J., Wang, F., & Wu, X. (2023). Speech Spoofing Detection Based on Graph Attention Networks with Spectral and Temporal Information. *ACM Multimedia Asia 2023*, 1–7.

Zhang, Y., He, L., & Dellwo, V. (2019). Speaker individuality in the durational characteristics of voiced intervals: the case of Chinese bi-dialectal speakers. In: *19th International Congress of Phonetic Sciences*, Melbourne, Australia, 5 August 2019 - 9 August 2019. Australasian Speech Science and Technology Association Inc, 3075-3079.

Zhang, Y., He, L., Kerdpol, K. & Dellwo, V. (2021). Between-speaker variability in intensity slopes: The case of Thai. Abstract presented at *the XVIIth AISV Conference* (Zürich, 4–5 February 2021).