Issues of Bilingualism in Likelihood Ratio-based Forensic Voice Comparison

Jing Hoi Lo

Doctor of Philosophy

University of York

Language and Linguistic Science

September 2021

Abstract

Situated at the intersection of forensic speech science and bilingualism, this thesis focuses on the issues of language and language mismatch in forensic voice comparison (FVC) and examines their effects on features commonly used in FVC within the framework of likelihood ratios (LRs). To this end, two experiments are presented which explore (1) the performance of the alveolar fricative /s/, long-term formant distributions (LTFDs) and automatic speaker recognition (ASR) software as speaker discriminants in same-language comparisons in Canadian English and French, and (2) the performance of the features above in cross-language comparisons, following a cross-linguistic acoustic analysis of the linguistic-phonetic features.

Although /s/ showed stronger language-independence acoustically than LTFDs, results from Experiment 1 show that /s/ performed more strongly as a speaker discriminant in French than in English, whereas the performance of LTFDs and ASR in the two languages was similar. Results from Experiment 2 show poorer performance across all features to varying extents in cross-language comparisons, which was exacerbated when appropriate reference data matching the language conditions of the case were not used. Individual-level analysis further reveals a complex mapping between acoustic and individual performance in cross-language comparisons. In particular, speakers for whom LTFDs provided the strongest discriminatory performance did not necessarily show the lowest within-speaker variation.

Overall, findings from the current study contribute to our understanding of cross-language comparisons, and more generally to the area of forensic speech science, by demonstrating quantitatively the impact of language mismatch on the discriminatory potential of different linguistic-phonetic and acoustic features within the numerical LR framework, as well as the significance of case-appropriate reference data in such cases. They also demonstrate the diagnostic value of individual-level analysis in system testing and indicate the need for a more nuanced conception of within- and between-speaker variability for FVC.

Contents

List of Tables

List of Figures

Acknowledgements

Many say that pursuing a PhD is a solitary journey. Doubtless even more so when you spend almost half of it in a global pandemic, when the idea of distance itself is at once desired and dreaded, disquietingly dominating and dreadfully demoralising. For me, it is and it isn't. As Eleanor Catton writes in *The Luminaries*, "Solitude is a condition best enjoyed in company." I have had the tremendous luck of enjoying the past four years in some of the best company I could hope for. The journey has been no less arduous, but I am thankful that it has certainly been made all the more worthwhile.

I do not even know where to begin with to thank my supervisors *Paul Foulkes* and *Vincent Hughes*. Thank you for your total and unwavering belief from the outset in this language-loving law graduate who dared come knocking on the door to forensic speech science and the wider world of phonetics. Thank you for letting me be a recklessly independent supervisee, taking all the freedom that I think I needed but cannot believe I deserved. Thank you for your mentorship and support, in words, in knowledge, in action, and in kindness, that have guided and inspired me at every step.

I would also like to thank *Tamar Keren-Portnoy*, for your insights on the Thesis Advisory Panel and for always being such a calming and grounded presence. Thanks to the staff of the Department of Language and Linguistic Science, for providing a nurturing environment that allowed me to grow as a scholar and as a human being. I wish to thank *Geoff Krause* in particular, for making every pesky problem go away so effortlessly, for bringing so much joy to the Department during your time here, and for just being such a jolly good fellow. Thanks also go to the *Forensic Speech Science Research Group*, for thought-provoking discussions about all things forensic; to *Colleen Kavanagh* and the RCMP Audio and Video Analysis Unit, for use of the *Voice ID Database*, which made this thesis possible; and to the *University of York Library*, for your ever helpful staff and especially for your truly stellar work during lockdown.

In a wondrously full-circle moment, this project owes its birth and its final form, perhaps in a circuitous but no less real way, to the city of Ōtautahi Christchurch, where I spent some of the most memorable months of my life during my final undergraduate year. I have *Chris Gallavin* to thank in particular, for introducing me in LAWS307 to this tiny nugget amidst a sea of evidence and case law that I would come to know as forensic phonetics, unwittingly but definitively setting me on the path to York that led all the way here. I also must thank *Sidney Gig-Jan Wong*, for making my first co-authoring experience so rich and illuminating with your supreme contributions and organisational skills, and for your stimulating conversations from the other end of the world, which helped me elucidate my ideas that eventually took their current shape on these shapes.

Doing a PhD over the last four years gave me a whole new world to explore aside from this research project, to try out possibilities and forge my own path. Every scary thing off the path, that I hadn't thought to explore, and that always turned out to be beautiful and exciting, made this journey all the more fruitful. Thank you to *all my students* at York, Huddersfield and the Centre for Lifelong Learning, for helping me learn and grow as a teacher, communicator and linguist. My gratitude goes to *Claire Childs* and *Eleanor Chodroff*, for taking me on to teach on your modules and letting me learn so much in the process, and especially for your support through the (tumultuous) period of pandemic teaching; to *Erica Gold*, for giving me the wonderful opportunity to teach at the University of Huddersfield, turning my final year completely upside down (for the better!); to *Sam Hellmuth*, for taking me on board to work on the York English Learning Toolkit and the MOOC; to the *Careers and Placements* team, especially *Nancy Baines*, and the *Centre for Lifelong Learning* at York, for letting me experience and engage with the wider University community beyond the confines of a department. My gratitude also goes to *Aisha Din*, for, in what I still think is a very brave decision, hiring me and nurturing me to work outside academia, after I graduated from my MSc and through the first year of my PhD, thereby providing me with the essential financial means to carry on, as well as for remaining a good friend.

With all the personal ups and downs, and societal upheavals and downturns, in the past four years, this work would not have been possible without the support of the fellow PGR researchers and GTAs I have had the fortune to work with, at the Department and far beyond, who have led the way with your own footsteps, who have walked this otherwise lonely road alongside me, and who have striven to uphold one another and worked tirelessly for the betterment of our working conditions as GTAs. I thank you all from the bottom of my heart, but I would be remiss not to express my particular thanks to:

> *Marina Cantarutti*, for so readily showing your care for fellow PGR researchers, and for selflessly imparting your acquired wisdom and experience, as well as your unquenchable, infectious passion for phonetics;

> *Sarah Kelly*, for your company in V/B/115 when we both worked there, for listening to my incessant rants and complaints, and for your incisive and ever spot-on advice whenever I asked;
>
> *Grace Cao* and *Bruce Wang*, for every conversation about Cantonese and forensics, and for being there to share our mutual anger and grief over a place that we all cared so deeply;
>
> *Rui Qi Choo*, for regaling me with innumerable Asian chitchat, for sharing my virtual office during the pandemic, and for holding me accountable to my writing and keeping me (as) on track (as possible);
>
> *Matthew Barlow*, for being the commute buddy I never knew I needed or wanted and making those interminable train rides more pleasant than unbearable;
>
> *Julien Plante-Hébert*, for your candid and honest friendship from the moment we met in Istanbul; and
>
> *Heather Turner* and *Sarah Lapacz*, for being my Jedi Council and sticking together, through the seasons of light and the seasons of darkness, through the best of times and the worst of times.

Last but not least, thank you, *Felix*, for keeping me sane, happy and loved, and for reminding me to live a life outside work. I can say with absolute certainty that I could not have done this without you by my side. I have no idea what I did to deserve you.

Author's Declaration

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

Parts of Section 2.2 have been submitted for publication, where I was the lead author and was responsible for the relevant arguments:

Lo, J. J. H., & Wong, S. G.-J. (submitted). Multilingualism and code-switching. In F. Nolan, K. McDougall & T. Hudson (Eds.), *The Oxford handbook of forensic phonetics.* Oxford University Press.

Parts of the arguments in Section 6.1.1 and results from Chapter 8 have been presented and accepted for publication:

Lo, J. J. H. (accepted). Seeing the trees in the forest: Diagnosing individual performance in likelihood ratio based forensic voice comparison. *Studi AISV.*

Parts of Chapter 8 have been published:

Lo, J. J. H. (2021). Cross-linguistic speaker individuality of long-term formant distributions: Phonetic and forensic perspectives. *Proceedings of Interspeech 2021*, 416–420. https://doi.org/10.21437/Interspeech.2021-1699

Chapter 1

Introduction

In 1992, a man called a merchant in Toronto, Canada and left a threatening message on his answering machine. The caller, apparently not wishing to be identified, appeared to have disguised his voice in parts of the call, lowering his pitch and using a breathy voice quality. What sounded clear though was that the caller spoke with some Cantonese accent throughout the English message. The recipient, who was from Hong Kong and spoke Cantonese as his first language, believed he knew who the caller was: Lo, a former employee who was also Hong Kong-born and Cantonese-speaking. Lo was arrested and charged, but maintained he was innocent. Following an analysis of the realisations of English coda /ɹ/ and /p t k/, among other phonetic variables, the defence expert came to the opinion that the accent of the voice in the call was not as strong as Lo's own Cantonese accent. The charges against him were eventually dropped.

An ocean and 14 years away, in 2006, a shipload of cocaine went missing in Ghana after being unloaded off the coast near its capital Accra, sparking a major drug scandal in the country that remained news headlines for years to come (Daily Graphic, 2014; Modern Ghana, 2010, 2011). At the centre of the saga were five men, including a high-ranking police official, who allegedly came together to discuss locating the whereabouts of the missing narcotics (Modern Ghana, 2007a). A highly incriminating recording of the meeting later emerged, and while it was not exactly clear who made the recording, it proved to be controversial and invoked much dispute, not so much of who was present, but of who said what. The recording, which contained a conversation that frequently switched between Ghanaian English, Twi and Hausa, was subjected to extensive forensic analysis and featured prominently in the outcome of the legal proceedings (Modern Ghana, 2007b), which led to two of the five men put to jail, only to be acquitted on appeal (Modern Ghana, 2009).

The cases described above are two examples of bilingualism finding its way into the spotlight in the realm of forensic phonetics. The first case, reported in Rogers (1998), involves the comparison of voices in a speaker's second language and highlights the issue of accent features due to cross-linguistic phonetic influences, specifically in the context of members of a minority community (Cantonese) speaking the majority language of the broader community (Toronto English). The second case, reported in detail in Foulkes et al. (2019), showcases the frequent occurrence of code-switching among multilingual speakers, complicating the whole analysis for the British–Ghanaian team, from transcription to speaker profiling to attribution.

In the UK, forensic materials containing speakers who display non-native features are estimated to amount to 5–10% of casework (C. Kirchhübel, personal communication, 2016). In other parts of the world where bilingualism is the norm, such as India, cases of FVC involving language mismatch in the materials are a much more regular occurrence (N. Suthar, personal communication, 2020). For some practitioners, speech samples involving more than one language could constitute as much as the majority of cases (Künzel, 2013).

That issues related to bilingualism regularly come up in forensic casework should not be surprising. After all, bilingualism is ubiquitous in most parts of the world (Tucker, 1998). Approximately 41% of English speakers worldwide are estimated to speak at least one other language (Crystal, 2003); the proportion of Spanish-speaking bilinguals is conservatively estimated to be at least 30% (Potowski, 2016). In a more recent survey of young people in the European Union, 80% of respondents reported being able to read and write in at least two languages (European Commission, 2018). Globally, many subnational communities speak one or more geographically localised languages alongside a *lingua franca* that is in wider use (e.g., Emlen, 2017; Fitzmaurice, 2019; Garibova, 2017; Horner & Weber, 2008).

The broad aim of the current study is to consider the theoretical and practical issues associated with bilingualism in the practice of forensic voice comparison (FVC). In doing so, this study takes a broad view of *bilingualism* to refer to the use of two or more languages (thus subsuming *multilingualism*) in everyday life (Grosjean, 2012), without making assumptions about proficiency or fluency of the individual speakers. Therefore, bilinguals do not simply refer to individuals with a balanced command of both languages, who are exceedingly rare (if existent at all), but can include those who learn both languages simultaneously from the onset of acquisition (simultaneous bilinguals) or one after the other (sequential bilinguals), those who acquire their languages in different settings (classroom instruction, home, immersion, etc.), and those who have learnt their second language to various levels of proficiency (see Edwards, 2006, 2012, for reviews of the complexity of defining bilingualism and multilingualism). The primary reason for adopting a broad conception of bilingualism is reality. As the cases above show, bilinguals of any background can be implicated in forensic speech analysis. Adopting a narrow view of what constitutes bilingualism thus risks limiting the consideration of issues of bilingualism that may be highly relevant for FVC.

1.1 Thesis structure

The remainder of the thesis is structured as follows.

Chapter 2 describes the task of FVC and summarises current issues of the area. It then discusses issues arising from bilingualism in FVC with reference to the literature on bilingual speech production. It concludes by setting out the specific research aims and questions (RQs) of the current study:

1. What is the effect of language on a feature's discriminatory potential? In other words, how language-specific is the speaker-specificity of any particular feature?
2. What is the effect of language mismatch between the questioned sample and the known sample on the discriminatory potential of a feature?
   - 2.1 In the context of cross-language comparisons, what is the effect of mismatch between the conditions of the case and the conditions of the reference data on system performance?

Chapter 3 considers in greater detail the two linguistic-phonetic variables that form the focus of the current study, the sibilant fricative /s/ and long-term formant distributions (LTFDs). In each case, their use in FVC, specifically in cross-language comparisons, is motivated and previous findings on their discriminatory potentials are reviewed. Each section concludes with specific RQs and hypotheses for any cross-linguistic differences in the production of /s/ and LTFDs in the investigated population (Canadian English–French bilinguals).

Chapters 4 and 5 form the acoustic-phonetic part of the study to address the RQs set out in Chapter 3. Chapter 4 begins by giving a description of the *Voice ID Database* (RCMP, 2016), which is the source of materials used throughout the current study. It then goes on to outline the procedure for preparing and extracting data from the recordings for /s/ and LTFDs, such that a cross-linguistic comparison of their realisations can be conducted. It concludes by presenting results from preliminary testing done on LTFDs to locate the most suitable setting to use in formant extraction for the present set of speakers, which informs the final methodological choices taken in the previous section. Results from the acoustic analysis are presented in Chapter 5. In the case of /s/, midpoint spectral moments and spectral dynamics are both analysed to investigate whether Canadian English–French bilinguals make acoustic distinctions between the two languages in their production. LTFDs are similarly compared between English and French to evaluate the effect of language on them. These results are discussed in the context of previous literature. The chapter concludes by considering the implications of these results for their use in FVC and outlining some specific predictions for the main RQs of this study.

Chapters 6 to 9 make up the forensic part of the study to report on two experiments designed to address the three RQs set out in Chapter 2, namely to assess the effect of (1) language, (2) language mismatch and (3) mismatch in the conditions of the reference data on the discriminatory performance of each variable. Chapter 6 outlines the methodology for system testing and evaluation within the framework of likelihood ratios, followed by the design of each experiment.

Experiment 1 seeks to address RQ1 and evaluate the effect of language on discriminatory performance. System testing is conducted in the context of same-language comparisons, separately in English and in French. A series of systems are tested for /s/, varying the combination of spectral moments and modes of input. System testing for LTFDs similarly encompasses different combinations of LTFs and modes of input. Samples in English and French are fed to the software *Phonexia Voice Inspector* to test the performance of automatic speaker recognition (ASR) software in different languages.

Experiment 2 is set up in the context of cross-language comparisons and is designed to address RQ2. To address the main RQ2 and assess the impact of language mismatch on the discriminatory performance of the same variables, two primary conditions where the language of the questioned sample does not match the language of the known sample are set up and compared with same-language comparisons (from Experiment 1). In one condition, the language of the reference data is matched with that of the known sample, while in the other, the language of the reference data is matched with that of the questioned sample instead. To address RQ2.1 and assess the effect of mismatch between the conditions of the case circumstances and the conditions of the training data, two secondary conditions, where calibration is performed using training scores from same-language comparisons instead of cross-language comparisons, are set up and compared with the primary conditions. Due to restrictions in *Phonexia Voice Inspector*, the set-up of Experiment 2 is adjusted to test the same questions for ASR.

Results from system testing are organised into three chapters, each presenting findings from both Experiments 1 and 2 for one of the examined variables. Chapters 7 and 8 present the results for /s/ and LTFDs respectively, while Chapter 9 presents the results for ASR testing. The three chapters follow the same structure. For each experiment, results from system performance on the global level are first presented. Results from individual-level analysis follow in each case. In the case of /s/ and LTFDs, the individual-level analysis is followed up by an acoustic analysis that explores the relationship between individual performance and the underlying data.

Chapter 10 returns to the main RQs of the thesis and summarises the results from the preceding three chapters. The implications of the findings in this study for forensic casework, along with some possible directions for future research, are discussed, before the thesis concludes.

Chapter 2

Forensic Voice Comparison and Bilingualism

This chapter first gives an overview of forensic voice comparison (FVC) in Section 2.1, highlighting in particular the different approaches taken by analysts to the task and issues surrounding the adoption of likelihood ratios as the framework of analysis. The main part of the chapter, Section 2.2, considers the relevance of bilingualism in FVC and the issues and challenges that considerations of bilingualism pose to this area of work and research. After previous related work in FVC is reviewed in Section 2.3, the chapter concludes by identifying the main research aims and laying out specific research question in Section 2.4.

2.1 Forensic voice comparison

As the main type of casework in forensic speech science in the UK (Foulkes & French, 2012), a typical case of FVC involves the expert being tasked with comparing an audio sample in which the identity of the speaker is disputed (the questioned sample, henceforth QS) with another sample obtained from a known speaker (the known sample, henceforth KS). Around the world, analysts employ a range of methods to perform comparisons, with a linguistic-phonetic approach combining auditory and acoustic analysis being the most common, especially in European jurisdictions (Gold & French, 2011; Morrison, Sahito, et al., 2016). While the adoption of automatic speaker recognition (ASR) technology in practice is on the rise, the use of automatic systems is typically accompanied by some form of human analysis (Gold & French, 2019). An analyst's methodological choice may also be influenced by the state of the law in the jurisdiction they work in. In England and Wales, for example, the eschewal of acoustic methods in favour of auditory analysis alone is not itself a sufficient ground for rejecting an expert's evidence (*R v Flynn & St John* [2008] EWCA Crim 970; cf. the position in Northern Ireland, per *R v O'Doherty* (2002) NICA 20). At the same time, the court in England and Wales has yet to espouse evidence based on ASR in the only case where such kind of evidence has been attempted (*R v Slade & Others* [2015] EWCA Crim 71).

2.1.1 Linguistic-phonetic approach

In a linguistic-phonetic approach, analysts examine a wide array of features from the component units of speech, including not only segmental and suprasegmental (phonetic) features, but also morphosyntactic, lexical, discourse-pragmatic and interactional features where relevant (Foulkes & French, 2012; French et al., 2010). In order to arrive at a conclusion on the evidence, the samples compared are analysed for the degree of similarity of each feature, contextualised by the typicality of the values obtained from those features within the speech community. As outlined by Nolan (1983), for a linguistic-phonetic feature to be effective in FVC, it should exhibit low within-speaker variability and high between-speaker variability. It should also occur frequently in spontaneous speech and be relatively easy to extract and measure, such that sufficient data can be extracted to obtain a reliable representation of the speaker's production. Further, it should be resistant to efforts of voice disguise and robust in transmission, such that any speaker-specific information contained is not lost when the audio signal is degraded and of poor quality in forensic conditions, such as telephone transmission and the presence of substantial background noise. The assessment of typicality in particular requires analysts to have an understanding of the prevalence and distribution of any chosen features within the relevant speech community. To make relevant assessments, analysts have traditionally drawn from sociophonetic research that has documented the patterns of production of those features, or relied on their own estimates where such descriptions are unavailable (Hughes & Wormald, 2020).

As forensic speech science emerges as an independent area of research, the efficacy of particular phonetic features for use in FVC remains one of the main ongoing themes of investigation. On the segmental level, studies exploring the discriminatory potential of consonantal variables have focused on the use of acoustic and temporal parameters of fricatives (Cicres, 2011; Kavanagh, 2012; Smorenburg & Heeren, 2020), nasals (de Boer & Heeren, 2021; Kavanagh, 2012, 2013; Smorenburg & Heeren, 2021; Yim & Rose, 2012), liquids (Kavanagh, 2012) and plosives (Earnshaw, 2014), though other studies have also sought to probe the distributional properties of socially conditioned variants for speaker-specific information (Earnshaw, 2020; Gavaldà, 2016). Among these consonants, nasals and fricatives are generally found to be more effective features for FVC (e.g., Kavanagh, 2012). As for vowels, while midpoint formant frequencies have been shown to carry speaker-specific information (e.g., G. de Jong et al., 2007; Jessen, 1997; Kinoshita, 2001; Loakes, 2004), dynamic representations of vowel formant trajectories have more recently gained much attention for their ability to capture another layer of speaker individuality, particularly in diphthongs, as speakers may display idiosyncratic behaviour not only in the phonetic targets but also in the timing of formant movement throughout the vowel (McDougall, 2006). Speaker-specificity of formant dynamics have since been extensively investigated for diphthongs in a variety of languages, including Australian English (McDougall, 2004; Morrison, 2009b), Thai (Pingjai et al., 2013), Mandarin Chinese (C. Zhang & Enzinger, 2013) and Cantonese (J. Li & Rose, 2012; Pang & Rose, 2012). Nevertheless, not all vowels can benefit from the amount of speaker-specific information brought by formant dynamics on top of static representations, and the evidence for the case of monophthongs, which do not have multiple phonetic targets, remains mixed. Although polynomial fits of formants from monophthongs produced in read connected speech can provide improved speaker-discriminatory performance over formant means (Fejlová et al., 2013), there is some evidence to suggest that, as the phonetic context becomes more uncontrolled and the range of within-speaker variation expands, formant dynamics no longer hold an advantage over midpoint formant frequencies in speaker-specificity (Heeren, 2020). Beyond consonants and vowels, in the case of tonal languages, such as Vietnamese and Cantonese, the realisation of lexical tones can form an additional useful dimension of speaker individuality (Carne & Ishihara, 2019; R. K. W. Chan, 2016).

Segmental properties are often exploited in discourse-pragmatic and interactional features, most commonly in filled pauses (FPs). Often considered a type of disfluency phenomena, FPs are said to derive their discriminatory power in part from their unconscious production, which also contributes to their resistance to voice disguise (Hughes et al., 2016). Speaker-specificity in the realisation of the vowel in FPs (*uh*, *um*), in combination with the vowel (and nasal) duration, is found to compare favourably to lexical vowels (Hughes et al., 2016). The investigation of speaker-specific information in formant dynamics has also extended to interactional features such as *yeah* and *like* (Gibb-Reid, 2018). More broadly, the rates of use of different categories of disfluency phenomena can also be strong carriers of speaker-specific information, with individual speakers maintaining largely consistent disfluency profiles across speech styles (McDougall & Duckworth, 2018). Non-phonological clicking behaviour, on the other hand, has not found support as a reliable speaker discriminant (Gold et al., 2013a).

On a suprasegmental level, fundamental frequency (f_0), rhythm and temporal measures (e.g., articulation rate) are among the most commonly attested features included in the analyst's toolkit (Gold & French, 2011). Findings from existing studies on the discriminatory value of some of these variables, however, have been decidedly unfavourable. f_0, for example, is known to exhibit high within-speaker variability (Braun, 1995) and, on its own, has not been found to be a strong speaker discriminant (Kinoshita, 2005). Similarly, articulation rate has been shown to have limited discriminatory potential (Gold, 2018). Other suprasegmental variables, such as long-term formant distributions (LTFDs; to be discussed below), fare better and have been found to provide stronger speaker-specificity. Voice quality and vocal setting, typically considered strong markers of speaker identity, are commonly analysed auditorily using a modified Vocal Profile Analysis scheme (Laver, 1980; San Segundo et al., 2019). In a recent likelihood ratio-based investigation of acoustic measures of voice quality, Hughes et al. (2019) found promising levels of discriminatory performance from f_0 together with spectral tilt and additive noise measures, but there have been otherwise relatively few studies on the discriminatory potential of voice source parameters (cf. Jessen, 1997; San Segundo & Gómez-Vilda, 2014).

Importantly, while studies have often focused on examining the discriminatory potential of individual phonetic features, the value of considering such features in isolation in FVC is necessarily limited, due to within-speaker variation and overlap of different speakers' ranges of variation. The significance of considering multiple features in concert is thus widely recognised in practice, even though some features may lend more weight to the analysis than others (Gold & French, 2011). Each feature constitutes a dimension along which speakers may vary and be discriminated in a multidimensional "speaker space" (Nolan, 1997), and a higher number of dimensions considered increases the possibility that each speaker occupies a distinct region in this space (P. Rose, 2002).

2.1.2 Automatic approach

In contrast to the linguistic-phonetic approach, the automatic approach to voice comparison does not depend on the identification and examination of individual components of speech. Instead, this approach relies on processing whole audio samples to obtain short-term acoustic features from overlapping windows of short duration (typically about 25 ms) throughout the samples (Hansen & Hasan, 2015). One of the most popular features used are Mel-frequency cepstral coefficients (MFCCs), which are obtained by applying a Mel filterbank, designed to emulate the higher sensitivity to lower frequencies in human listening to a log-transformed power spectrum of the short windowed signals, then performing a discrete cosine transform (DCT) and retaining a number of highly decorrelated coefficients of the resultant cepstrum (Hansen & Hasan, 2015; Jurafsky & Martin, 2009). Through this string of operation, MFCCs are said to be able to decouple the glottal source corresponding to f_0 from the signal (Jurafsky & Martin, 2009), and as such provide a good representation of the supralaryngeal vocal tract. Recent research, however, suggests that the source is not in fact fully decoupled from the filter in MFCCs, and as such they are predictive of not only formants but also f_0 (Hughes et al., 2020).

MFCCs have enjoyed much success in ASR, and as modelling techniques continue to develop (Dehak et al., 2011; Reynolds et al., 2000; Snyder et al., 2018), ASR performance has continued to improve (see, e.g., Morrison & Enzinger, 2019, and papers in the same special issue). However, as the state of the technology stands, cepstrum-based features do not display any degree of interpretability that is comparable to linguistic-phonetic features (P. Rose, 2003), despite recent attempts to interpret ASR output using acoustic features such as formants and f_0 (Hautamäki & Kinnunen, 2020).

2.1.3 Conclusion frameworks

Another area which has generated much discussion in FVC is the framework used by analysts to express conclusions, which in the UK saw a fundamental overhaul in recent decades. Following the ruling in *R v Doheny & Adams* [1996] EWCA Crim 728 by the Court of Appeal in England and Wales, the use of impressionistic likelihood scales to express expert opinion by forensic speech scientists is considered to be inappropriate and has been replaced in the UK by a two-stage framework outlined in a Position Statement (UKPS; French & Harrison, 2007), where samples are first analysed for their consistency and then, upon an initial finding of being "consistent with having been spoken by the same person", assessed for the distinctiveness of the analysed features. While the UKPS codifies the importance of considering the typicality of the features observed, it has been criticised for failing to allow the analyst to express the degree of similarity between the samples (P. Rose & Morrison, 2009). The UKPS remained one of the dominant approaches adopted by practitioners in the UK until the early 2010s (Gold & French, 2011), but by the second half of the decade the field has seen a shift towards adopting the framework of likelihood ratios (LRs; French, 2017). Hailed as the logically correct framework to evaluate forensic evidence (Morrison, 2009a), LRs provide a gradient measure of the evidence and offer a framework for empirically testing the performance (validity and reliability) of the set of features chosen, as well as the specific methods adopted and reference databases used, collecting known as *systems* (Morrison, 2013, 2014). From a regulatory perspective, the logic of the LR framework and its use in evaluative reporting have been endorsed by the European Network of Forensic Science Institutes in its Guidelines

(ENFSI, 2016) and by the UK Forensic Science Regulator in the current codes of practice and conduct (Forensic Science Regulator, 2021). For these reasons, it is this framework of LRs that will be adopted throughout the current study.

At its core, the framework of LRs conceptualises judicial decision making as Bayesian reasoning, where each piece of evidence acts to inform the trier-of-fact's degree of belief in the prosecutor's and the defence's propositions (see Champod & Meuwly, 2000). Each piece of evidence is evaluated against these two competing propositions to assess (1) $p(E|H_P)$, the probability of observing the evidence under the prosecution's hypothesis (i.e., that the speakers are the same, in the context of FVC), versus (2) $p(E|H_D)$, the probability of observing the evidence under the defence's hypothesis (i.e., that the speaker in the QS is not the known speaker but some other speaker in the relevant population). The outcome of the comparisons is given in the form of an LR, which is the odds of the two probabilities: $\frac{p(E|H_P)}{p(E|H_D)}$. An LR of 1, meaning that the two probabilities are equal, offers no support for either side. An LR greater than 1 indicates greater probability of observing the evidence given the prosecution's hypothesis than given the defence's hypothesis and thus offers support for the prosecution. The greater the magnitude of the LR, the stronger the degree of support. On the contrary, an LR between 0 and 1 indicates support for the defence's proposition, and the closer the LR is to 0, the stronger the degree of support there is in this direction. The analyst may then express their conclusions in the form of a numerical LR or convert them to a verbal likelihood ratio (see, e.g., Mullen et al., 2014).

In a fully Bayesian paradigm, as the trier-of-fact incorporates each piece of evidence, their prior beliefs are updated to reflect the strength of the evidence, resulting in a posterior probability that reflects their state of belief after all the evidence has been incorporated into consideration. In practice, the implementation of LRs is far less straightforward and not without issues. Not all forms of evidence are amenable to numerical treatment for LRs to be derived. LRs, whether numerically or verbally expressed, have been shown to be prone to misinterpretation by jurors, legal professionals and other users of forensic evidence alike, their treatment biased by the type of evidence to which they are attached (Martire et al., 2014; Mullen et al., 2014; Thompson & Newman, 2015). While issues of understanding LRs in court are far from resolved and continue to be debated, the use of LRs in evaluating forensic evidence crucially underscores the role of the forensic expert, which is to give an opinion on the strength of the evidence on the basis of expert knowledge, rather than comment on the posterior probability of whether the sources of the samples are one and the same, a question that should be left to the trier-of-fact. In other words, the remit of the forensic expert is confined to the specific piece of evidence.

The adoption of LRs as a conceptual framework further reaffirms and makes explicit the significance of evaluating the evidence with respect to multiple rather than single hypotheses.

When the conclusion for some evidence is given in an LR, its numerator and denominator corresponds respectively to the similarity between the QS and the KS and the typicality of the features observed in the samples. An appropriate definition of the relevant population is thus crucial when estimating how typical or distinctive the observed values of the features are within that population. While in some cases the defence proposition may be narrowly delineated (e.g. in a closed-set comparison), often it is broadly constructed to simply exclude the suspect from consideration, which may be unhelpful for the analyst. Given that it is extremely rare that the parties to the proceedings will come to an agreement on a specific definition before forensic analysts have to undertake their work (Hughes & Rhodes, 2018), analysts are left to exercise their judgment to delineate the relevant population pragmatically, on the basis of logically relevant factors displayed in the QS. Crucially, only characteristics of the QS, and not characteristics of the KS, should have a role in defining the relevant population (Robertson & Vignaux, 1995). A key issue in doing so, however, is the conceptual paradox that, precisely because the identity of the questioned speaker is unknown, the "correct" definition is inherently indeterminate (Hughes & Foulkes, 2015a). A narrow definition of the relevant population may increase the strength of evidence over a broader alternative, but only if the circumscription turns out to match the actual characteristics of the questioned speaker; if the analyst's decision leads to a mismatched population, the validity of the conclusions reached will be adversely affected (Hughes & Foulkes, 2015a, 2015b). Decisions regarding the relevant population based on assumptions of group-level characteristics of the questioned speaker thus directly impact the validity of the outcome.

It has been argued that such assumptions to narrow down the relevant population can be applied without forming part of the expert's conclusion, as long as the assumptions are made clear to the court, as the relevant characteristics from the questioned speaker will usually be sufficiently salient to the trier-of-fact that no forensic expertise needs to be accessed (Morrison, Enzinger, & Zhang, 2016, 2017). However, the nature of voice evidence, as Hughes and Rhodes (2018) argue, prevents that from being the case, and untrained listeners are ill equipped to make a determination of such information and its evidential value. Instead, Hughes and Rhodes (2018) advocate a two-level framework, in which voice evidence is conceptualised as comprising group-level (Level 1) characteristics and individual-level (Level 2) characteristics. The analyst may arrive at a conclusion to each proposition defined in the two levels, although how the evidence is presented in the end may depend on the particular circumstances of the case. The evidential value of group-level observations can be assessed at Level 1, which in essence asks the rarity of the accent in the QS. More idiosyncratic properties of the voice can then be assessed at Level 2, using a relevant population refined on the basis of Level 1 observations.

An altogether different approach to the issue of the relevant population is proposed in Morrison, Ochoa, et al. (2012), who recommend the use of a panel of lay listeners to select the composition of the relevant population based on sufficient perceptual similarity with the questioned speaker. They argue that the default alternative proposition that the analyst should adopt is that: "The suspect is not the speaker on the offender recording, but is someone who sounds sufficiently similar to the voice on the offender recording that a police officer (or other appropriate individual) would submit the offender and suspect recordings for forensic comparison" (p. 64). A number of issues with this proposal have been identified in Gold and Hughes (2014), who point out that listeners can be highly variable in the acoustic elements they are sensitive to when assessing voice similarity, and as such this approach could be unreplicable and remain opaque to the trier-of-fact. Further, the process of selecting recordings to be presented to listeners itself requires screening from the expert based on some relevant, broad criteria and is thus not unsusceptible to the expert's influence.

After a relevant population has been defined, relevant features are then selected and subjected to the expert's chosen method of analysis. Any similarities and differences are evaluated against the typicality of the features in the relevant population, which may be assessed with reference to population statistics where available (see Gold & French, 2019), such as for f_0 (Hudson et al., 2007; Jessen, 2008; Künzel, 1995; Skarnitzl & Vaňková, 2017) and articulation rate (Gold, 2018; Jessen, 2008). In a numerical, data-driven LR approach, feature values are statistically modelled and compared to compute output scores in the form of LRs. In voice evidence, continuous data such as formant frequencies and consonantal spectral measures are commonly modelled and compared using the multivariate kernel density or Gaussian mixture model–universal background model approach (see Chapter 6 for details). Much voice evidence, however, comes in the form of discrete data (e.g., occurrence of paralinguistic clicks) and may require the development of specialised statistical models so that they can be integrated in a numerical LR approach (e.g., C. Aitken & Gold, 2013; Bolck & Stamouli, 2017).

To obtain an accurate estimate of the typicality of features, such an approach relies heavily on a suitable database of recordings containing a representative sample of the relevant population to be used for modelling. One proposal is for the expert to "go and get" a case-specific reference sample (P. Rose, 2007), but this exercise may be costly in terms of financial resources and, more importantly, time; in cases where the relevant population is not sufficiently narrowly defined, the difficulty for the expert to collect a representative sample of the population may be prohibitive. There have been efforts by researchers to develop large-scale corpora for forensic purposes, which allow a substantial amount of pre-existing data to be taken off the shelf (e.g., Gold et al., 2018; Jessen et al., 2005; Nolan et al., 2009), but to date only a small number of such databases are available.

Case-appropriate large-scale databases are also important to the next step in the derivation of LRs. Although output scores, whether calculated by human or ASR systems, take the form of LRs and can indicate the relative strength of evidence, their absolute values are not directly interpretable as LRs *per se* in the circumstances of the specific case (Morrison, 2013). Rather, they need to be converted to (interpretable) LRs via calibration using scores calculated from a set of training data. In order to achieve good-quality calibration, the training data should come from the relevant population and match the QS and the KS in their respective conditions (Morrison, 2013), and any mismatch could lead to miscalibration, causing performance to deteriorate and the validity of the resultant LRs to be reduced (Morrison, 2018).

2.2 Issues of bilingualism in FVC

Issues arising from bilingualism in forensic speech science are not a new topic. In the area of language analysis in asylum proceedings, concerns relating to the treatment of speakers who engage in the practice of language mixing, or who have low proficiency in the language of the interview, have been much discussed (Language and National Origin Group, 2004). Bilingualism has also received some attention in earwitness identification, where the ability of bilingual listeners to identify speakers in different languages from a voice line-up has been a main matter of concern (Mok et al., 2015; Sullivan & Schlichting, 2000). However, despite the regular occurrence of cases in FVC that involve bilingual speakers or language mismatch between samples (Künzel, 2013; N. Suthar, personal communication, 2020), related issues in FVC remain underexplored. Theoretical and practical issues related to bilingualism in this type of casework have yet to be systematically considered, and very few studies have sought to address these issues. Meanwhile, the Code of Practice of the International Association for Forensic Phonetics and Acoustics (IAFPA), the professional body for the discipline, explicitly advises its members to exercise caution when carrying out analysis “in a language that the analyst does not have native-level competence” (3.9) or when conducting “cross-language comparisons” (3.10;

IAFPA, 2020).

The beginning of Chapter 1 introduced two main scenarios in which bilingualism is highly relevant to FVC. In one, the QS apparently displays features from a non-native speaker of the language. In the other, speakers in the QS engage in code-switching. Though not illustrated, a third scenario, where the language of the QS differs from that of the KS, has also been attested as not unusual. Of course, the ways in which bilingualism surfaces in forensic recordings listed above are by no means mutually exclusive. Drawing on the wealth of literature on bilingualism and second language acquisition, the remainder of this section is dedicated to outlining a number of key issues associated with bilingualism in these scenarios, before returning to review previous related research in the area of FVC in the next section. The first part of the discussion in Section 2.2.1 centres on considerations related to the choice of features, and the second part (Section 2.2.2 turns to specific issues for FVC within the LR framework. While this section mainly centres on the impact of such considerations on the linguistic-phonetic approach to FVC, undertaken by the majority of practitioners in the field (Gold & French, 2011), some of the theoretical and practical issues are no doubt of relevance to analysts who adopt automatic methods as well.

2.2.1 Bilingualism & choice of features

2.2.1.1 Non-native-sounding features

For speakers who learn another language after acquiring their first language (L1), it is well known that the phonetics and phonology of their L1 can have considerable influence on their second language (L2). Where sounds in the L2 are not present in the L1 inventory, speakers may substitute them with sounds found in their L1 instead. For instance, the English voiceless dental fricative /θ/ (and its voiced counterpart /ð/), which is absent from many of the world's most spoken languages, may be variably substituted with [t] by L1 speakers of Dutch, Canadian French and Thai (Lombardi, 2003; Picard, 2002; Wester et al., 2007), [s] by L1 speakers of Japanese, Mandarin and Polish (Hancin-Bhatt, 1994), or [f] by L1 Afrikaans and Cantonese speakers (A. Y. Chan, 2006; Deterding et al., 2008; Watermeyer, 1996). Transfer of acoustic-phonetic cues across similar segments is another common outcome in both consonants and vowels, perhaps most commonly studied in terms of voice onset time (VOT) of stops (see reviews in Davidson, 2011; Zampini, 2008). As languages employ different VOT implementations of phonological voicing contrast in their stop series (T. Cho & Ladefoged, 1999; Lisker & Abramson, 1964), speakers who go on to learn another language with a different voicing distinction (e.g., English learners of French) may then produce stops in their L2 with VOT in between the monolingual norms in the L1 and the L2 (Flege, 1987; Flege & Eefting, 1987).

The difficulty for learners to acquire certain new sounds in the L2 is more generally predicted in prevailing models of L2 sound learning. The Speech Learning Model (SLM, Flege, 1995), recently revised in Flege and Bohn (2021), posits that speakers maintain the ability to form novel phonetic categories across their lifespan. When they are learning sounds in an L2, these sounds are evaluated against established L1 categories in a common phonetic space, and result in the formation of new sound categories when sounds in L2 are perceived to be sufficiently dissimilar from L1 phonetic categories. The Perceptual Assimilation of Second Language Speech Learning (PAM-L2, Best and Tyler, 2007; see also PAM, Best, 1995), is similarly underpinned by the idea of perceiving and categorising L2 sounds with reference to L1 categories. Depending on the degree of consistency between the L1 and L2 sounds, a sound in the L2 may be assimilated to an L1 category with varying degrees of goodness, or left uncategorised if no consistency is found with any L1 categories. (A sound may also be considered non-assimilable if it is not heard as linguistically relevant.) Notably, the basis for categorisation is not necessarily phonetic or gestural, but can be phonological, such as the case of /r/ for L1 English speakers (who typically have some form of alveolar realisation) learning French (which has uvular [ʁ]). The respective mapping of two distinct L2 sounds with respect to L1 categories then governs the ability of an individual to detect a contrast between them and, depending on the initial mapping, new phonetic or phonological categories may be developed as the learner continues to receive input in the L2.

Segments aside, the prosodic or suprasegmental aspect is thought to be particularly challenging in L2 acquisition, in terms of both phonological categories and phonetic implementations (Mennen, 2007). L2 learners experience difficulties in the acquisition of intonational peak alignment (Graham & Post, 2018), rhythm (A. Li & Post, 2014), intonation (Santiago & Delais-Roussarie, 2015), stress, etc. Evidence for L1 phonetic transfer, for example, can be found in the realisation of word-level stress and phrase-level accent by the intermediate Arabic learners of English in Almbark et al. (2014), whose L2 speech rarely exhibited phonological transfer in stress and accent assignment, but was characterised by underuse of f_0 to mark accent, as well as overuse of f_0 and lack of vowel reduction to mark stress. At the same time, challenges in L2 prosody acquisition are not always straightforwardly attributable to L1 influence, as many learners of different L1s appear to experience similar difficulties, thus leading to the view that there may be a universal pathway for the development of prosody in L2 (A. Li & Post, 2014).

Transfer from L1 to L2 may further be found in the use of disfluency phenomena, which are regularly examined in FVC. While speakers generally produce more disfluencies in L2 speech as they face increased planning difficulties and cognitive load (Fehringer & Fry, 2007), learners across different levels of proficiency largely retain a similar disfluency profile from their L1 (Fehringer & Fry, 2007; Olynyk et al., 1987; Wiese, 1984). Phonetically, where the choice of FP vowels is different in the learner's L1 and L2, L2 learners commonly deviate from the normal vowel produced by L1 speakers of the language, although learners of higher proficiency may approach the norms more closely (R. L. Rose, 2017).

Overall, when a speaker is perceived as non-native, it involves a constellation of segmental and suprasegmental features deemed to deviate from the native norms in their L2. While cues from segmental, suprasegmental and disfleuncy features all contribute to the perception of a non-native accent (Boula de Mareüil & Vieru-Dimulescu, 2006; Magen, 1998; Pellegrino, 2013), the relative contribution of individual cues, or between segmental and suprasegmental features, remains unclear. Findings from some studies indicate a more consistent role of suprasegmentals than segments on accent ratings (Anderson-Hsieh et al., 1992; Magen, 1998), whereas others have found evidence that differences in the production of segments outweigh suprasegmental cues such as intonation and timing in their influence on ratings of accentendness (Ulbrich & Mennen, 2016) and identification of L2 varieties (Vieru-Dimulescu & Boula de Mareüil, 2006).

One potential favourable factor for the speaker-discriminatory potential of non-native-sounding features is the high degree of individual variation within L2 and bilingual varieties. In spite of systematic influence from the L1 sound system, a wide range of factors contribute to a highly varied set of acquisition outcomes. Earlier studies in L2 acquisition have focused on the role of age of acquisition, particularly with reference to a hypothesised "critical period" (Lenneberg, 1967), after which native-like L2 acquisition is considered no longer possible, in addition to factors such as length of residence in the L2-dominant environment, L1 and L2 use, and formal instruction (Piske et al., 2001). More recently, however, it has been hypothesised that the primary driver for the differences between early and late L2 learners is the quantity and quality of input: Speakers who receive more input through interaction with L1 speakers (typically earlier-arrival immigrants) become more likely to acquire L1-like pronunciation (Flege, 2019). Indeed, it is not infrequent that late learners can be judged by native listeners to sound native (e.g., Birdsong, 2007). Even among highly proficient speakers, however, there can often be a full range of variability, as found by Sewell and Chan (2010) in the number of local consonantal features produced by their Hong Kong English speakers. As each bilingual individual, be it heritage speaker, early bilingual or adult learner, comes from a different set of backgrounds, having acquired the language in a variety of settings, achieved different levels of proficiency and experienced a wide array of patterns of language use (de Bruin, 2019), the strength of an L2 accent often cannot be easily correlated with particular subgroups of bilinguals.

Between-speaker variability in L2 varieties can also be socially stratified as speakers develop sociolinguistic competence in their L2 (see Bayley & Regan, 2004). Variationist studies on L2 speech such as Adamson and Regan (1991) have found that immigrant speakers can develop gendered targets for sociolinguistic variables in a similar way L1 speakers do, such as for English (ING), where female L2 speakers, like female L1 speakers, favour the more prestigious variant [iŋ], while male L1 and L2 speakers alike favour the use of [ɪn]. However, L2 speakers often do not acquire the same set of internal (linguistic) and external (social) constraints governing local L1 communities (Howard et al., 2006; Mougeon et al., 2004); instead, they partially replicate some of the same constraints as L1 speakers, while at the same time introducing their own constraints that are irrelevant for local peers (Schleef et al., 2011).

Another important factor to consider is speakers' motivations and attitudes towards varieties of the (second) language, and how they interact with L1 influence. While L2 speakers commonly regard prestigious L1 varieties, such as Received Pronunciation (RP) and General American English in the case of English, as having higher status and/or being more attractive (e.g., Carrie, 2017; Dalton-Puffer et al., 1997), they may not necessarily orient towards such varieties as their targets in production. Among Norwegian learners of English, who rated RP as of higher status and linguistic quality than General American English, Rindal (2010) found an almost even split between learners who aimed for RP and General American English as their model of pronunciation. Yet, although each speaker's target accent served to modulate their production of American versus British variants, Norwegian learners produced a mix of American and British phonetic variants with a preference for the former, regardless of their own target, possibly due to cross-linguistic influence from their L1 (Rindal, 2010). In other places, the local (L2) variety may become nativised, such that speakers turn from exogenous norms to endogenous ones (Schneider, 2003). Hong Kong English is arguably in such a state of flux, with the variety generally agreed to have reached completion of nativisation and beginning to show signs of endonormative stabilisation (Evans, 2009; Schneider, 2007; Sung, 2015), as attitudes towards the variety becoming increasingly positive among its speakers (Cao, 2018; Hansen Edwards, 2015, 2016).

Thus, far from behaving in any way resembling homogeneity, L2 learners who share an L1 form an extraordinarily heterogeneous speech community. Features that give the percept of non-nativeness, however, can also originate from other groups of speakers. Simultaneous bilinguals, often heritage speakers, are judged to speak with a foreign accent in their non-dominant language, particularly for speakers who use the language less in their daily life, though their accent is not judged to be as strong as L2 late learners (Kupisch et al., 2014). Simultaneous bilinguals and heritage speakers may diverge from monolingual norms in production, with effects of cross-linguistic influences demonstrated in the production of stops (Fowler et al., 2008; Sundara et al., 2006), vowels (Guion, 2003), prosody (J.-Y. Kim, 2019; Queen, 2001) and FPs (Lo, 2020). The bidirectional effects of phonetic transfer in bilinguals mean that divergences from native norms can further come from L1 speakers themselves who then go on to acquire another language, giving rise to the perception of a non-native accent (e.g., de Leeuw et al., 2010; Hopp & Schmid, 2013). SLM in particular posits the operation of both mechanisms of category assimilation, by which a merged phonetic category of similar L1 and L2 sounds emerges over time, and category dissimilation, which arises out of the pressure to keep L1 and L2 sound categories distinct, such that the L1 does not stay immutable but remains open to influences from subsequently acquired sounds over the lifespan (Flege, 2007). In the longer term, this form of phonetic plasticity is observed in migrants who left their home countries to live in an L2-dominant place, resulting in gradient phonetic shifts (de Leeuw et al., 2013) or even loss of phonemic contrasts in their L1 (de Leeuw et al., 2018). This sort of fluctuation is not limited to long-term contact with the L2, but has also been observed in contexts of short-term immersion, where a speaker newly embarking on learning an L2 nevertheless shows convergence from the L1 to the L2 in a matter of weeks (Chang, 2012, 2019).

These findings from various groups of bilingual speakers have implications for the selection of linguistic-phonetic features in FVC, as whether a particular feature is useful may depend on the particular speech community in question. However, since non-native-sounding features are not within the exclusive purview of L2 late bilinguals, but permeate through bilinguals from a wide range of backgrounds, it may become more difficult to pinpoint the language background of the speaker when such features are encountered in the QS.

Further, while considerable between-speaker variation is attested within L2 varieties, the speaker-discriminatory potential of a feature in FVC is predicated on not only high between-speaker variability within the variety of which the questioned speaker is assumed to be a member, but also importantly low within-speaker variability. One related claim in the literature is that L2 speech is more variable than L1 speech (Wade et al., 2007). While some evidence has been found in support of this in speaking rate (Baese-

Berk & Morrill, 2015), other work comparing within-category variability of vowels and stops has shown that greater variability within L2 speakers is not necessarily found on a general level, but rather confined to specific features and acoustic cues (Vaughn et al., 2019; Xie & Jaeger, 2020). It must be noted, however, that this body of research has thus far focused solely on read speech, and L2 spontaneous speech may yet vary in different ways.

Apart from the precision of sound categories within specific (controlled) styles, intra-speaker variation crucially arises as a consequence of style-shifting. L2 speakers can respond to shifts in style by adopting more native-like variants (Beebe, 1980) or favouring different variants in the L2 variety (Lin, 2003). They can also accommodate to L1 speakers of different varieties on selected variables (Cao, 2015, 2018). L1 varieties may be assigned different social meanings in an L2 community, such that speakers orient towards different models of pronunciation when style-shifting across different social situations (Rindal, 2010). With a broader repertoire of variation available to them, bilinguals, especially those who are highly proficient or resident in an L2 setting, are capable of manipulating features from their L1 to express social meaning in their L2 (Gafter & Horesh, 2020). The stylistic range of bilingual speakers thus has implications for FVC, if the QS consists of more informal conversation and the KS is in the form of a formal interview.

A final challenge for the use of non-native-sounding features is that, despite the wealth of studies conducted on bilingual speakers in general, many L2 varieties remain underdescribed. Much of L2 research focuses only on a small set of language pairings, predominantly with English as the target language (Gut, 2009). Many studies also focus on segmental aspects of acquisitions, with suprasegmental features relatively underresearched (Rasier & Hiligsmann, 2007). It may thus be more challenging to locate particular features of interest if the variety has not yet been well documented.

2.2.1.2 Code-switching

As the Ghana case (Foulkes et al., 2019) described in Chapter 1 illustrates, a hallmark of spontaneous, conversational speech of bilinguals is the occurrence of code-switching. Within the same conversation, or even within the same clause, bilinguals regularly incorporate lexical items or longer phrases from another language. Once considered symptomatic of linguistic incompetence (Weinrich, 1953), code-switching is now known to be rich in sociolinguistic and pragmatic meaning and constitutes communicative resources that bilingual speakers can utilise (Gumperz, 1977). Speakers may engage in code-switching as audience design, and the occurrence and frequency of the phenomenon may be conditioned by a range of factors governing the speaker's sociolinguistic relations with their interlocutors on the macropolitical, social and interactional levels (Gardner-Chloros, 2009). For forensic purposes, while the switching of codes between conversations may give rise to cases of cross-language comparison, the presence of code-switching within a sample does not necessarily present fundamental issues, since same-language materials are present in both QS and KS. Materials from each language can be extracted for comparison, provided there are sufficient materials from each language for analysis. To be sure, which language is the L1 or the L2, or whether that is apparent from the samples, remains a separate issue.

The separation of code-switched materials into each language should not, however, be taken for granted. Switches can take place word-internally, with the other-language item taking on inflectional marking and word order of the surrounding language (Sankoff et al., 1990), while variably following the phonology of either language, but not both (Bessett, 2017; Stefanich & Cabrelli Amaro, 2018). Indeed, many researchers do not consider such cases to be code-switching altogether, but differentiate them as borrowing instead, citing the irregular treatment of phonetic integration and the general pervasiveness of phonetic influence throughout bilingual speech as evidence that, unlike morphosyntactic criteria such as case marking and conjugation, phonetic criteria are poor diagnostics of the underlying process (Poplack and Meechan, 1998; Poplack et al., 2020; cf. Deuchar, 2020).

Apart from the potential difficulty of identifying code-switched items, studies that investigate the phonetics of code-switching have found fine-grained effects at or around the site of the switch. Largely focused on word-initial plosives, most such studies have found that VOT in a code-switched word typically undergoes minor shifts towards the norm of the other language. In most cases, the effect of transfer is an asymmetric one, having effect on one language (Antoniou et al., 2011; Balukas & Koops, 2015; Olson, 2016b), or effecting shifts of different magnitudes in each language (Bullock & Toribio, 2009; Olson, 2016b). Some have specifically noted an effect of language dominance, whereby switching to the dominant language is particularly prone to transfer effects from the non-dominant language (Olson, 2013), and as such balanced bilinguals are unaffected (Tsui et al., 2019). One account attributes the phonetic effects of code-switching to incomplete or partial inhibition of the other language (Olson, 2013). Studies investigating spontaneous speech data have further raised cognitive load as a further potential influence on code-switching productions (Piccinini & Arvaniti, 2015).

Beyond VOT, fine-grained phonetic effects of code-switching have been found in the realisation of other consonants and vowels, as well as prosody (Khattab, 2013; Muldner et al., 2019; Olson, 2016a). There is further evidence that partial inhibition (or activation) starts to take effect some time prior to the actual switch, resulting in subtle acoustic cues perceptible to listeners in anticipation of the upcoming switch (Fricke et al., 2016). Language-specific phonological processes, such as voicing assimilation and stop spirantisation in Spanish, may be licensed even across language boundaries (Olson, 2019). Thus, while the primary site of interest is where the switch takes place, the effects of code-switching extend further beyond.

This body of research highlights, among other things, that recordings containing code-switching defy simplistic isolation of materials in each language, particularly when the switch involves only single lexical items, which represent the typical objects of analysis in these studies. Treating such materials as independent in separate acoustic analyses could risk confounding the effects of language-specific systems and code-switching behaviour, especially for sounds in the vicinity of the switch site. The finding that increased cognitive load may be a contributing factor in the phonetic effects of code-switching warrants further investigation, for it may have forensic implications when police interviews remain a key source of KS.

2.2.1.3 Availability of features in cross-language comparison

In cases of cross-language comparison, the limited pool of features that can be utilised for comparison poses a considerably greater challenge. For analysts who at least in part rely on auditory-perceptual and acoustic-phonetic methods, segmental analysis is far from straightforward, as phonological inventories and phonotactic rules are both specific to each language. Coarticulatory patterns across segments likewise display language-specificity (e.g., Beddor et al., 2002; Bombien & Hoole, 2013; Mok, 2010; Zsiga, 2000). Commonly, segments which are ostensibly "equivalent" across languages display fine-grained cue-specific acoustic differences (Chodroff et al., 2019; Deterding & Nolan, 2007). Suprasegmental features also undergo systematic shifts and may therefore be incomparable across languages. For bilingual speakers, then, language-specific targets could well preclude reliable, direct comparison of acoustic measurements obtained across languages, even when effects of cross-linguistic phonetic influences described above in Section 2.2.1.1 are taken into account.

Künzel (2013) goes so far as to assert that language-related features, as opposed to voice-related ones, must be excluded from analysis, as the language-specificity of "high-level" features, such as "dialect, sociolect, intonation patterns, phonetic and linguistic parameters of hesitation" (p. 25), will render their analysis all but impossible; automatic approaches, on the other hand, primarily make use of cepstral coefficients and are considered to be impacted only in a quantitative way, as these "low-level" features, which characterise the general filtering behaviour of the vocal tract, involve considerably less influence of the language.

Indeed, language-specificity of long-term measures of voice quality and vocal setting is commonly reflected in the speech of bilinguals when they speak in each language. Bilingual speakers of numerous language pairs have been reported to distinguish overall f_0, in both level and span (Altenberg & Ferrand, 2006; Baird, 2019; Cheng, 2020; Järvinen et al., 2013; Schwab & Goldman, 2016). Similarly, (largyngeal) voice quality and (supralaryngeal) vocal setting, while generally regarded as quasipermanent properties of the voice (Laver, 1980), do not remain unchanged when bilinguals switch languages. Recent studies that target acoustic correlates of laryngeal voice qualities (e.g., spectral tilt for creaky phonation) provide support for the possibility of voice quality transfer from the more dominant language to the weaker language. Evidence can be found in the use of creaky voice in phrase-final position in Spanish by heritage or L2 speakers whose dominant language is American English, in which there is frequent use of the feature especially among young speakers, whereas the feature is only rarely found in monolingual speakers of Mexican Spanish (J. Y. Kim & Soto-Corominas, 2017). Articulatory evidence on the subject is to date scarce, but research on interspeech postures (ISPs, as proxy for articulatory settings) in English and French suggests that speakers who are perceived as native in both languages adopt distinct ISPs when speaking in monolingual mode (Wilson & Gick, 2014). Acoustic studies, which typically focus on examining measures derived from long-term average spectra (LTAS) or LTFDs, similarly provide evidence for language-specific settings used by bilinguals. Proficient L1 Cantonese–L2 English bilinguals, for example, have been found to produce higher mean spectral energy and lower spectral tilt in Cantonese than in English, pointing to a tenser larynx and breathier voice quality in their L1 (Ng et al., 2012). A small-scale comparison of LTAS from Korean–English bilinguals by S. Cho and Munro (2017) found spectra obtained from read speech in Korean were generally associated with lower intensity by 3–5 dB in the frequency range of 2–4 kHz and by 5–10 dB in higher frequency ranges than the corresponding spectra obtained in English. Interestingly, the authors also observed that the LTAS in both languages were much more similar for speakers who were perceived to be clearly more dominant in one language. Further, inspection of LTFDs in the same study revealed overall slightly lower F2 peaks in Korean than in English, reflecting the same tendency previously found for vowels in the two languages.

The same applies to the distributional and phonetic properties of FPs and disfluency phenomena on a broader level. L2 speech generally contains a higher number of disfluencies when compared to L1 speech, as speakers are faced with increased cognitive load and planning difficulties (Fehringer & Fry, 2007). As in other linguistic domains, disfluency profiles of L2 learners across different levels of proficiency have been found to transfer from L1 to L2 (Fehringer & Fry, 2007). Forms of disfluency can also transfer, such that forms typically only found in the dominant language can surface in the speech of the other language (Hlavac, 2011). Phonetically, current acoustic evidence suggests that bilingual speakers maintain subtle distinctions realisations of FPs across languages. In line with earlier studies within each language, L1 Dutch speakers who are highly proficient in L2 English produce FPs with a higher and more backed quality in English than in Dutch (de Boer & Heeren, 2020). In a similar vein, FPs by New Zealand English–Māori bilinguals tend to have higher F1 and F2 in English than in Māori (Wong & Papp, 2018). In the case of simultaneous bilinguals, language-specificity is mediated by the relative dominance of the languages: German-dominant German–French bilinguals, for example, produce FPs in French that are more open, and hence closer to the canonical vowel quality of FPs in German, than their French-dominant counterparts (Lo, 2020).

Although the studies above indicate that direct, quantitative comparison of linguistic-phonetic features is likely challenging in case of language mismatch, such "language-related" features arguably still have great capacity to carry speaker-specific information. Theories of sound learning such as SLM introduced above make clear that, for bilinguals, sounds in the L1 and L2 that are perceptually similar may be assimilated and form "merged" sound categories. These sounds are thus not independent of each other in representation, and their perception and production across languages remain closely linked within the individual. In addition, bilinguals have been found to show similar structure of voice variability across languages, contributed by common sets of formant-based and source-based acoustic variables (Johnson et al., 2020). Indeed, despite cross-linguistic differences, bilinguals are found to produce highly similar LTFDs across languages (see Chapter 3.2.1), and variation in LTFDs is said to be lower within individual bilinguals across languages than between different speakers (S. Cho & Munro, 2017; Heeren et al., 2014), thus raising the possibility that they can preserve some speaker-specificity cross-linguistically. Production of different types of disfluency phenomena in the L2 has also been found to be at least in part dependent on the speakers' behaviour in the L1 (N. H. de Jong et al., 2015). Further, idiosyncrasy may be found in how bilinguals shift between different languages, for segmental variables such as stop VOT (Johnson, 2021a), or in the distribution and realisation of FPs (de Boer & Heeren, 2020; Lo, 2020).

Based on the reasons above, it is argued that the claim that all language-related features must be excluded warrants closer scrutiny. At the same time, it can be questioned to what extent low-level features such as cepstral coefficients are resistant to systematic effects of language. MFCCs in principle encode information of the supralaryngeal vocal tract and not the glottal source, but in fact do show strong correlations with both formant frequencies and f_0 (Hughes et al., 2020). As bilingual speakers have been found to develop language-specific patterns of voice quality and articulatory settings, language would then in theory have an effect on not only higher-order acoustic features aimed at capturing general filtering behaviour, such as LTAS and LTFDs, but also these cepstrum-based features.

2.2.2 Bilingualism & the LR framework

The discussion in the previous section has focused on general issues that bilingualism brings to the analysis of linguistic-phonetic features. This section turns to address some specific challenges associated with bilingualism that arise within the LR framework.

2.2.2.1 The evidential value of group-level characteristics

Within an LR framework, the typicality of such features is to be assessed within the relevant population, as defined by the defence proposition. As mentioned above, however, it is extremely unlikely that the analyst will have a specific defence proposition agreed by both parties to work with. Instead, alternative propositions may have to be formulated by assessing group-level characteristics (e.g., regional background and speaker sex) in the QS to refine the relevant population for the evaluation of individual-level evidence. As outlined above, Hughes and Rhodes (2018) argue that group-level characteristics have much evidential value and any judgments should not be delegated to lay listeners, but should remain within the domain of the forensic expert, citing as example lay listeners' poor ability in identifying regional background. In extension, it is argued here that, where language background is concerned, whether general (monolingual vs. bilingual) or specific (which L2 variety), the determination of group-level characteristics does require expert knowledge and presents critical evidential value. In addressing this issue, this section leaves code-switching within the QS out of the present discussion, as in such cases the relevant population must logically be bilinguals in those languages, unless there is reason to doubt that the same individual produced those utterances in the different languages.

In general, perceptual studies have found that both native and non-native listeners are relatively good at identifying whether a speaker is native or not (e.g., Bent et al., 2016; Major, 2007), even when presented with very short stimuli (Flege, 1984; Park, 2013). As described in Section 2.2.1.1, however, L2 speakers with the same L1 may display a wide range of accent strength, while L1 speakers who have acquired another language early (e.g., simultaneous bilinguals) or have been immersed in L2-dominant environments may also develop what is perceived as a foreign accent. The cues that distinguish one group of bilingual speakers from another may therefore be subtle. In situations of long-term language contact, such as between Welsh and English, there is some evidence that listeners from the community can tell whether a speaker is bilingual or not from speech in one language, but their performance is not much higher than chance level, with a significant minority of listeners failing to perform above chance (Mayr et al., 2019).

Identifying the specific variety, namely the source of cross-linguistic influence, is a different story. When faced with forced-choice or closed-set categorisation tasks, listeners can be reasonably accurate in distinguishing L1 and L2 varieties, but their ability tends to be more limited when it comes to unfamiliar varieties (Derwing & Munro, 1997; Jarvella et al., 2001). Vieru-Dimulescu and Boula de Mareüil (2006) have found French listeners to perform better at identifying Arabic-accented French, stemming from the main immigrant group in France, than speakers with a Romance language as their L1, with particularly strong confusion between L1 speakers of Spanish versus Italian and English versus German. In that sense, it has been suggested that speakers of the target L2 variety could have some advantage over listeners of other L1 varieties (Shen & Watt, 2015). Listeners further demonstrate above-chance ability to identify L2 accents in forced-choice tasks where the stimuli are degraded to convey temporal and not spectral characteristics (Kolly & Dellwo, 2014). However, in cases of FVC, the question of variety is rarely a matter of closed options. Pinpointing specific origins prove considerably challenging to lay native listeners in free identification tasks, who often confuse L2 varieties with others in the same broad geographical region, even for varieties that are well represented in the listeners' community (Gnevsheva, 2018; McKenzie, 2015).

Beyond lay listeners, Foulkes and Wilson (2011) tested other groups of listeners in their ability to identify a Ghanaian English accent. While native Ghanaian listeners performed the best, trained phoneticians could perform on a similar level, once unsure answers were factored out. Students with some phonetic training also performed relatively well, indicating the potential value of considering phonetic information in detail. Ongoing work by Gombe Muhammad et al. (2021) on Nigerian English further emphasises the difficulty of identifying unfamiliar L2 varieties. In an experiment to identify the L1 of four varieties of Nigerian English speakers, native Nigerian speakers regardless of linguistic training were more accurate at identifying their own L1 than other L1s. While UK phoneticians as a group were not significantly outperformed by Nigerian listeners, their performance varied considerably across different varieties, with accuracy reaching as high as 80% for L1 Yoruba and below 30% for L1 Igbo.

Speech in a different language also impedes listener's ability to make judgments about other group-level characteristics. Linguistic unfamiliarity has been shown to adversely affect listeners' ability to estimate age from speech, as listeners rely on both physiological and language-specific sociolinguistic cues (Jiao et al., 2019; Nagao, 2006). L1 listeners are less accurate when estimating the age of speakers in an unfamiliar language (Nagao, 2006) or L2 accent (Gnevsheva & Bürkle, 2020). Similarly, L2 listeners are better at age estimation of other L2 speakers with the same L1 than those with different L1 (Jiao et al., 2019). The role of experience is highlighted by L2 listeners who develop familiarity with L1 accents and perform equally well with L1 and L2 speakers (Gnevsheva & Bürkle, 2020). Age estimation may form only a small part of the forensic evidence, typically on a categorical level (young vs. old), but the findings above imply that judgments about this group-level characteristic can be interlinked with the background of the speaker and are best left to expert analysts who have been properly validated.

The literature reviewed in this section shows that relying on lay listeners to adjudicate the delimitation of the relevant population in a bilingual context may prove problematic. It also adds further evidence to the concerns identified in Gold and Hughes (2014) for a perceptually based approach which collects reference data through lay listener-judged similarity. The evidence here thus provides support, in the context of bilingual speakers, for the position in Hughes and Rhodes (2018) that group-level characteristics in voice evidence can have high probative value and should be assessed by the expert as part of the evidence.

2.2.2.2 Assessing typicality

With the division of voice evidence into group-level and individual-level comes the need for both to be evaluated for similarity and typicality. To estimate the size of the relevant bilingual community within the larger population, census data may appear to be an appealing option to provide demographic statistics. Returning to the case of Cantonese-accented English in Toronto at the beginning of Chapter 1 as an example, Rogers (1998) noted that the size of the Chinese community in the Greater Toronto Area, mostly Cantonese-speaking, was estimated to be about 125,000. Contemporaneous census data estimate this figure to amount to about 3% of the Toronto population at the time (Statistics Canada, 1992).[1] If this accent-level evidence is addressed in LR terms, the LR would be $\frac{1}{0.03} \approx 33$,

[1] In fact, the 1991 census records over 175,000 speakers who reported Chinese as their mother tongue, out of the whole population of just under 3,900,000 in Toronto (Statistics Canada, 1992).

meaning that it would be 33 times more likely to find this (Cantonese) accent if the samples were produced by the suspect than if they originated from a different person in Toronto. To be clear, Rogers did not adopt the LR framework or directly reference census data. Nor did he explicitly appeal to a two-level approach to evidence akin to that later advocated in Hughes and Rhodes (2018). Yet, the consideration of group-level evidence separate from individual-level evidence was plain in his report: "The simple presence of a certain foreign accent might be useful in identifying a speaker in a small community [but] was not helpful in this case" (p. 204). In reality, there can be many shortcomings in the way census data are collected which considerably diminish their value for forensic use.

In England and Wales, which only first included questions on language use in the 2011 National Census (continued in the 2021 edition), each respondent is asked to indicate their 'main language' and only permitted to give a single answer: 'English' or 'Other' to be specified, despite being permitted to name multiple national and ethnic identities (Office for National Statistics, 2016, 2020) The assumption that a single dominant, main language must be identified, even in multilingual households, is symptomatic of a monolingual ideology in operation that discourages multilinguals, particularly those of stigmatised languages, from self-reporting as such, stifling effective collection of data on actual language use and leading to underestimation of the extent of multilingualism in the population (Mehmedbegovic & Bak, 2017).

Another issue is that state entities may choose to treat closely related language varieties as a single category, even when such varieties have major systemic differences between them and low mutual intelligibility, as in the cases of Arabic and the Chinese languages. Here the case of Cantonese in Toronto is once again enlightening. In the publicly available profiles of the Canadian censuses, Chinese languages were reported as a single "Chinese" category until 2001, when the figures for Cantonese, Mandarin and Hakka became separately reported (Statistics Canada, 2003). As L1 Cantonese speakers do not share the same accent in English as L1 Mandarin speakers (Deterding, 2006; Deterding et al., 2008), relying on census data back then could have led to an inflated estimate of the size of the Cantonese-speaking community and misestimation of the strength of the group-level evidence.

Assessing typicality for individual-level evidence can also be challenging. As mentioned above, L2 (and other bilingual) varieties and the sociolinguistic variation within are often underdescribed. As L2 speakers often only partially replicate the sociolinguistic constraints that govern structured variation in the community and may style-shift in very different ways from monolingual speakers (see Section 2.2.1.1), an understanding of how sociolinguistic variables are used in the specific groups of bilingual speakers is thus crucial. In this regard, sufficient population statistics for L2 speakers and bilingual varieties are generally lacking (cf. those available for homogeneous L1-speaking communities in Section 2.1.3), and many studies in L2 research may only have limited referential value for forensic use due to small sample sizes (Loewen & Hui, 2021; Plonsky, 2013). While quite a number of L2 speech corpora exist, their use in forensic research may be suboptimal due to restricted speaker characteristics (e.g., level of proficiency) and/or speech styles (see, e.g., CECL, 2019). In setting out a proposed protocol for collecting forensic audio databases, Morrison, Rose, et al. (2012) recommend that at least two non-contemporaneous recordings be collected from every speaker. Importantly, it is recommended that the recordings should consist of multiple natural speech styles and involve no elements of role play beyond mock police interview. The use of such pre-existing corpora to assess typicality in FVC must then be carefully considered in each case to evaluate the possible effects of mismatch not only in the population but also in speech styles, channel conditions, etc.

2.2.2.3 Cross-language comparisons

Thus far, this section has considered issues within the LR framework that relate to bilingualism at large, but mostly pertaining to the scenario where non-native features are perceived to be present in the QS, without considering yet any complications that bilingualism may give rise to in relation to the KS. This section thus turns to cross-language comparisons in particular to discuss specific pertinent issues.

As mentioned above, when assessing group-level (Level 1) evidence, the formulation of alternative propositions should be guided by the QS alone and not include characteristics of the known speaker (Robertson & Vignaux, 1995). This logical point is especially acute in cases of cross-language comparisons, as mischaracterising the group-level evidence may lead to highly overstated or understated evidence on the individual level (Level 2).

To illustrate the potential magnitude of the impact of refining the relevant population inappropriately, imagine a hypothetical case in the UK involving a QS in English and a KS in Cantonese. In accordance with the QS, and without further background information, a broad operative definition of the relevant population may include all speakers in the UK. Leaving aside other potential factors such as sex and age, the relevant population may be refined to all English speakers only on the basis of the language spoken. A specifically narrow definition of L1 Cantonese–L2 English bilinguals is only warranted if there is clear evidence of transfer from Cantonese in the QS. If L2 features can be detected but there is insufficient information to justify the choice of any specific variety, it may be more appropriate to conceive a less narrow relevant population of L2 English speakers in the UK.

If, in another hypothetical case in the UK, the languages in the two samples are reversed, such that the QS is in Cantonese and the KS is in English, the relevant group-level observations must change accordingly. Without taking other potentially relevant factors into account, the relevant population would be all Cantonese speakers in the UK. The possibility of a narrower definition depends on the presence of any accent features that can be observed in the QS. If Cantonese were the L1 or dominant language of the speaker in the QS, there might not be sufficient information in the QS to indicate bilingualism as a logically relevant factor.

The two scenarios above illustrate the significance of defining the relevant population based on the QS in cases of cross-language comparison. Guided by the language in each QS, the analyst may arrive at very different definitions in the two cases. If the KS is also taken into account when refining the relevant population, the analyst may end up with the same definition in both scenarios in spite of the completely different circumstances. The different definitions above can have further consequences downstream in the analysis of voice evidence. In the first case, the value of Level 1 evidence can vary drastically depending on a broader definition (English speakers) or a narrower one (L1 Cantonese–L2 English speakers), whereas in the second case, the observation of Cantonese speaker within the UK itself may offer the potential of strong Level 1 evidence.

The potential impact of taking into account the KS is not limited to conjoining it with the QS in defining the relevant population. If analysts are exposed to both recordings before approaching the analysis and defining the relevant population, knowledge of the involvement of multiple languages in the samples may introduce cognitive bias to the analysis (Dror, 2020), such that analysts may be unintentionally prejudiced to consider a (generally or specifically) bilingual population, as well as to select or preferentially weight features for analysis that are particularly relevant to varieties spoken by the individual in the KS. Adopting measures that seek to mitigate the effects of the KS as a source of cognitive bias, such as linear sequential unmasking (Dror et al., 2015; French & Fraser, 2018; Gold & French, 2019), is thus necessary to guard against an unduly narrow definition of the relevant population (e.g., a specific definition of Cantonese–English bilinguals in the case above).

Aside from the definition of the relevant population, in a numerical, data-driven approach to LR, the implications of language mismatch reach even further. Once the relevant population has been refined for assessment of individual-level characteristics, the relevant population then needs to be appropriately modelled for scores to be calculated and converted to LRs.

As Morrison (2013, 2018) argues, the background data should reflect the particular circumstances and conditions of the KS, so that the mismatch between the QS and the KS, which affects the numerator of the output score, is not different from the mismatch between the QS and the background data, which affects the denominator of the score. Considerations would include, among other things, the channel of transmission, sample duration and speech style (Morrison, 2013). This position is now adopted as the consensus among a group of forensic scientists (Morrison et al., 2021). Morrison (2018) further provides an empirical demonstration of the quantitative impact of not matching the conditions of the background data with those in the KS, showing that using high-quality data to model the background population instead of, inter alia, landline transmission in a reverberant environment results in much poorer performance of the system.

In cases of cross-language comparison, the reliance on the characteristics of the QS in refining the relevant population on the one hand, and the requirement to depend on the conditions of the KS on the other, can be at odds with each other. To be able to fully balance the mismatch between the two samples with the mismatch between the QS and the background data, the language of the background data used to model the relevant population has to match with that of the KS. Doing so, however, presumes that the relevant population has been defined in a way that not just encompasses speakers of the language of the QS, but more specifically people who also speak the language of the KS. If such a definition is not warranted by the evidence in the QS, pressing on along this route of analysis becomes logically problematic. If the language of the background data is matched with that of the QS instead, the mismatch between the QS and the KS, at least in terms of language, becomes different from that between the QS and the background data (where there is in fact no mismatch), thus potentially resulting in worse system performance.

Aside from the logical issue, the availability of reference data in such cases may be severely limited. To properly convert output scores from comparisons to interpretable LRs, they have to be calibrated using training data that match the conditions and circumstances of the case. Cross-language comparisons would thus require a system to be trained on comparisons involving language mismatch as well, entailing the necessity for reference materials from both languages from the same set of speakers, with the same channel and audio characteristics as the samples compared, to be available for training in the first place. To date, existing speech corpora designed for forensic purposes have focused on collecting speech in a single language, and provide little information as to whether their speakers are bilingual, or make no express consideration of such bilingual possibilities (e.g., Jessen et al., 2005; Morrison, Rose, et al., 2012; Nolan et al., 2009). Large-scale corpora of spontaneous speech that contain samples in multiple languages from the same bilingual individuals are few and far between (e.g., Bradlow, n.d.; Campbell et al., 2004; J. King et al., 2010; Kupisch et al., 2012; Patil & Basu, 2008), and those designed to elicit natural code-switching are yet rarer (Deuchar et al., 2018; Johnson, 2021b; Lyu et al., 2015).

Where a pre-existing reference database of recordings is not available, one preferred approach that has been put forward is for the analyst to gather their own data specific to the needs of the case (P. Rose, 2007). However, practical considerations such as financial resources and time pressure, which are compounded when speech from multiple languages is to be collected, often preclude such a procedure from being carried out. If the population of the circumscribed community is small, obtaining a sufficiently large sample of the population could be extraordinarily difficult. At the same time, if the group-level characteristics available from the questioned recording do not permit the relevant population to be narrowly refined to some bilingual community for the evaluation of individual-level evidence, collecting speech data in the language of the KS would only be possible for a subset of the population who can actually speak the language and would unlikely result in a representative sample.

If an analyst only has access to reference recordings from one of the languages in the samples, proceeding with an FVC analysis would result in a mismatch between the circumstances of the case, which involve cross-language comparisons, and the conditions of the training data, which can only involve same-language comparisons. The discrepancy may result in score distributions in the training data that are unsuited to the case and miscalibrate the output comparison scores, leading to poor performance and unreliable evidence. Such an approach is logically undesirable and could be forensically damaging, but the quantitative impact remains to be empirically determined.

2.2.3 Summary

In this section, a number of issues of bilingualism in FVC have been identified. While L2 speakers are often said to produce features that make them sound non-native, research has shown that there is a great deal of variability within each L2 variety. The fact that L1 speakers of a language can also develop non-native-sounding features makes it even more challenging to make use of such features as the basis for narrowing down the relevant population when they occur in the QS, as well as to properly assess typicality in the evidence. Further, potential code-switching by bilinguals has implications not only for forensic analysis of any single samples, but also for the possibility of cross-language comparisons. Language mismatch between the QS and the KS poses particular challenges to each step of FVC analysis, not only affecting the pool of features potentially available for analysis, but also causing theoretical and methodological issues to an LR-based approach to FVC.

2.3 Previous work

To date, there have been few studies focused on issues of bilingualism in the context of FVC, especially within the LR framework. One study that made such an attempt is Frost and Ishihara (2015), who investigated the viability of LR-based FVC in L2 Hong Kong English. Working on the assumption that speakers tend to exhibit higher within-speaker variation in a more fluid L2 variety, the authors examined the discriminatory potential of formant dynamics from a small set of target words such as "hello" and "yes" among 15 speakers of the variety. Comparing the performance of the same diphthongs in Hong Kong English with those in Australian English, they found that validity metrics were at similar levels and concluded that there was no major disadvantage for conducting the analysis in the L2 variety. While this conclusion could be encouraging for the wider use of LRs in different speech communities, there were a number of shortcomings with the study that likely led to overly optimistic results. The low number of speakers means that only limited between-speaker variability was represented in the sample. Only a very small number of tokens (five per word) were used, with the implication that within-speaker variability may not be adequately estimated (Hughes & Foulkes, 2015b). The choice to restrict the analysed context to single lexical items further meant that only limited within-speaker variability could be accounted for.

Within the context of bilingual speakers, rhythmic and temporal measures are among acoustic-phonetic features that have been examined the most. Speaker-specificity of such parameters across languages has been argued to arise from idiosyncratic ways of coordinating articulatory movement, which should survive language shift to some extent (Dellwo & Schmid, 2016). This idea has received limited support from Y. Zhang et al. (2019), who examined the percentage of voiced interval duration (%VO) in bilingual Mandarin–Wu Chinese speakers.[2] Using read sentences and passages, the authors found %VO variability to be largely dominated by language-specific patterns, where speakers

[2] It is acknowledged that Y. Zhang et al. (2019) considered these speaker bidialectal rather than bilingual, although the authors did point out that Mandarin and Wu are mutually unintelligible and have different phoneme inventories. See, e.g., Mair (1991) for a discussion of the language/dialect debate in the Chinese context.

consistently produced higher %VO in Wu, which historically retained voiced initial obstruents, than in Mandarin. Although a small number of speakers exhibited little cross-linguistic within-speaker variability, the magnitude of cross-linguistic shift was highly inconsistent across speakers. It was thus concluded that speaker individuality of %VO was not cross-linguistically robust. Kolly et al. (2015) similarly found less consistent effects of language on the number and total duration of pauses produced by 16 L1 German speakers in German, English and French. While speakers overall produced the least and the shortest pauses in L1 German and the most and the longest pauses in French, the language they were least proficient in, seven speakers did not show any significant cross-linguistic differences in their pausing behaviour, suggesting that there is low within-speaker variability across languages for this set of features.

The potential use of temporal measures in cross-language comparisons has also been explored within the LR framework in Tomić (2017), who elicited non-contemporaneous spontaneous L1 Serbian and L2 English speech from 10 female bilinguals. LRs were calculated separately for each parameter (including articulation rate, speaking rate and measures of pause duration), using English speech as the QS, Serbian speech as the KS and reference data pooled together from all speakers' Serbian speech. LRs from uncorrelated parameters were further combined through naïve Bayes multiplication. Overall, the chosen parameters were found to perform rather poorly, with high error rates in both same- and different-speaker comparisons. While this could be taken as an indication of the difficulty of applying a numerical LR approach to cases of language mismatch, there are a number of limitations to the study. The small number of speakers was likely insufficient to build a robust representation of the relevant population, and may lead to much instability in the scores obtained. LRs were also uncalibrated, meaning that performance was not optimised within these parameters and the "LRs" should not have been directly interpreted in their absolute values. The choice of features may have further contributed to the poor performance, as the utility of articulation rate is limited even in the absence of language mismatch (Gold, 2018) and the discriminatory potential of speaking rate is even weaker (Künzel, 1997).

Another long-term acoustic-phonetic variable which has been considered for its discriminatory potential in cross-language comparisons is LTFDs. Despite findings of cross-linguistic differences, as mentioned above, bilinguals typically produce highly similar LTFDs, and within-speaker variability across language has been found to be lower than between-speaker variability (S. Cho & Munro, 2017; Heeren et al., 2014). Further, an attempt to carry out an LR-based examination of LTFDs in cross-language comparisons has been carried out in Tomić and French (2019), using a similar paradigm to that in Tomić (2017). Studies on LTFDs are reviewed in greater detail in Chapter 3.

A separate line of inquiry has investigated the speaker-specificity of FPs in bilingual speakers. Comparing the distribution and realisation of FPs in 21 bilingual speakers of New Zealand English and te reo Māori, Wong and Papp (2018) found both significant within-speaker specificity and between-speaker variability in their acoustic realisations. FPs can thus be considered high on the cline of transferability from one language to another, thereby potentially serving as good cross-linguistic speaker discriminants. Cross-linguistic discriminatory performance of FPs was similarly tested by means of linear discriminant analysis (LDA) for 20 Dutch–English bilinguals in de Boer and Heeren (2019), where systematic differences in the production of the phonetic target in L1 Dutch and L2 English led to deteriorated performance when the language tested was different from the one trained in the LDA.

The cross-linguistic speaker-specificity of segmental variables has yet to be investigated in the LR framework, but has been considered in Marquina Zarauza (2016), who compared formant frequencies for [a l], VOT for [k], and spectral moments (and intensity) of [s], alongside articulation rate and f_0, among 22 male Catalan–Spanish bilinguals. Using generalised linear mixed models, the author analysed the effect of the factors of language, speaker and repetition (as each speaker was recorded twice per language) on each acoustic parameter, followed by post hoc comparisons between each pair of speakers. The author found that, of all the parameters tested, the vast majority exhibited high between-speaker variability and low within-speaker variability within each language, as a high proportion of post hoc comparisons between different speakers were returned as significant, while almost no speaker showed significant differences with themselves. Furthermore, with the exception of F2 for [a] and [l], significant cross-linguistic differences were found for less than a third of the speakers in all other tested parameters. The author thus concluded that the use of these acoustic-phonetic variables were not invalidated in cross-language comparisons. While these results are suggestive of the speaker-specificity of these variables, their reference value for application in FVC is highly restricted. Post hoc pairwise comparisons only go toward ascertaining the similarity between two samples, but do not take the typicality of the values into account. Estimating the discriminatory potential of features this way could thus be misleading, as a focus on only similarity belies the actual strength of evidence that crucially depends also on such similarities being evaluated against the relevant population. In addition, the effect of language was only evaluated within each speaker but not between speakers, and comparisons between different speakers were only conducted within the same language. Thus, while many speakers appeared to demonstrate cross-linguistic consistency, there was no assessment of whether (or how) speakers could be discriminated from each other. Support for the claim that the tested variables would be suitable for cross-language comparisons is limited at best.

As for ASR systems, earlier studies generally report degraded performance in speaker recognition when there is some form of language mismatch in the data (e.g., Akbacak & Hansen, 2007; Auckenthaler et al., 2001; Ma et al., 2007). The issue of language mismatch between training and test data has been highlighted in the NIST Speaker Recognition Evaluations, most recently in the 2006 edition, which is reviewed in Przybocki et al. (2007). The authors pointed out a clear impact of language mismatch on system performance; notably, non-English comparisons were more heavily affected than comparisons in English. More recently, Künzel (2013) set out to examine the effect of language mismatch on ASR within the FVC context. He tested the *Batvox 3.1* system by Agnitio on 75 bilinguals with assorted language backgrounds, using a case-specific normalisation procedure in the system designed to reduce errors due to channel and linguistic similarities. Cross-language comparisons resulted in close to no deterioration in performance when compared to same-language comparisons, although this may have been facilitated by limitations in the data, including the low number of speakers in each language pair and the homogeneity of the speech material.

Research on the speaker-specificity of acoustic-phonetic features in bilingual speakers, and more specifically across languages, has thus so far only considered a small number of variables within small datasets. Parameters relevant to auditory-perceptual and acoustic-phonetic approaches are left underexplored, inviting further research on the topic. As these prevailing approaches remain dominant practices in the field (Gold & French, 2011), there is a compelling need to critically evaluate forensically relevant features on all levels in order to examine the practical implications for current approaches. The current study thus seeks to contribute to the field by investigating such features in depth.

2.4 Current study

The literature reviewed in this chapter has clearly shown that issues of bilingualism challenge various components of LR-based FVC analysis. In particular, cross-linguistic differences have made the application of a fully numerical LR-based approach in FVC difficult in cases of cross-language comparison, especially outside the use of ASR software. Thus, the current study focuses on the issue of language mismatch in FVC. Specifically, the primary aim of the project is to investigate the performance of linguistic-phonetic features in cross-language comparisons within a bilingual community. Operating within the numerical LR framework, this study seeks to quantitatively assess the impact of language mismatch on the discriminatory power of the features. In face of the difficulties of obtaining suitable reference databases for system training and calibration in cross-language comparisons, this study additionally considers the practical impact of mismatch between the case conditions and those in the reference data on system performance.

With a focus on bilingual communities, a secondary aim of the project is to assess the effect of language on the speaker-discriminatory potential of features. As the literature review throughout this chapter demonstrates, research into the discriminatory potential of many features has been carried out in a number of languages and varieties. Investigations using disjoint speaker populations and different conditions across studies, however, render the effect of language difficult to be isolated from other factors. Leveraging the use of the same bilingual speakers in different languages, this study seeks to examine the extent to which the discriminatory power of the same features is language-specific.

2.4.1 Research questions

The current study asks the following research questions:

1. What is the effect of language on a feature's discriminatory potential? In other words, how language-specific is the speaker-specificity of any particular feature?

2. What is the effect of language mismatch between the QS and the KS on the discriminatory potential of a feature?

   2.1 In the context of cross-language comparisons, what is the effect of mismatch between the conditions of the case and the conditions of the reference data on system performance?

A particular consideration in exploring the effects of language mismatch is that the speaker-discriminatory potential of features that are highly speaker-specific and language-independent is not expected to be adversely affected in cross-language comparisons, when compared to same-language comparisons. Features that are largely constrained by language-specific effects and systematically distinguished by bilinguals across languages, on the other hand, will unlikely offer much discriminatory value in cross-language comparisons.

Therefore, to investigate these questions, the present study focuses on two linguistic-phonetic features, chosen for their attested speaker-discriminatory potential in the literature, within the population of Canadian English–French bilinguals. These are, on the segmental level, the sibilant fricative /s/ and, on the suprasegmental level, long-term formant distributions (LTFDs).

As the literature reviewed in the next chapter demonstrates, /s/ and LTFDs are expected to show different levels of language dependency among bilingual speakers of English and French, thus in theory allowing both sets of predictions above for cross-linguistic speaker-discriminatory potential to be tested. /s/ is not predicted to exhibit systematic cross-linguistic differences, and as such its effectiveness in FVC is not expected to suffer from effects of language mismatch. Meanwhile, LTFDs are expected to show language-specific tendencies. Their performance is accordingly predicted to deteriorate in cross-language comparisons. In light of claims in the literature that ASR software is minimally affected by language mismatch (Künzel, 2013), the discriminatory performance of the two linguistic-phonetic features is additionally compared with that of a commercially available ASR software (*Phonexia Voice Inspector v4.0*).

Chapter 3

Research Variables

The current chapter outlines the motivations behind the choice of /s/ and LTFDs as potentially useful features in FVC in different languages, as well as more specifically in cross-language comparisons. Section 3.1 first surveys the literature on /s/ with reference to the criteria for the ideal speaker discriminant set out in Nolan (1983), then reviews previous forensic research on its discriminatory potential, before turning to consider the production of /s/ in English–French bilinguals in the current study and lay out specific research questions and hypotheses. Section 3.2 reviews the existing literature on LTFDs, as well as related research on cross-linguistic articulatory settings, before concluding with specific predictions for the production of LTFDs in the current study.

3.1 Alveolar fricative /s/

3.1.1 Within- & between-speaker variability

The alveolar fricative /s/ is produced by raising the tip or blade of the tongue towards the roof of the mouth between the dental to alveolar region, forming a narrow constriction that causes the airflow passing through to result in turbulent noise, which is enhanced by the airflow striking the upper teeth further downstream (Ladefoged & Maddieson, 1996, p. 145). As a strident and sibilant, /s/ is acoustically characterised by sound energy that is concentrated in the high frequency region, as well as relatively high intensity when compared to other fricatives (Strevens, 1960). In order to maintain a narrow aperture to generate high frequency noise, articulators such as the jaw and the front of the tongue body have to be aligned with a high degree of precision. While the articulators do not remain static during the production of /s/ (Iskarous et al., 2008), the gesture and positioning of articulators for individual speakers are constrained with a high degree of variation from one token of /s/ to another, providing a favourable basis for low within-speaker variability.

As the spectral shape of the turbulent noise in /s/ is largely determined by the shape of the cavity in front of the point of constriction, the morphology of a speaker's vocal tract has an important role in how /s/ is produced. The precise shapes of the alveolar ridge, the lower jaw, the upper teeth, as well as the location of the ridges along the palate, have all been said to play a part in shaping the aerodynamic and acoustic output of the turbulent airstream (Ladefoged & Maddieson, 1996). Regardless of the precision of articulatory configuration that is paramount to the production of /s/, as the shape of the vocal tract varies from speaker to speaker, a wide range of gestural and articulatory possibilities is available to speakers, some with more pronounced acoustic consequences than others. Findings from Fletcher and Newman (1991) suggest that the width of the groove and the place where it is formed may be traded off without significant impact on the acoustics of /s/. The curvature of the palate, however, significantly influences the positional variability of the tongue blade in /s/ production (Rudy & Yunusova, 2013). The role of the palate is highlighted by the ability of speakers to adapt to artificially introduced changes to the palate, so that they can achieve a similar centre of gravity (CoG) in their /s/ with different places of articulation (Thibeault et al., 2011). Other aspects in the gesture for /s/ can also be varied without impairing the ability to achieve the requisite turbulent noise. In many languages, there is no phonological contrast between dental and alveolar, or apical and laminal /s/ (exceptions include Toda, which contrasts laminal dento-alveolar /s̻/ and apical alveolar /s/, see Gordon et al., 2002). As such, the choice of place and manner of articulation can act as a source of between-speaker variability. Evidence of individual preference for apical or laminal /s/ has been found in speakers of English and French (Dart, 1991; Toda, 2009), as well as for Beijing Mandarin (Lee, 1999). Acoustically, such preference can be reflected in the different shapes of the spectra for apical and laminal variants of /s/ (Dart, 1991, pp. 83–85). The speaker dependency of apicality, in English at least, has been further supported by articulatory data from magnetic resonance imaging (Narayanan et al., 1995). The speakers of English and French in Dart (1991) similarly made use of a range of places of articulation along the dental–alveolar continuum, contrary to the dichotomous divide depicted in traditional descriptions.

There is also some evidence for the relatively low within-speaker variability of /s/ in articulatory research. In English, Dart (1991, 1998) notes that all but one of her participants maintained their tongue-tip gesture across different tokens of /s z/, where both apical and laminal articulations are available options. M. J. McAuliffe et al. (2001) also found, in an electropalatographic (EPG) study where the vowel context was kept constant, that /s/ consistently demonstrated lower intraspeaker variability than /t l k/, attributing its relatively consistency to an anchoring effect provided by full lateral contact between the tongue and the palate. Studies in other languages yielded similar results. In German, for example, sibilants, in particular /s/, display the lowest variability among coronal consonants in midpoint tongue-tip and jaw positions within each speaker (Mooshammer et al., 2007). For most speakers, an increase in vocal effort, which is common in forensic materials involving a noisy background or telephone speech (French, 1998; Jessen et al., 2005), makes no difference to their jaw positions during their production of sibilants (Mooshammer et al., 2007), making them potentially more desirable candidates in casework.

Within-speaker variation in /s/ is predicted to be low relative to other consonants because it is less susceptible to lingual coarticulatory influences (Recasens & Espinosa, 2009; Recasens et al., 1997). Its high coarticulatory resistance found early empirical support in Bladon and Nolan (1977), who showed that /s z/, which are generally laminal for their English speakers, spread their laminality to neighbouring alveolar consonants /t d n l/ in CC clusters or CVC sequences, which are otherwise typically apical, while remaining steadfastly laminal themselves. Acoustic and EPG data from Tabain (2001) provide further corroboration of the coarticulatory resistance of sibilants. The Australian English speakers in the study produced both alveolar (/s z/) and postalveolar (/ʃ ʒ/) sibilants in accented CV syllables with very low variability in spectral and EPG centres of gravity (CoGs), in contrast with the wide range found in dental fricatives /θ ð/. /ʃ/ showed even less extensive coarticulatory effects than /s/, due to the involvement of a raised tongue body in the former.

Coarticulatory effects from other articulators, especially the lips, are well attested. Indeed, the prominent effects of adjacent rounded vowels or consonants on /s/ are well known. The gesture of lip protrusion is spread to /s/ (Shadle & Scully, 1995), resulting in a general downward shift of acoustic energy, such that spectral measures like spectral peak and CoG decrease as a consequence (Tabain, 2001). Iskarous et al. (2013) quantify the susceptibility of /s/ to lip protrusion with the concept of mutual information (MI), a measure of how dependent an articulator's position for a particular consonant is on its position in neighbouring segments. MI for the lower lip in American English /s/ is found to be lower and to increase much later in the vertical dimension than in the horizontal dimension, indicating that /s/ is impervious to variation that influences the width of the constriction, but much less resistant to effects from lip protrusion. A dynamic examination of MI for the lip, jaw and tongue tip further shows that coarticulatory effects from the following segment set in only just before the end of the fricative. Coarticulation, however, may nevertheless act as a further source of speaker-specific information, as Yu (2016) found considerable individual variation in the degree of influence that the following vowel has on the realisation of /s/ in Cantonese.

Other pre-consonantal environments in English, most notably /stj/ and /str/ clusters, are also known to induce /s/ retraction, such that the phoneme acquires a more [ʃ]-like quality. /str/-retraction is now widely documented in numerous varieties of North American English (Baker et al., 2011; Durian, 2007; Rutter, 2011; Stuart-Smith et al., 2019; Wilbanks, 2017), British English (Altendorf et al., 2021; Bailey et al., 2019), as well as Australian English (M. Stevens & Harrington, 2016) and Trinidadian English (Ahlers & Meer, 2019). Competing theories have been propounded as to its phonetic motivations, including long-distance assimilation to /r/ (Shapiro, 1995) and assimilation to /t/ that is itself retracted and affricated (Lawrence, 2000). While current evidence has yet to rule out either account, parallel synchronic and diachronic patterns of retraction in /str/ and /stj/ provide support for assimilation of affricated /t/ over long-distance assimilation (Altendorf et al., 2021; Bailey et al., 2019).

In addition to phonological factors, within- and between-speaker variability of /s/ also arises from sociolinguistic variation. Most notably, /s/ is known to be a marker of gender (Fuchs & Toda, 2010; Stuart-Smith, 2007, 2020), with a cross-linguistically widespread association of fronted /s/ with feminine or non-masculine gender norms (Bekker & Levon, 2017; Pharao et al., 2014). /s/ has been found to vary in topic-based style shifting among gay male speakers, among whom fronted /s/ is argued to function as a linguistic resource in the construction of a counter-hegemonic gay persona (Boyd, 2018). In Southeast England, the quality of /s/ is also found to index the level of interactional threat (as per Brown & Levinson, 1987), such that female speakers adopt a more backed /s/ when engaging in speech activities that carry greater face-threat (e.g., confrontation) and a more fronted /s/ in less threatening activities (e.g., information sharing) (Holmes-Elliott & Levon, 2017). Other sources of variation between speakers may also arise in forensically relevant scenarios: Different face-concealing garments attenuate energy in the higher frequencies more than others, significantly impacting the CoG of voiceless fricatives, but individual speakers may adopt different strategies to compensate for the effects brought about by the facewear, such that the resultant spectrum reveals different properties (Fecher, 2014).

In summary, the body of research reviewed above shows that /s/ offers a rich source of between-speaker variation within a population. At the same time, within-speaker variability of /s/, while compared unfavourably to /ʃ/, has been found to be relatively low compared to other consonants. /s/ is therefore considered to be a good segmental candidate for use in FVC as a speaker discriminant.

3.1.2 Practical considerations

This section considers other practical considerations that also contribute to the suitability of a feature for the task of speaker comparison. Characterised by the presence of high frication energy, sibilants are typically salient on spectrograms and can be easily distinguished from surrounding segments in most cases. Turk et al. (2006), in their methodological guide, found sibilants to be among the most reliably segmented consonants, demarcated by "the onset and offset of frication energy" (p. 10), though their segmentability is compromised in clusters that are homorganic or share a manner of articulation. Sibilants are also typically of higher intensity than non-sibilants, meaning that their acoustic properties are not as compromised in noisy conditions.

In languages that have the phoneme, /s/ is exceedingly common in everyday speech. /s/ is placed as the third and fourth most frequently occurring consonant in two independent analyses of spontaneous American English speech, accounting for 4.89% and 4.61% of all phoneme occurrences (Hayden, 1950; Mines et al., 1978). Its high frequency is also borne out in a wide array of languages, such as French (6.00-6.12%; Malécot, 1974; Wioland, 1985), German (3.31%; R. King, 1966) and American Spanish (9.4%; Guirao & García Jurado, 2009). By contrast, occurrence of the postalveolar /ʃ/ is decidedly rare than its dento-alveolar counterpart, in languages that have both sounds in their phonemic inventories. Data from R. King (1966) suggest that, in German, /ʃ/ occurs only about half as frequently as /s/ does. The disparity between /s/ and /ʃ/ is even more considerable in English and French, where /ʃ/ accounts for less than 1% of all phonemes (Hayden, 1950; Malécot, 1974; Mines et al., 1978; Wioland, 1985). This contrast between /s/ and /ʃ/ is reflected in the phonetically balanced materials used in the corpus in this study (see Appendix A), meaning that there are tokens of /s/ in abundance, but only extremely few instances of /ʃ/.

The ready availability of /s/ in speech, as well as its relative ease of extraction, further buttress the proposition that /s/ may be a highly suitable candidate for use in FVC, not only in English but also in many other languages. It must be noted, however, that forensic materials are generally not conducive to the transmission of /s/, since many questioned samples in casework are in the form of telephone speech, which is typically transmitted in a narrow bandwidth range of around 300 to 3400 Hz. As most of the sound energy of /s/ is concentrated above this range, only a small portion of acoustic information of /s/ survives in telephone speech. The loss of information consequently leads to reduced intelligibility of /s/ (Fernández Gallardo & Möller, 2015), and could further severely limit the discriminatory potential of /s/ in forensic casework if little speaker-specific information of the sound resides within the telephone bandwidth.

3.1.3 Discriminatory potential of /s/

Early approaches to ascertaining the discriminatory potentials of /s/ from the field of ASR primarily adopted some form of filterbank analysis (Bonastre & Méloni, 1994; Bonastre et al., 1991; Magrin-Chagnolleau et al., 1995), where the full frequency range is divided into a number of narrow ranges (bands), and a spectrum of /s/ is parameterised with a series of numbers representing the average sound energy level within each band. In these systems, /s/ and other fricatives consistently underperformed relative to other tested variables, such as vowels and nasals. An example of such an approach comes from van den Heuvel (1996), who compared the speaker-specificity of /s/ in Dutch speakers against vowels /a i u/ and nasals /m n/ using smoothed, Bark-scaled filterband spectra. Out of all tested phonemes, /s/ performed the worst in a linear discriminant analysis (LDA) of 15 speakers, scoring a classification rate of 52.0% when four discriminant functions were used (compared with 73.7% for /n/, the best-performing consonant). Classification accuracy rose to 76.0% when the maximum of 14 functions were used (cf. /n/: 90.6%).

While van den Heuvel (1996) did not discuss why /s/ offered comparatively weak speaker-specificity, methodological decisions in the design of the experiment may have contributed to its relatively poor performance. The test was conducted using repeated tokens of pseudowords, in the form of /C_1VC_2ə/, spoken in isolation, which might have elicited more careful pronunciation that is not reflective of the speakers' behaviour in connected speech. In addition, each vowel and consonant only occurred in a very limited set of contexts, such that there was very little scope for the high coarticulatory resistance of /s/ to be put to test against other segments, which would conceivably undergo more extensive coarticulation when placed in a wider range of environments. A further possible reason is that, while filterband spectra can effectively discriminate between different fricative consonants (Akpanglo-Nartey, 1982), filterband energy levels themselves may not be well-suited to capture speaker-specific properties of /s/. In van den Heuvel (1996), the five filterbands that made the most significant contribution to the discriminant functions corresponded to the frequency ranges of approximately 630–1160 Hz and 2810–3950 Hz. On the other hand, frequencies between 4 and 8 kHz, where most of the acoustic energy is found in /s/, were compressed into four filterbands, resulting in an extremely low frequency resolution. Any speaker-specific information within that range is likely highly compromised.

More recently, studies have incorporated spectral moments and other acoustic properties in their investigations of /s/ as a speaker discriminant to greater success, suggesting that better representation of speaker-specificity for /s/ could indeed lie beyond filterbands. Cicres (2011) made use of a wide range of spectral properties, including the first four spectral moments and spectral peaks, as well as normalised long-term average spectrum (LTAS) bands (similar to filterbands), and found excellent performance for all Spanish voiceless fricatives (/f θ s x/). LDA classification rates reached approximately 90% for all fricatives and, in a cross-validated procedure, remained above chance level (16.7%), ranging from 58.4% for /f/ to 79.4% for /s/. All four spectral moments contributed significantly to the first two discriminant functions, confirming their utility in parameterising /s/ for speaker discrimination. LTAS bands corresponding to the frequency range of 4–8 kHz also contributed to the first discriminant function, albeit with much lower weighting, alongside some higher LTAS bands. The small contribution of LTAS bands corroborates the suggestion above that much of the speaker-specific information in /s/ is not encoded in these bands, but in the spectral moments that characterise the shape of its spectra. The discrimination task was conducted on only six speakers, and so the extent of between-speaker variability in this small set may not accurately represent that in the population. Nevertheless, these results can be viewed as an affirmation of the discriminatory potential of /s/, and in particular the value of spectral moments.

/s/ similarly finds promise as a consonantal discriminant, alongside the nasals /m n/, in Kavanagh (2012), who conducted a two-pronged analysis with both LDA and LR-based testing. In a cross-validated LDA involving 30 speakers of British English, where the signal was lowpass filtered at 8 kHz, the use of all four spectral moments measured at midpoint yielded a classification rate of 28%. Including normalised duration achieved a marginally higher classification rate of 29%. While the classification rates were low relative to that obtained in Cicres (2011), they were nonetheless well above chance level (3.3%). When a combination of the four spectral moments and normalised duration was tested within the LR framework, the best-performing system yielded a relatively low EER of 17% and C_{llr} of 0.52, indicating that /s/ was indeed useful for distinguishing between speakers. However, there was much fluctuation in its performance as the bandwidth condition and the number of speakers varied (Table 3.1). The lowest bandwidth condition (4 kHz) outperformed most other systems when the number of speakers stayed the same, while the highest C_{llr} (1.08) came from the system with the most speakers in the higher bandwidth condition. Therefore, speaker discrimination by /s/, somewhat surprisingly, did not seem to benefit much from extra spectral information above 4 kHz, even though most of the acoustic energy lies in higher frequencies. In line with findings from van den Heuvel (1996), this comparison provides encouraging results for application in forensic casework, where the majority of materials are in the form of telephone-transmitted audio. At the same time, the low number of speakers and tokens in Kavanagh (2012) raises concerns over the robustness of the resultant validity measures (Hughes, 2014; Kinoshita & Ishihara, 2014): The number of tokens from any individual speaker ranged from 17 to as low as 6, which was further halved in the process of building the individual speaker model.

No. of speakers	Bandpass (kHz)	EER (%)	C_{llr}
18	16	17	0.52
18	4	17	0.55
18	8	28	0.64
18	22	17	0.77
30	4	23	0.79
30	8	23	1.08

Table 3.1: System validity in Kavanagh (2012), arranged in ascending order of C_{llr}.

The idea that /s/ retains speaker-specificity in degraded conditions receives further support from testing conducted using a larger corpus of spontaneous telephone speech. Using the technique of multinomial logistic regression for speaker classification, Smorenburg and Heeren (2020) demonstrate the classification accuracy of both /s/ (19.5%) and /x/ (18.4%) in Dutch to be well above chance level (2.3%), with the spectral moments CoG and SD making the most substantial contribution to classification accuracy, while duration and amplitude made little difference to the outcome. Crucially, even with a limited bandwidth of 340–3400 Hz, /s/ outperformed /x/, despite the fact that most of the sound energy falls within this range only for the latter.

The prospects of dynamic measurements providing richer speaker-specific information were also explored in Kavanagh (2012), by subjecting combinations of spectral measurements taken from the onset, midpoint and offset of each /s/ token to LDA classification. No improvement over the performance of static measurements was found, as the best-performing systems resulted in classification rates almost identical to those using static measurements. It was pointed out that the usefulness of dynamic measurements in speaker comparison should not be easily dismissed, as their performance was very likely restricted by the aforementioned low number of tokens per speaker, which directly limited the number of predictors that could be employed in LDA. Nevertheless, the addition of dynamic information in Smorenburg and Heeren (2020), in the form of coefficients from quadratic curves fitted to the CoG trajectory, was similarly shown to be ineffective. Existing work thus suggests that, despite the dynamic nature of the sibilant, its speaker-specific information can be largely captured by static spectral measures.

3.1.4 Summary

Overall, the literature reviewed in this section suggests that, cross-linguistically, /s/ satisfies many of the key desiderata for analysis in FVC and is likely a useful segmental feature for this task. Previous explorations into the speaker-specificity of /s/ indicate that its spectral properties show particular promise as good speaker discriminants, although the benefits of including dynamic information appear to be limited. In spite of the loss of acoustic information in telephone speech, existing evidence suggests, rather unexpectedly, that the performance of /s/ does not suffer much in limited bandwidth conditions. /s/ therefore remains a highly relevant feature for examination as a speaker discriminant in forensic casework.

3.1.5 Current study

The current study seeks to extend the investigation of /s/ in FVC to cross-language comparisons in an English–French bilingual community. Since there has been no research on the acquisition of /s/ by bilingual speakers of these two languages to date, an acoustic analysis of /s/ produced by English–French bilinguals is first carried out in the present study as a precursor to the forensic investigation. This section first describes the production of /s/ in English and French, then summarises the literature on /s/ in bilingual speakers, before turning to the specific research hypotheses for this study.

3.1.5.1 /s/ in English and French

Coronal consonants in English, including /s/, have traditionally been described as having an alveolar articulation. This contrasts with French coronal consonants, which are often thought to be dental rather than alveolar. A survey of early literature on the subject in Dart (1991) found that coronal stops /t d/ were indeed unanimously described to be alveolar in English and dental in French. There was also consensus that English /s z/ were alveolar, whereas descriptions for the fricatives in French were equally divided between dental and alveolar. As for the precise gesture of the tongue, in both English and French, most sources reported that /t d n l/ were articulated with an apical gesture, but the treatment of /s z/ in this regard was again somewhat different from that of the other coronal consonants. For sources that commented on the gesture of /s z/, these fricatives were laminal, not apical, in French, but in English there was no clear-cut agreement one way or another for /s z/, with many early sources contending that they could be either apical or laminal.

Dart's own articulatory study by and large confirmed the place of articulation and tongue-tip gesture reported in the literature for /t d/. With the aid of linguagrams and palatograms, she found that most Parisian French speakers opted for an apicolaminal or laminal gesture and a highly dental place of articulation when producing the stops, while West Coast American English speakers by far adopted an apical gesture and a farther back place of articulation. The purported contrasts between /s z/ in English and their counterparts in French, however, were not borne out. Her English speakers produced 57.5% of /s z/ in a laminal manner and 42.5% in an apical one, while her French speakers produced 68.4% of /s z/ in a laminal manner and 31.6% in an apical one. It was thus concluded that the apicality or laminality of the fricatives does not serve as an important distinguishing variable between the two languages. As for the place of articulation, French /s z/ tended to be articulated with closure contacting some part of the upper incisors, whereas English speakers had a tendency to produce /s z/ at a slightly farther back place of articulation than that of the French speakers, though no significant variation was found from one language to another. Laminal productions of /s z/ seemed to correlate with a farther back place of articulation in English than in French, but no such difference was observed for apical fricatives. Between the two languages, there appears to be more similarities in their /s z/ than was assumed. Speakers of both languages similarly employ more laminal than apical gestures, and a place of articulation that is just contacting or just behind the upper incisors.

These findings receive support from subsequent articulatory studies. In a cross-linguistic MRI examination of sibilant fricatives, Toda (2009) found considerable variability of place of articulation between different French speakers, ranging from dental to alveolar, and in the case of one speaker, even farther back at alveolo-postalveolar. The French speakers were also equally divided in whether their /s/ was apical or laminal. The English speakers in this study behaved very similarly to the French speakers, and likewise exhibited much between-speaker variability in their production of /s/, providing evidence in corroboration of Dart (1998) that the apparent divide between dental and alveolar, and apical and laminal /s/ in English and French is not as well established as previously believed. While corresponding data from Canadian varieties are as yet lacking, kinematic analysis of /t d/ in a preliminary investigation of Canadian English and French suggests that the behaviour of speakers of these varieties largely conforms with that reported in previous research, at least with regard to tongue-tip gesture: Coronal stops in Canadian English tend to be articulated with an apical gesture, while they tend to be laminal in Canadian French (Brajot et al., 2013).

The similarities of /s/ between English and French were also borne out in the acoustic portion of Dart (1991), where the spectra of /s/ had similar shapes in the two languages. The main acoustic differences were found instead between different places of articulation and tongue gestures. In both languages, the highest concentration of sound energy was located in the highest frequencies (near 8 kHz) for apical fricatives, though there were systematic differences between /s/ produced with dental versus alveolar articulation. Spectra of dental /s/ were almost flat up to about 5 kHz, whereas spectra of alveolar /s/ started rising in relative intensity at a much lower frequency. Spectral differences between places of articulation were found to be particularly striking in laminally produced /s/. While in laminal dental /s/ the spectra showed a rise similar to, if not steeper than, that found in /s/ with an apical articulation, laminal alveolar /s/ had spectra that were essentially flat.[1]

3.1.5.2 /s/ in bilingual speech production

Cross-linguistically, the acoustic quality of /s/, characterised by its high-frequency noise, is highly similar. Nevertheless, current evidence suggests that some bilingual speakers may be attuned to fine-grained acoustic-phonetic differences and accordingly shift their realisation when switching between languages.

Kitikanan et al. (2015) showed that L1 Thai–L2 English speakers produced /s/ with significantly higher spectral peak in their L1 Thai than in their L2 English, which in turn had a higher spectral peak than /s/ produced by British English speakers. Cross-linguistic differences were also reported in spectral moments to some extent, with /s/ in Thai exhibiting higher CoG and (for female speakers only) lower skewness than /s/ in English. Although no differences between L1 Thai, L2 English and L1 English were found for SD and kurtosis, the differences in the other spectral measures were taken to be indication of the learners' ability to discern subtle differences of /s/ in the two languages, resulting in their L2 realisation converging towards the native speakers' norms in an attempt to maintain separate phonetic categories for L1 and L2 /s/. Similarly, Quené et al. (2017) demonstrated that highly proficient Dutch–English bilinguals maintained a contrast in their production of /s/ in L2 English and L1 Dutch, in which /s/ is generally more retracted, articulated with a flatter tongue body and as such associated with a lower CoG. In Boyd (2018), both L1 French and L1 German speakers produced /s/ with

[1] Acoustic studies of apico- versus lamino-alveolar /s/ in other languages that do contrast these tongue gestures, such as Basque, have found that differences between the apico-alveolar /s̺/ and the lamino-alveolar /s̻/ can be captured by CoG, with the latter showing a higher CoG than the former (e.g., Beristain, 2021; Jaggers & Baese-Berk, 2019).

a higher CoG and lower skewness in their L2 English than in their own L1, with French gay speakers realising a more pronounced contrast across languages than their straight counterparts. Notably, here, the speakers' adoption of a more fronted variant in French (and German) than in English is in line with the traditionally claimed distinction drawn between the two languages.

Findings stemming from studies involving bilinguals other than late learners of English provide more mixed evidence. In the case of early Cantonese-English bilinguals, Johnson and Babel (2019) found no difference in the overall level and only minor differences in the trajectory of peak ERB_N between English /s/ and Cantonese /s/. When measured by spectral moments, sibilants in the two languages were also highly similar, although the results did not perfectly replicate the differences in peak ERB_N. English /s/ reported significantly higher CoG than Cantonese /s/, as well as a more convex trajectory. Skewness and kurtosis showed no cross-linguistic differences in the shape of the trajectory and only differences of extremely small magnitude in levels. While only five speakers were analysed in the study, the overall picture that emerges is that there may be fine-grained differences between English /s/ and Cantonese /s/, especially in their dynamic trajectories.

Schertz et al. (2019) investigated two bilingual Korean-Mandarin communities along the Chinese–North Korean border and found no differences in CoG between dento-alveolar fortis /s/ in Korean and dental /s/ in Mandarin Chinese. Intriguingly, Korean and Mandarin /s/ exhibited parallel fronting in apparent time, when only fronting in Mandarin was expected. The parallel change in Korean, then, was interpreted to be likely contact-induced, as a result of the assimilation of Korean fortis /s/ and Mandarin /s/ to a single category.

Further evidence for the role of acoustic similarity in cross-linguistic transfer comes from Beristain (2021), who studied contact between L2 Spanish, which has only one sibilant /s/, and three varieties of L1 Basque which variably merge the apico-alveolar /s̺/ and the lamino-alveolar /s̻/. Speakers of Basque varieties that either do not merge the two sibilants or merge them to the apico-alveolar /s̺/ were found to assimilate the Spanish /s/ to the apico-alveolar sibilant. In contrast, speakers of varieties that merged the Basque sibilants to the lamino-alveolar /s̻/ maintained an acoustic distinction between their Spanish /s/ from Basque /s̻/, suggesting that the sibilant fricatives in the two languages were sufficiently acoustically dissimilar for a separate category to be formed for the L2 sound.

3.1.5.3 Research question & hypothesis

The current study asks whether Canadian bilinguals of English and French distinguish their acoustic realisation of /s/ in the two languages. Earlier studies reviewed in the section above point to acoustic similarity as a potential main factor in whether speakers will establish distinct sound categories. As English /s/ and French /s/ have been found to be articulatorily and acoustically similar, it is hypothesised, following the (revised) Speech Learning Model (Flege & Bohn, 2021), that English–French bilinguals will not form separate phonetic categories for /s/ in the two languages and hence will not distinguish them in their production.

3.2 Long-term formant distributions

Consideration of LTFDs as a independent feature of interest originated not in the field of phonetics in general, but in the area of forensic speech science itself. Nolan and Grigoras (2005) first advance the case for their viability as acoustic-phonetic speaker discriminants, proposing an analysis of the whole collection of formant estimates from all voiced sounds in the sample. The peaks and the shapes of the distribution of each long-term formant (LTF) can thus be compared between samples to ascertain the similarity between them and their distinctiveness. Along with overall f0 and long-term average spectra (LTAS), LTFDs are among a bundle of features that aim to capture the overall characteristics of the speaker in a sample by extracting acoustic measurements not from individual sets of sounds but from across the entire sample.

Whereas f_0 conveys laryngeal information and LTAS encapsulates characteristics from both the laryngeal and supralaryngeal parts of the vocal tract, the formant-based LTFDs are primarily concerned with information from the supralaryngeal vocal tract. By considering the aggregate distribution of formant frequencies over whole speech samples, LTFDs are argued to not only reflect the physiology of the individual vocal tract, but also capture idiosyncratic behaviour in the speaker's overall articulatory habits (Nolan & Grigoras, 2005). Indeed, LTF1 means have been found to correlate with raised or lowered larynx, as rated in a vocal profile analysis, and higher LTF2 means have been shown to correlate with the setting of fronted tongue body, thus giving evidence in support of the relationship between LTFDs and idiosyncratic vocal settings (French et al., 2015). LTFDs have also been found to be independent of other commonly used features in voice comparison, such as f_0 and speech rate (Moos, 2008), making them a potentially useful set of additional features in the forensic speech analyst's toolkit. LTFDs are further considered to be advantageous over other suprasegmental features, f_0 and LTAS, as f_0 is subject to extensive within-speaker variation from a variety of sources over the course of speech, and LTAS can only convey peaks in the spectrum with a low level of precision due to averaging over numerous frames in which the position of the peaks constantly varies (Nolan & Grigoras, 2005).

LTFDs, which have become a mainstay in FVC analysis (Jessen, 2020), are nonetheless not immune to challenges posed by issues commonly found in forensic materials. In telephone speech, F1 is generally raised due to the narrow channel bandwidth and low frequency cut-off (Byrne & Foulkes, 2004; Künzel, 2001), implicating not only formant measurements of individual vowels but also LTFDs. Increased vocal speech in Lombard speech similarly results in significantly higher LTF1, but is not found to have a consistent effect on LTF2 or LTF3 (Jessen & Becker, 2010). LTFDs have further been found to be affected by style, whereby LTF1–3 all experience an upward shift in read speech when compared to spontaneous speech, although the differences across styles are numerically small for each formant: below 20 Hz for LTF1, and between 40 and 70 Hz for LTF2–3 (Moos, 2010).

Studies conducted mostly within the LR framework provide empirical corroboration for the discriminatory potential of LTFDs, with low error rates reported in both English and German (Becker et al., 2008; French et al., 2015; Gold et al., 2013b). Where the performance of individual LTFDs is concerned, higher formants (LTF3–4) generally outperform lower formants (Asadi et al., 2018; Gold et al., 2013b), providing evidence for the suggestion that higher formants encode more speaker-specific information than lower formants, which are said to be mainly responsible for encoding linguistic information. Speaker-specific information from multiple formants has been shown to combine effectively to enhance the discriminatory potential of LTFDs, as including additional formants (up to four in total) progressively reduces error rates progressively at each step (Becker et al., 2008; Gold et al., 2013b).

Even stronger performance of LTFDs can be obtained through including formant bandwidths as additional acoustic parameters (Becker et al., 2008), although improvements in performance may only be marginal (Hughes et al., 2017). Formant bandwidths, which are defined as the range of frequencies around the peak of each formant up to 3 dB below it, or where the power of the frequency has dropped to half of that of the peak (Kent & Read, 2002), are the consequences of acoustic energy loss, or damping, as sound passes through the vocal tract (K. N. Stevens, 1998). The more energy is absorbed, the wider the formant bandwidth is. Theoretically, the size of the bandwidth is the summation of the resistive contributions of each source of damping, which are primarily determined by the vowel formant frequency itself (K. N. Stevens, 1998, p. 136, 153). At low frequencies, the main sources of impedance come from the walls of the vocal tract and the configuration of the glottis, whereas at higher frequencies (above 2 kHz) radiation impedance of the mouth opening overtakes as the dominant contributor to formant bandwidth (p. 153, 258–259). The scope for individual variation in most resistive components contributing to formant bandwidths is relatively low, although it is suggested that some variability between speakers may be found in their glottal configuration and subglottal pressure (p. 260).

3.2.1 Cross-linguistic LTFDs

While the suprasegmental, holistic nature of LTFDs lends itself to high discriminatory power in the absence of language mismatch, their discriminatory potential in cross-language comparisons may be limited if LTFDs are language-dependent. To date, the effect of language on LTFDs has been rarely examined in phonetic studies and results obtained so far have been mixed. On the one hand, cross-linguistic comparisons of LTFDs from spontaneous telephone speech in German, Albanian and Russian have found LTF2 and LTF3 distributions to be comparable across languages (Jessen & Becker, 2010). On the other hand, studies examining intraspeaker variability in bilingual speakers have found them to display language-specific tendencies. Dutch–Turkish bilinguals, for example, do not produce LTF2 and LTF3 with significantly different means in each language, but the shape of their LTF2 distributions does exhibit cross-linguistic differences (Heeren et al., 2014). In a small-scale study of five Korean–English bilinguals, S. Cho and Munro (2017) found that these speakers largely maintained the shapes of LTFDs across languages, but produced LTF2 with peaks at lower frequencies in Korean than in English, following a pattern in the same direction as the vowels of the two languages. Along with Heeren et al. (2014), S. Cho and Munro (2017) similarly reach the conclusion that variability of LTFDs is lower within individual speakers across languages than between different speakers. With a view of establishing the potential of comparing LTFDs cross-linguistically for bilingual speakers, researchers have looked into other language pairs, including German and French (Krebs & Braun, 2015), and Serbian and English (Tomić & French, 2019). While these studies reported similar trends of minor differences in means but typically strong within-speaker correlations across languages, the emphasis on investigating central tendencies in past studies means that speaker-specific information in the shape of LTFDs has yet to be adequately investigated.

Related findings on phonetic settings also lend support to the idea that LTFDs may be subject to language-specific effects. Articulatory studies on inter-speech postures (ISPs) demonstrate language-specificity in the phonetic settings adopted by monolingual speakers of different languages (Gick et al., 2004), as well as bilinguals in each of their languages (Wilson & Gick, 2014). Similar tendencies have been found in LTAS of bilinguals (Ng et al., 2012), pointing to the differential use of voice quality according to the language spoken. The precise nature and origin of language-specificity in articulatory settings, however, whether as a consequence of the inventory and frequency patterns of sounds, or as otherwise learned targets, remain to be explored (Gick et al., 2004).

To date, Tomić and French (2019) represents the only attempt to explore the cross-language discriminatory potential of LTFDs for bilingual speakers within the LR framework, using a corpus of 35 female Serbian–English bilinguals. In particularly, known samples in L1 Serbian were compared with questioned samples in L2 English in a cross-validated procedure (see Section 6.1), with data from the background population also in Serbian. Cross-language comparisons were found to be highly unreliable, with poorly calibrated systems producing increased error rates and unable to capture useful speaker-specific information. Combining multiple formants in cross-language comparisons lowered error rates but resulted in increasingly worse system validity as measured in C_{llr} (see Section 6.1.1 for explanation of C_{llr}). However, a particular issue of this study is that output scores were not converted to LRs using training data appropriate for the specific case circumstances, but were themselves directly interpreted as LRs, meaning that biases in the score distributions from same- and different-speaker comparisons were not calibrated and, as a result, performance was not optimised for these parameters. While speakers in the study were found to produce language-specific LTFDs, the lack of calibration may have been a principal contributor to the poor performance of LTFDs in case of language mismatch. Another potential contributing factor to the poor performance of LTFDs lies in the chosen method of extracting and modelling LTFDs for comparison (see Section 6.2.2), which likely resulted in substantial loss of useful data. The current study aims to address these methodological shortcomings by employing a larger database (of 60 speakers) and performing score calibration using appropriate case-specific training data.

3.2.2 Current study

As the current study seeks to assess the cross-linguistic use of LTFDs in FVC, it is necessary to first establish the effect of language on LTFDs among Canadian English–French bilinguals. The literature reviewed above suggests that LTFDs produced by bilinguals are expected to show distinct, language-specific patterns. It is therefore predicted that language will have an effect on the long-term distributions of formant centre frequencies. Specifically, it is predicted that French will have overall higher LTF2 than English, as French has a more crowded front vowel space consisting of pairs of unrounded (/i e ɛ/) and rounded front vowels (/y ø œ/) when compared to English.

Although cross-linguistic comparisons of overall formant bandwidth distributions are rarely conducted, their language-specificity is also investigated in the current study as forensic research and work frequently make use of this set of acoustic parameters. Language is also expected to have an effect on formant bandwidths, as French has a set of nasal vowels in its sound inventory (/ɛ̃ œ̃ ɑ̃ ɔ̃/), which typically have wider bandwidths than oral vowels, due to increased acoustic energy loss (particularly at low frequencies) from wall impedance as a consequence of the introduction of the nasal cavity (K. N. Stevens, 1998, p. 193). As English does not have phonemic contrast by nasality, the prediction follows then that overall formant bandwidths should be higher in French than in English, although differences between the two languages may be mitigated by the fact that vowels in English regularly undergo allophonic nasalisation when they are found before coda nasals (see, e.g., Chen, 1997; Cohn, 1993; Krämer, 2019).

Chapter 4

Methodology: Acoustic Phonetics

This chapter outlines the methods used in acoustic analysis to address the questions raised in the previous chapter regarding the production of /s/ and LTFDs in bilinguals. Sections 4.1 and 4.2 first describe the corpus that is used throughout the study. The procedure for data preparation and analysis for the two variables is detailed in Sections 4.3–4.5, followed by a report on the results of preliminary testing conducted for LTFDs in Section 4.6.

4.1 Materials

The current study used recordings from the Voice ID Database (henceforth the Database; RCMP, 2016), an audio corpus collected by the Audio and Video Analysis Unit of the Royal Canadian Mounted Police (RCMP) at the University of Ottawa. Samples were collected in English and French, the two official languages of Canada. The Database predominantly consists of Canadian-born and raised speakers, but also includes speakers from a wide range of regional and linguistic backgrounds. Crucially, for the purposes of this study, a substantial proportion of participants (46%) who speak both English and French were recorded in both languages. The version of the Database used here contains recordings from 927 speakers in total, although data collection has continued and the Database has since expanded.

For each language that a speaker contributed to, they were recorded four times, each on a different channel (in no particular order): high-quality microphone, GSM-transmitted mobile, landline telephone and covert room bug (Kavanagh, 2014). The current study focused only on recordings obtained from the high-quality microphone condition, which was recorded on a Marantz PMD 670 recorder (44.1 kHz, 16 bit, stereo) via a microphone placed approximately 15 cm from the speaker's mouth (Kavanagh, 2014). In each recording, speakers read a list of 20 phonetically balanced short sentences and, for recordings that took place from 2012 onwards, an additional phonetically balanced passage. Of the English materials, the sentences were extracted from the Harvard Sentences (IEEE, 1969), and the passage was an abridged version of *The Rainbow Passage* (Fairbanks, 1969). The French sentences were created by combining shorter phonetically balanced sentences from Vaillancourt et al. (2005), while the passage was *La bise et le soleil* (*The North Wind and the Sun*). An orthographic and phonemic transcription of all the read materials can be found in Appendix A.

It must be acknowledged that the read nature of the materials and the high quality of the recordings chosen are not representative of typical, forensically realistic conditions, and may lead to measures of performance that are more optimistic than those obtained from spontaneous speech and the other poorer quality conditions. However, the primary focus of the present study is on the effects of language itself. The controlled nature of the materials thus offers its advantages. Keeping the speech materials uniform provides maximal comparability between different speakers, ensuring that variations in acoustic measurements and in speaker-discriminatory performance can be attributed to language, speaker physiology and behaviour, rather than differences in speech content. This is especially the case for LTFDs, as long-term measures depend not only on the inventory of sounds and their phonetic implementation, but also on the frequency of occurrence of each sound within each language and each sample (Mennen et al., 2010). As this is one of the first projects exploring the effects of language and bilingualism on the strength of evidence in FVC, results from high-quality data can be of considerable benchmark value to future work evaluating the effects of language.

4.2 Speakers

60 adult male Canadian English–French bilinguals were selected from the Database, applying a simple set of selection criteria based on available metadata. Metadata on language background were limited, including only the languages spoken, the age the speaker was first exposed to their L2 (0 in the case of simultaneous bilinguals), and whether English and/or French was their mother tongue. A small number of speakers reported their level of proficiency in languages other than English and French, or their dominant language, but such data were not systematically collected. As such, bilingual speakers in this study included those who participated in recording for both languages and who did not report knowledge of any other languages. The latter criterion was applied to exclude influence from languages other than English and French.

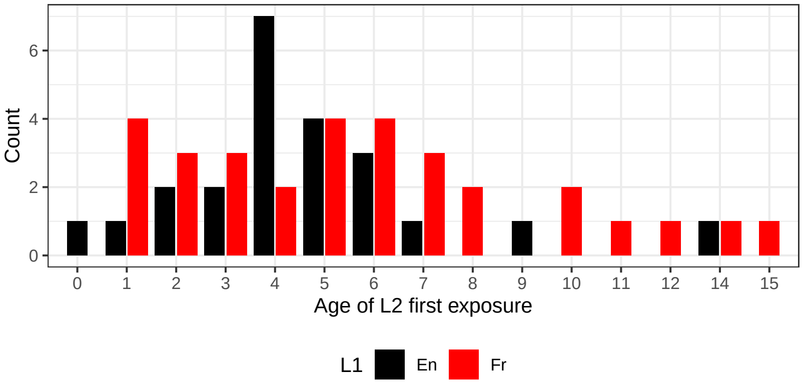

Figure 4.1: Distribution of age of first exposure to L2 among self-reported sequential bilinguals.

Out of all 60 speakers, 23 (38%) and 31 (52%) reported English and French as their L1 respectively, while the remaining 6 (10%) reported to have acquired both languages simultaneously. Among the self-identified sequential bilinguals, a wide range was reported for the age of first exposure to L2, from before the age of 1 to 14 for L1 English speakers and from 1 to 15 for L1 French speakers. Figure 4.1 shows the distribution. While the age of first exposure to L2 does not imply regular L2 input from that point, it is extremely likely that this group consists of both early and late bilinguals. In terms of age distribution (median = 23, mean = 27.7, SD = 12.4), the group comprised mostly young speakers below the age of 28 (46 out of 60), although it also included a small number of older speakers aged between 33 and 71.

The present sample clearly cannot be, and is not intended to be, representative of the whole linguistic landscape in Canada, which is highly diverse in itself, be it on the national, provincial/territorial or lower level. Nevertheless, the mix of linguistic backgrounds in this set of speakers can be considered to be a reasonable approximate of the English–French bilingual community around Ottawa (where the corpus is collected), a highly bilingual area sitting on the border of English-speaking Ontario and French-speaking Quebec (Figure 4.2). As of 2016, 17.9% of the Canadian population are bilingual in English and French (Statistics Canada, 2017b). 56% of the population report English to be their L1 while the proportion of L1 French speakers is lower at 20.6%. Officially, only 0.5% of the population report to be simultaneous bilingual speakers of English and French.[1] Within the English–French bilingual population, L1 French speakers constitute the majority, at 53.2% (Statistics Canada, 2017c). The level of bilingualism is highest in

[1] There is reason to believe that the proportion of simultaneous bilinguals may have been underestimated. The relevant question is formulated as "What is the language that this person first learned at home in childhood and still understands?" (Statistics Canada, 2017e), and an individual is understood to have two L1s only if "the two languages were used equally often" (Statistics Canada, 2017d).

the provinces of Quebec (44.5%) and neighbouring New Brunswick (33.9%), while Ontario, with 69.5% L1 English speakers and 4.3% L1 French speakers, only has 11.2% of its population bilingual in the two languages. In the metropolitan area of Ottawa-Gatineau,[2] the proportion of English–French bilinguals stands at 44.8%. L1 English and L1 French speakers amount to approximately half (51.4%) and one-third (32.5%) of the speakers in the area.

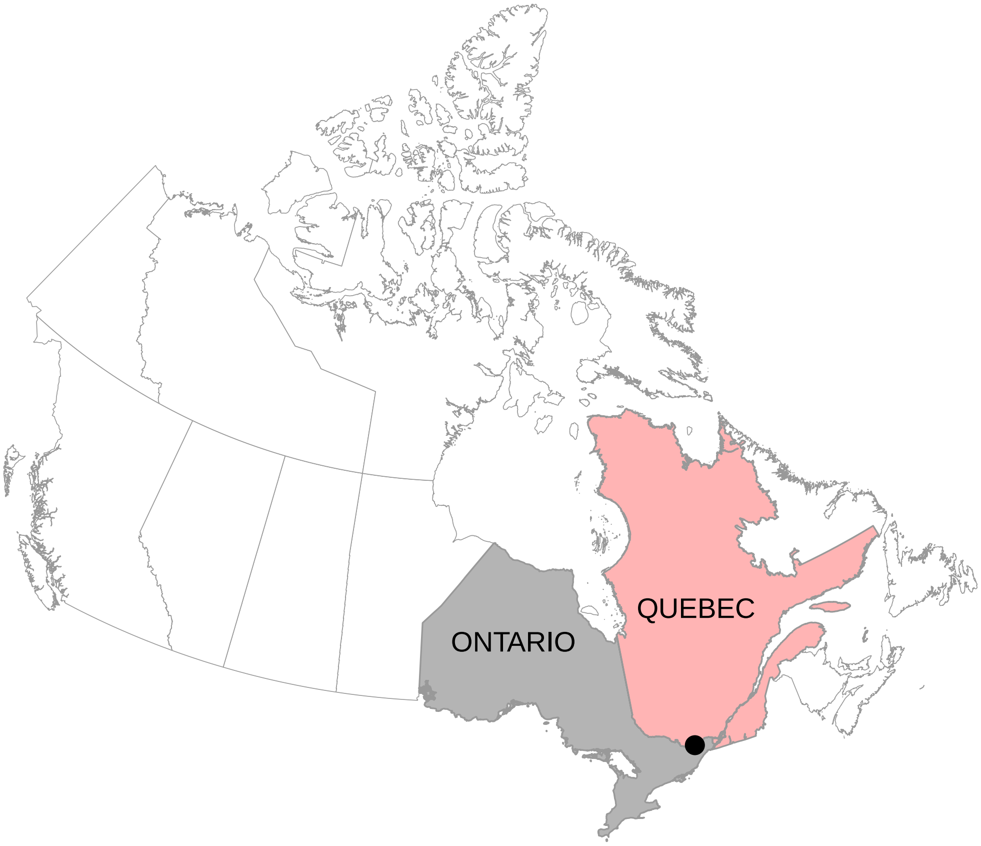

Figure 4.2: Map of Canada showing location of Ottawa. Produced using boundary data from Statistics Canada (2017a).

[2] The metropolitan area of Ottawa-Gatineau straddles both provinces of Ontario and Quebec and comprises the two cities of Ottawa and Gatineau, which face each other on the banks of the Ottawa River, and associated suburbs.

4.3 Data preparation

All recordings were first orthographically transcribed in Praat (Boersma & Weenink, 2016). Hesitations, repetitions, mispronunciations and deviations from the set material were retained as far as possible, although partial words were excluded. Automatic segmentation was performed using the Montreal Forced Aligner (MFA; M. McAuliffe et al., 2017), using an acoustic model trained from scratch on the recordings and preconstructed English[3] and Quebecois French[4] pronunciation dictionaries from Prosodylab-aligner (Gorman et al., 2011), modified to include out-of-dictionary words in the materials and alternative pronunciations. The phonetic transcription obtained was then checked for errors in the alignment and manually corrected where the forced alignment was clearly erroneous, with a particular focus on /s/, vowels and glides, as these are the phonetic variables of interest in the current study. Boundaries of sibilants were identified at the onset and offset of aperiodic noise. The demarcation of vowels was guided by the presence of a clear formant structure, and midpoints of formant transitions were used to indicate boundaries between vowels and adjacent glides and liquids (/j w l r/ in both languages and /ɥ/ in French). As all time-aligned intervals by MFA had a precision level of 10 ms, any manually corrected or created boundaries were placed in accordance with the same level of precision. While this means that boundaries may be out of line with the precise location of suitable acoustic landmarks by up to 5 ms, maintaining consistency over the level of precision ensures that no segments were disadvantaged against others in subsequent analysis.

4.4 Data extraction & analysis: /s/

4.4.1 Acoustic parameters

A number of parameters have been proposed to capture the acoustic characteristics of fricative consonants. These include measures of duration (Strevens, 1960), intensity (Koenig et al., 2013), formant transitions (Soli, 1981), as well as a wide range of spectral measures, such as spectral peak (Strevens, 1960), spectral slope or spectral tilt (Jesus & Shadle, 2002), cutoff frequency (Stuart-Smith et al., 2003) and discrete cosine transform coefficients (Jannedy & Weirich, 2017; Watson & Harrington, 1999). The current study focuses on the first four spectral moments (Forrest et al., 1988), which are obtained by treating the frequency spectrum of the sound as a random probability distribution and

[3] https://github.com/prosodylab/Prosodylab-Aligner/blob/master/eng.dict

[4] https://github.com/prosodylab/prosodylab-alignermodels/blob/master/FrenchQuEu/fr-QuEu.dict

which have become the most commonly used parameters in analysing the acoustic properties of /s/ and other fricatives.

The first spectral moment is the spectral centre of gravity (CoG). It refers to the mean frequency of the spectrum, on either side of which sound energy is equally distributed. A higher CoG indicates that a higher proportion of sound energy is concentrated in the high frequency range and, like spectral peak, is inversely correlated with the size of the cavity in front of the point of constriction. Higher CoG is thus generally interpreted as a more fronted articulation, all else being equal.

The second spectral moment, which refers to the variance, measures the spread of acoustic energy across the spectrum. The far more common practice, however, is for the second moment to refer to the square root of the variance, namely standard deviation (SD), as variance carries a unit of Hz^2 and is not as readily interpretable as SD. Higher SD corresponds to a more diffuse frequency spectrum, which has been interpreted as an acoustic correlate of laminality (F. Li et al., 2009).

The third spectral moment, skewness, is a dimensionless quantity that measures the asymmetry of the spectrum. A skewness of 0 indicates a symmetrical distribution of energy around the mean. Positive skewness means that the distribution is right-skewed, with a long tail in the positive direction, and vice versa. Like CoG, skewness has also been linked with place of articulation, where /s/, with a bulk of sound energy in the higher frequencies and thus longer tail in the lower frequencies, has a more negative skewness than /ʃ/, which has a more backed place of articulation (Forrest et al., 1988; Jongman et al., 2000).

The fourth spectral moment, kurtosis, is also dimensionless, although a straightforward phonetic or even spectral interpretation is not forthcoming. While it has often been described as a means of characterising the “peakedness” of the spectrum, Westfall (2014) contends that the longstanding interpretation is in fact erroneous, arguing instead that kurtosis relates to tail extremity, or the existence of outliers. A normal distribution has a kurtosis of 3, which is often shifted and reported as an (excess) kurtosis of 0 for convenience of interpretation. A leptokurtic distribution, with a kurtosis greater than 3 (i.e., positive excess kurtosis), tends to have outliers outside the central portion of the distribution curve, whereas a platykurtic distribution, with a kurtosis between 0 and 3 (kurtosis cannot be negative), or negative excess kurtosis, lacks such outliers. As kurtosis is often correlated with SD (Harrington, 2010), the kurtosis of a fricative has been generally referred to in the literature as a correlate of sibilance (Kitikanan et al., 2015) or more specifically as another correlate of apicality/laminality (F. Li et al., 2009). By default, kurtosis reported in Praat is in fact excess kurtosis. In the current study, all measurements of kurtosis were thus converted to kurtosis by adding 3 to each value, and “kurtosis” in this study refers to the classic definition and not “excess kurtosis” throughout.

4.4.2 Inclusion criteria

/s/ in all phonological contexts were in principle included in the present study, but contexts were excluded from consideration where the segmentation of /s/ was rendered unreliable due to the absence of well-defined acoustic cues between the target /s/ and its neighbouring segments. These include:

1. /s#s/

   Successive occurrences of /s/ across word or morpheme boundaries are often pronounced as a single fricative event in natural continuous speech, as exemplified in the words *makes sweet* in the top panel of Figure 4.3. Segmentation is problematic when there are no two separate fricative events, as there is no clear place for a reliable boundary. The fricative noise cannot be treated as a single phoneme, as frication from underlying double /s/ may indeed be distinguished from a single /s/ by increased duration (Klatt, 1974). Conversely, where a period of silence is present to allow demarcation of two distinct fricative events, as illustrated in the bottom panel of Figure 4.3, the relevant tokens were both retained.

2. /s/ adjacent to /z/

   Sequences of /s/ and /z/ pose challenges to segmentation as they are homorganic fricatives and share very similar spectral properties. The occurrence of voicing assimilation, which results in /s/ becoming voiced or /z/ becoming devoiced, either partially or completely, means that the conventional acoustic cue of voicing can no longer serve to distinguish the two in a continuous segment of frication. Regressive voicing assimilation is well documented in French (Abdelli-Beruh, 2012; Hallé & Adda-Decker, 2011), though some progressive voicing assimilation has also been observed (Niebuhr et al., 2011). While regressive voicing assimilation is also prevalent in English, the process is limited to devoicing only and may indeed be speaker-specific (Myers, 2010; Niebuhr et al., 2011).

3. /s/ adjacent to /ʃ ʒ/

   The dento-alveolar [s] and the postalveolar [ʃ ʒ] are normally distinguishable spectrally, as sound energy in the latter is concentrated in a lower frequency region than that in the former. However, place assimilation in sequences of /s/ and /ʃ/ is

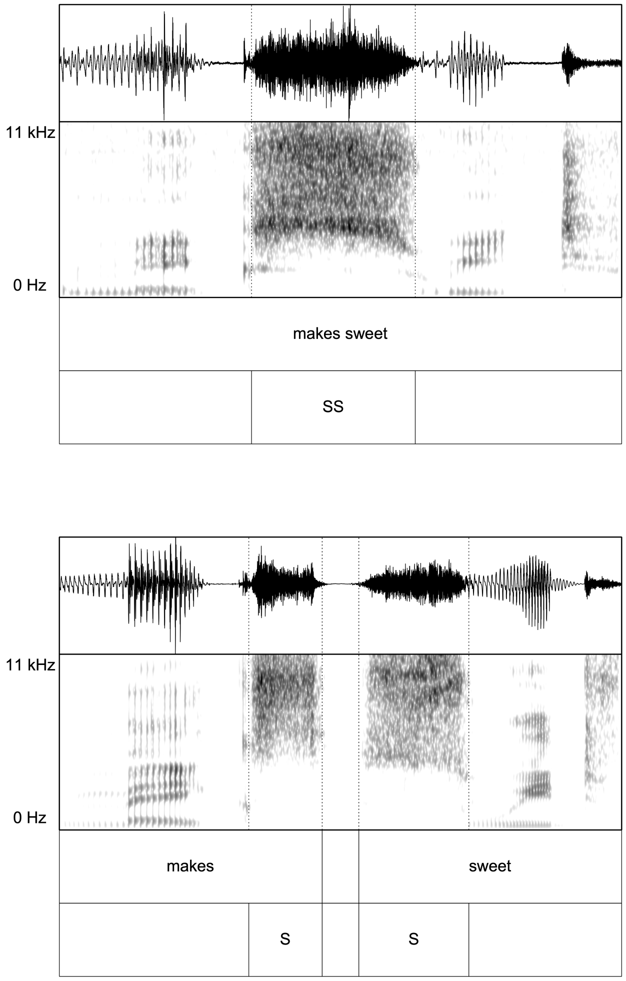

Figure 4.3: Two instances of *makes sweet* (top) by speaker 436 with single continuous fricative event and (bottom) by speaker 385 with separate fricative events.

well known to take place in English and has also been shown to occur in French (Niebuhr et al., 2011). When assimilation occurs, the whole sequence is realised as a noise transitioning between [s]-like and [ʃ]-like properties, in the case of gradual assimilation, or as a stable [ʃ]-like noise, in the case of complete assimilation (Holst & Nolan, 1995). Place assimilation in English is strictly regressive and primarily towards the postalveolar, whereas assimilation in French may occur in both regressive and progressive directions, but only /s/ assimilates towards /ʃ/ and not the other way round (Niebuhr et al., 2011). In addition, voicing assimilation can independently co-occur with place assimilation, so that /s/ may also assimilate towards the voiced postalveolar /ʒ/ (in French, as this phoneme is not found word-initially in English except in loanwords).

Table 4.1 summarises the contexts in English and French which were excluded as a result of the criteria above. In total, 2,366 tokens in English and 2,725 tokens in French were included.

Table 4.1: Contexts in English and French excluded from analysis. /sʒ/ placed in parentheses in English as possible but not naturally occurring context.

	/ss/	/sz/	/zs/	/sʃ/	/sʒ/	/ʃs/	/ʒs/
English	+	+	+	+	(+)		
French	+	+	+	+	+	+	+

4.4.3 Spectral analysis & coding

Both static and dynamic measurements (see below) were extracted from each eligible token of /s/ in a procedure automated with Praat scripts. In each case, any residual voicing was first removed with a Hanning bandpass filter between 500 and 11000 Hz. Fast Fourier Transform was then applied to windowed intervals of each segment to compute frequency spectra, from which the four spectral moments were extracted. The duration of each fricative was also extracted. Tokens with CoG $<$ 2500 Hz were deemed to be too low for /s/ and consequently removed. Tokens with a duration of 50 ms or below were also excluded to prevent overlap of multiple (3+) windows in dynamic measurements. In total, 187 tokens (7.9%) in English and 119 tokens (4.4%) in French were removed, leaving 2,179 and 2,606 tokens respectively for subsequent analysis.

Each remaining token of /s/ was coded for its preceding and following phonetic environments. Preceding phonetic environment was coded in four categories: pause, consonants, rounded vowels and unrounded vowels. Following phonetic environment was similarly coded, with the addition of /str/ as a category in English separate from other consonants, due to possible /s/-retraction in this environment documented in the literature. Vowels and glides coded as rounded in English include /o ɔ u ʊ w/[5] and the diphthong /aʊ/ when it precedes /s/. Here /ɔ/ represents the low back vowels for cot and caught, which are stably merged (Boberg, 2011; Clarke et al., 1995) as a rounded vowel in Canadian English, rather than as an unrounded vowel in other similarly merging varieties of North American English, such as California English (Hagiwara, 2006). In French, vowels and glides coded as rounded include the oral vowels /y ø œ u o ɔ/, the nasal vowels /œ̃ ɔ̃/ and the glides /ɥ w/. Particularly of note is the nasal vowel /œ̃/,

[5] Following the North American tradition, vowels in words of the lexical sets face and goat are transcribed as /e/ and /o/ in this study.

which is merged with /ɛ̃/ in Metropolitan French but remains a separate phoneme in Canadian French (Walker, 1984).

4.4.3.1 Static measurements

Static measurements for each token were taken from a single 40-ms Kaiser2 window centred at the midpoint of the fricative (see Figure 4.4 for an example), following the window size used in Jongman et al. (2000) and Kavanagh (2012). This window size is considered sufficiently large to provide good resolution in the frequency domain for analysing the whole segment. While some researchers have opted for wider windows of up to 100 ms (e.g., Wrench, 1995), doing so would impose a higher threshold on the duration of the segment, incurring the trade-off of restricting shorter tokens from analysis. Of the window functions available in Praat, Kaiser2 was chosen as it has been evaluated as one of the top-performing windows in reducing spectral leakage (Harris, 1978).

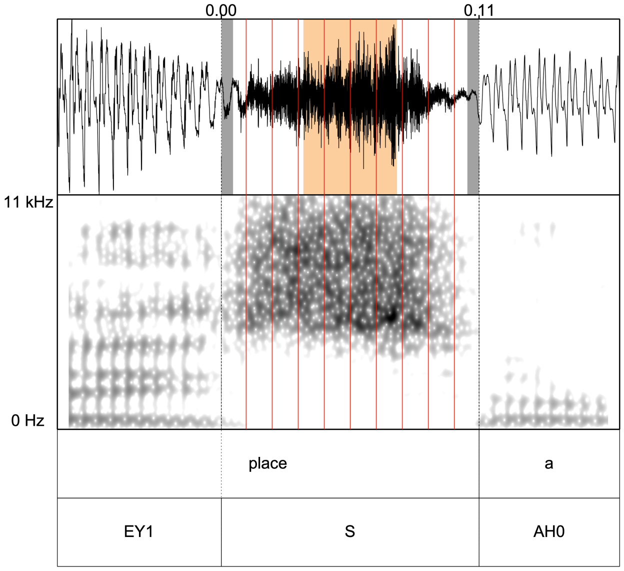

Figure 4.4: Waveform, spectrogram and TextGrid of the words *place a* with orthographic and ARPAbet phonemic transcription, showing boundary placement and window locations for /s/. 5 ms excluded from each end shaded in grey. 40-ms static window shaded in orange. 10-ms dynamic windows centred at each red line.

4.4.3.2 Dynamic measurements

For dynamic analysis, nine sets of spectral moments were extracted from nine 10-ms Kaiser2 windows, evenly spread out across the duration of the token, excluding 5 ms from each end to counter potential misalignment due to the precision level of the aligner mentioned above (also illustrated in Figure 4.4). Narrower windows were used to increase temporal resolution, at the expense of some frequency resolution, so that successive windows could represent different portions of the segment and not considerably overlap, which could lead to overrepresentation of highly overlapped regions and underrepresentation of spectral movement, especially in the case of shorter tokens.

4.4.4 Statistical analysis

Both static and dynamic measurements were analysed in R (R Core Team, 2018) to evaluate any cross-linguistic differences in /s/ acoustics. Static measurements were analysed with linear mixed effects models (LMEMs) via the *lme4* package (Bates et al., 2015), while dynamic measurements were analysed using generalised additive mixed models (GAMMs; Sóskuthy, 2017; Wood, 2017).

Static measurements from each spectral moment obtained from all 60 speakers were fitted with a separate LMEM. In the case of kurtosis, measurements were first log-transformed to correct for the heavily positive skew. Language spoken (English vs. French) and speaker L1 (English, French and both), as well as their interaction, were included in the initial full model as fixed predictors to assess their effects on the realisation of /s/. To control for the effects of phonetic factors on /s/ production, preceding and following environments and duration were also included as fixed effects. Random intercepts were included for the factors of speaker and word, and language-by-speaker random slopes were included to account for any variation between speakers in how they respond to language shift. Weighted effect coding (te Grotenhuis et al., 2017) was adopted for all fixed predictors, so that effects can be interpreted with respect to the grand mean of all observations rather than to particular reference levels. An advantage of weighted effect coding over traditional (unweighted) effect coding is that it can account for any imbalance in the number of observations across different categories.

The significance of fixed effects was determined by a series of likelihood ratio tests (LRTs) in a step-down approach (implemented via ANOVAs in R), whereby each predictor was dropped in turn to form a reduced model that was compared to the full model, and predictors were only retained if their inclusion led to a significantly improved fit at $\alpha = .05$. Only the best-fitting model for each spectral moment is reported in Chapter 5.

The trajectory of spectral moments over the course of /s/ was analysed by means of GAMMs, which are useful for modelling non-linear relationships without having to stipulate a fixed complexity of the curve fitted to the data. Further, through the incorporation of random smooths, GAMMs allow for the modelling of non-linear random effects.

All models were fitted in R with the bam function from the *mgcv* package (Wood, 2017). For each spectral moment as the outcome variable, a set of GAMMs were fitted over the set of nine measurements taken from each token of /s/. Parametric and smooth terms (denoted by s()), which respectively model overall differences in level (or height) and shape, were included for the main variables of interest, language and speaker L1. Preceding and following phonetic contexts were similarly included to control for their effect on the overall trajectory. Following Sóskuthy (2017), they were coded as ordered factors, in order for their effects to be modelled with difference smooths, meaning that a smooth was fitted to a reference level (e.g., Language = English) and further smooths represented the non-linear difference with other levels in the predictor variable (e.g., between French and English). Duration was also included as a predictor, and any effect on the shape of the trajectory was modelled using the class of smooths known as tensor product interactions (ti()). As in the LMEMs above, speaker and word were included as random effects. Random intercepts were included for word, while factor random smooths were used to model the non-linear effect of speaker on /s/ over its duration. An AR1 model was also included to reduce residual autocorrelation within the trajectory of each token. To summarise, the full model for each spectral moment can thus be expressed with the following formula:

```
moment ~ s(Window) +
            Language + s(Window, by = Language) +
            SpeakerL1 + s(Window, by = SpeakerL1) +
            PrevContext + s(Window, by = PrevContext) +
            FollContext + s(Window, by = FollContext) +
            s(Duration) + ti(Window, Duration) +
            s(Word, bs = "re")
            s(Window, Speaker, bs = "fs")
```

Likelihood ratio-based model comparison was performed using the compareML function from the *itsadug* package (van Rij et al., 2017), to test the significance of language and speaker L1 (see Sóskuthy, 2021, for an evaluation of different methods of significance testing). All models thus had to be fitted using maximum likelihood estimation (ML), rather than restricted ML or fast restricted ML. Following Sóskuthy et al. (2018), a two-step process of model comparison was employed to avoid false positives from evaluating the parametric and smooth terms separately. The overall effect of each variable was first tested by comparing the full model with a nested model that excluded both parametric and smooth terms for the variable (e.g., Language + s(Window, by

`= Language))` from the full model. A second comparison, specifically testing the significance of the effects on shape, was only conducted in the presence of a significant overall effect, by excluding the smooth term but retaining the parametric term.

4.5 Data extraction & analysis: LTFDs

4.5.1 Inclusion criteria

There is as yet no agreed set of criteria for sounds that are included or excluded in the calculation of LTFDs. Nolan and Grigoras (2005), who first proposed the use of LTFDs, extracted measurements from all voiced frames. Other researchers subsequently reported different inclusion criteria: Becker et al. (2008) removed all consonants and portions with unclear formant structure; Moos (2010) visually selected the vocalic stream based on formant structure, and as such retained laterals and approximants but excluded nasals or sounds released with strong nasality; Tomić and French (2019) similarly used only vowels with clear formant structure; Gold et al. (2013b) and Hughes et al. (2017) included only sounds automatically identified as vowels. Previous studies have relied on automatic extraction rather than identification of relevant segments. In addition to drastically reducing the time needed for analysis, the automatability of LTFD extraction has been claimed to offer another practical advantage, namely to enable analysts who do not speak the language of the materials to perform forensic analysis (Becker et al., 2008), despite strong caution against analysis by non-native speakers in the IAFPA Code of Practice (IAFPA, 2020).

While the stability of LTFDs under these various criteria is yet to be addressed in research, the inclusion or exclusion of certain classes of sounds in their derivation would predictably have an effect on the resultant distributions. In languages like Mandarin Chinese, which makes heavy use of rhotics and rhoticised vowels, excluding them would remove their lowering effect on F3 and likely lead to higher LTF3. In a similar vein, excluding sounds with strong nasality from languages like French and Portuguese would mean that only oral vowels and not nasal vowels are represented in the resultant LTFDs. As nasal vowels are known to widen formant bandwidths when compared to oral vowels (K. N. Stevens, 1998), one potential consequence of excluding nasal vowels is a higher concentration of lower formant bandwidths. Indeed, a brief comparison of LTFDs obtained from the present data spanning (i) all vowels, (ii) oral vowels only, and (iii) nasal vowels only demonstrates precisely such an effect. In Figure 4.5, the distributions of F1 and F2 bandwidth (BW1, BW2) for nasal vowels are both shown to be higher than those

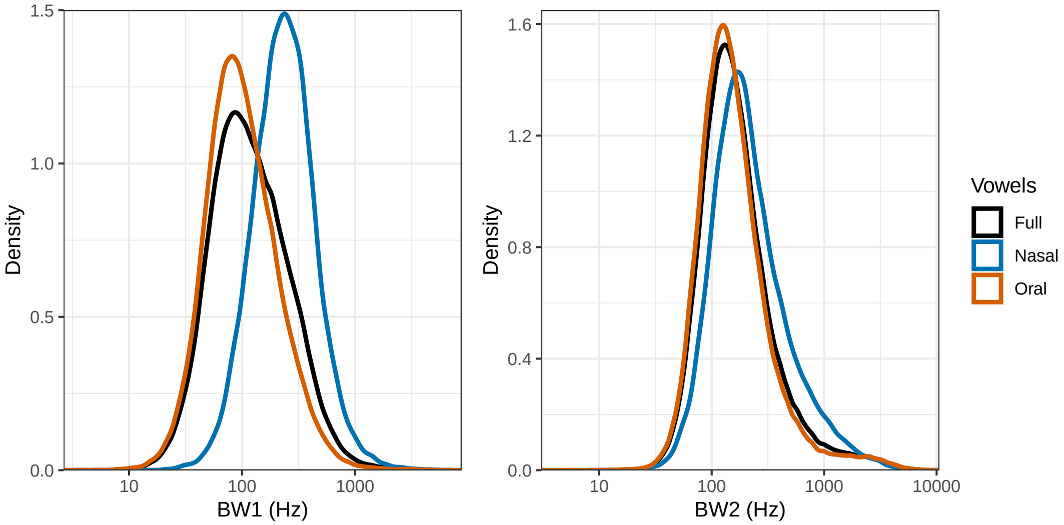

Figure 4.5: Overall BW1 and BW2 distributions of all vowels, oral vowels only and nasal vowels only in French, pooled across all 60 speakers.

for oral vowels. Including both oral and nasal vowels in the calculation of LTFDs thus results in less sharply peaked overall BW1 and BW2 distributions, with a heavier tail in the positive direction, than including oral vowels only, although the size of effect is small due to the comparatively low frequency of nasal vowels.

As the aim of the present study is to examine the effect of language on LTFDs, segments were chosen to reflect the sound inventories of each language. As a result, all vowels were included, including nasal vowels in French. Glides were also included due to their acoustic similarity with vowels. Other consonants, such as nasals and approximants, were not included. Table 4.2 summarises the phonemes included for formant extraction in each language. In total, an average of 26.2s of vocalic materials in English and 34.6s of vocalic materials in French were identified per speaker.

Table 4.2: Target phonemes for phoneme extraction in English and French.

Language	Vowels	Glides
English	i ɪ e ɛ æ ə ɑ ʌ ɔ o ʊ u aɪ aʊ ɔɪ	j w
French	Oral: i y e ø ɛ œ ə a o u	j w ɥ
	Nasal: ɛ̃ ɑ̃ œ̃ ɔ̃	

4.5.2 Formant extraction

Formant centre frequencies and bandwidths for the first four formants were extracted in Praat from the onset to the offset of all instances of eligible segments at intervals of

10 ms. Contrary to Moos (2010), an edited vocalic stream was not used to extract formant estimates. Instead, vowel formants and bandwidths were extracted *in situ* from the recordings themselves. Formant extraction was automated with a custom Praat script, using the Burg algorithm for linear predictive coding, with the formant tracker set to search for 6 formants up to a maximum formant frequency of 5500 Hz in frames of 25 ms. These settings were determined by preliminary testing, reported in Section 4.6, and remained fixed for all speakers. While specifying different formant settings for each speaker and indeed for each token would result in the most accurate measurements for each individual vowel (Harrison, 2013), it is considered that LTFDs are devised as a semi-automatic linguistic-phonetic variable and involve detecting how often each frequency is estimated to be a formant, as chosen by the formant tracker (Nolan & Grigoras, 2005). Given the large amount of data involved in the analysis of LTFDs, tailoring formant settings to specific speakers and tokens is likely not often practicable. Indeed, the analysis in Nolan and Grigoras (2005) focuses simply on the frequencies “chosen as the estimate of a formant by a linear prediction formant tracker” (p. 162) in voiced frames.

4.5.3 Statistical analysis

As in the case of /s/ spectral moments, LMEMs were used to analyse the effect of language on LTFDs. A separate model was fitted to all data points from each LTFD and BW, with language spoken, speaker L1 and their interaction included as fixed effects. Random by-speaker intercepts and language-by-speaker slopes were included to model individual speaker variation. Significance testing was again conducted by means of likelihood ratio tests in a step-down approach. To correct for skew in the data, LTF1 and all bandwidth distributions BW1–4 were log-transformed before modelling. It should be noted that, as multiple sets of data points were extracted from the signal in close proximity, the degree to which data points from the time series were independent of one another would be variable and the assumption of independence in LMEMs might be slightly violated. Additionally, the inclusion of vowel categories or phonological features as predictors in the model would have helped control some of the variation in the formant estimates. However, as the data were phonetically balanced and, as mentioned above, the common forensic practice is not to segment by individual vowel, findings from the current approach are considered to be more informative for current forensic methodologies.

4.6 Preliminary testing: LTFDs

This section reports on the preliminary testing conducted to examine the reliability of different settings in Praat across the whole range of vocalic sounds, in the present corpus of Canadian English and French. Following the procedure described in Section 4.5.2, the maximum number of formants (5 vs 6) and the maximum formant frequency (5000 vs 5500 Hz) were varied to test four sets of settings in total (hereafter abbreviated as 5-5000, 5-5500, 6-5000 and 6-5500). The window length was fixed at 25 ms in all cases. Within each setting, F1–F4 centre frequencies and bandwidths were extracted at intervals of 10 ms from all vowels and coded by the source phoneme for further analysis. In this part of preliminary testing, nasal vowels were also included in French. Estimates at the boundary of successive vowels were considered to have come from the earlier vowel. Only complete sets of estimates, where Praat did not return N/A on any formant or bandwidth estimate, were retained, but no further exclusionary criteria were applied to discard any data.

4.6.1 Overall formant centre frequencies

As can be observed in Figure 4.6, which displays LTFDs from all settings pooled across all speakers, the choice of setting had negligible impact on F1. With the exception of the setting 5-5500, there is a virtually complete overlap of the LTF1 distributions. While the peak from the setting 5-5500 is very similar to the others in terms of frequency, the heavier tail above 700 Hz found in both languages indicates a greater proportion of high formant estimates.

The effects of formant settings were more apparent on the distributions of LTF2. While peaks were located at similar frequencies regardless of setting, there was evidence of a shift toward higher frequencies from the setting 6-5000, through 6-5500 and 5-5000, to 5-5500, as the distribution became more negatively skewed.

The impact of setting on LTF3 and LTF4, following a similar trend to that in LTF2, was much more considerable than that on the lower formants. The setting 6-5000 showed a more negatively skewed distribution, with a peak at the lowest frequency. In the order of 6-5500, 5-5000 and 5-5500, the LTFD peaks shifted towards higher frequencies, by around 200 Hz and 400 Hz in LTF3 and LTF4 respectively. At the same time, the distributions became less negatively skewed, indicating the presence of a lower proportion of low-frequency estimates alongside the general upward shift.

Another observation from Figure 4.6 is the significant overlap between LTF3 and LTF4 derived from different settings. In French, the setting 6-5000 displays a bimodal

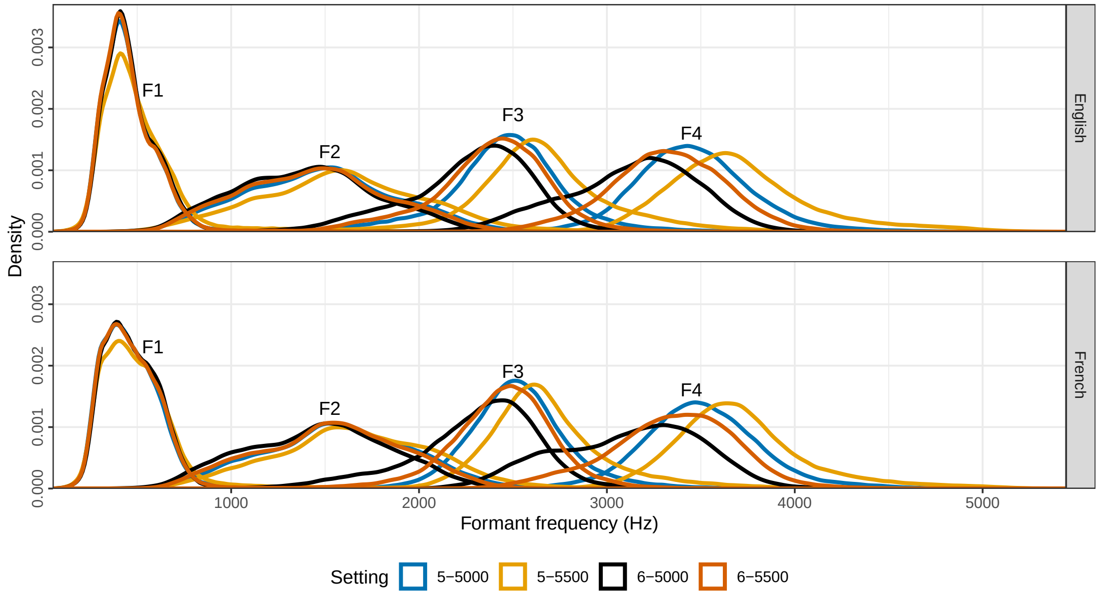

Figure 4.6: Overall LTF1–4 distributions in English and French by formant setting.

LTF4 distribution, with a secondary peak at around 2600 Hz, coinciding with the peak of the LTF3 distribution from the setting 5-5500.

Overall, these results suggest that the effect of formant settings were consistent in direction across formants, but were generally much smaller on lower formants than higher formants. By attempting to detect the highest number of formants in the narrowest frequency range, the setting 6-5000 recorded the lowest F2–F4 estimates when compared to other settings. This was followed by the other setting that searched for the same number of formants in a wider frequency range (6-5500). The five-formant settings followed, with the one using the wider frequency range (5-5500) consistently giving estimates with the highest frequencies.

Two particular points of concern can be raised at this juncture. The substantial proportion of low estimates of F4, and of F3 to some extent, from the setting 6-5000 suggests the possibility of formant overfitting, where the formant tracker sought to fit a greater number of formants than were present in the indicated frequency range. Similarly, the abrupt shift of a significant proportion of F2 estimates from below 1200 Hz to above 2000 Hz in the setting 5-5500, in comparison with the other settings, indicates possible formant underfitting, where this combination of parameters failed to accommodate low F2 estimates that were close to the corresponding F1 (e.g., in the case of high back vowels). The following section thus examines the effects of formant settings on individual vowels to investigate further the potential sources of error.

4.6.2 Formant centre frequencies by vowel

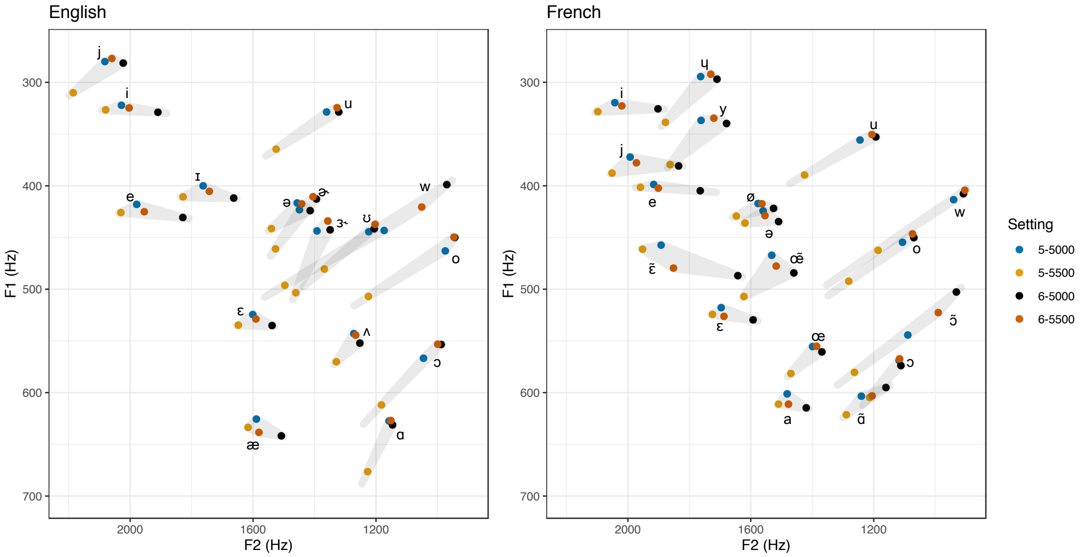

Figure 4.7: Formant plot of all monophthongs in English and French by formant setting, averaged across all 60 speakers.

An inspection of the vowel space, illustrated in Figure 4.7, provides a more detailed picture of the discrepancies in F1 and F2 across the tested formant settings. Mean F1 and F2 estimates for most vowels from three settings (5-5000, 6-5000 and 6-5500) were tightly clustered within a narrow range, with the notable exceptions of English /w/ and French /ɔ̃/. Estimates from the setting 5-5500 were generally set far apart from the other sets of estimates, in the direction of higher F1 and F2, though it is clear that back vowels were affected to a greater extent than front vowels. Both /u/ and /w/, as well as /ɔ̃/ in French, were estimated in this setting as highly central along the front-back dimension, when auditory judgment was in accordance with the more backed realisation implied by the other settings.

When the full F2 distributions of these back vowels were analysed more closely (Figure 4.8), it became clear that such apparent centralisation was mostly not the result of a gradient shift of formant estimates. Instead, the bulk of estimates were very similar to those obtained from other settings, but a small yet significant proportion of tokens were estimated to have an F2 of over 2000 Hz, suggesting that the setting 5-5500 was at times unable to detect a low F2 in close proximity to F1.

Turning to the higher formants, the impact of formant settings is similarly apparent in Figures 4.9–4.12. Gradient shift between formant settings was observed in all vowels, although, as in the case of F2, a subset of vowels can be seen to be responsible for the more extreme estimates.

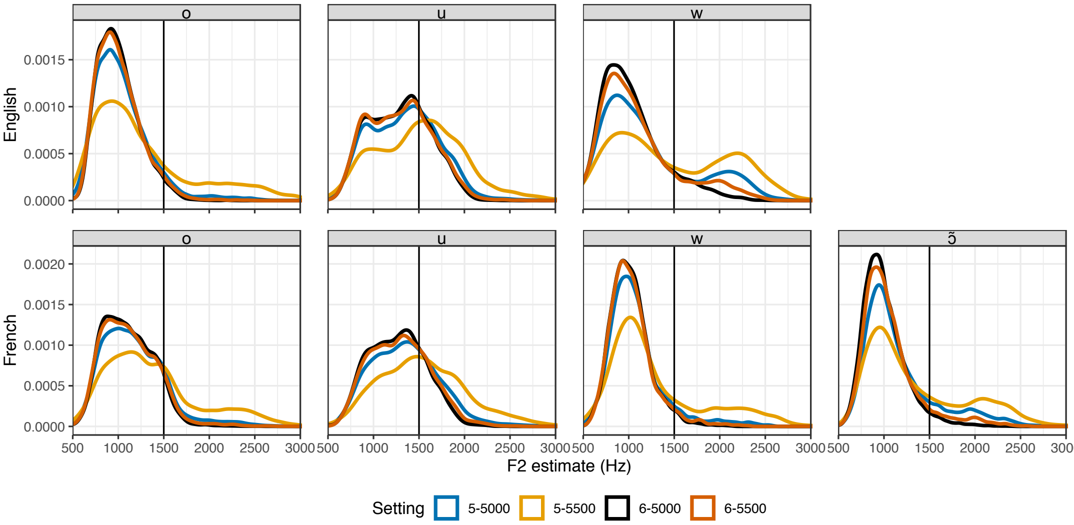

Figure 4.8: F2 distribution of /o u w/ in English and /o u w ɔ̃/ in French by formant setting. Vertical line at 1500 Hz added for reference.

Figures 4.9 and 4.10 show that high F3 and F4 from the setting 5-5500 was in part caused by the same vowels with unusually high F2, namely back (rounded) vowels. Other vowels contributing a high proportion of high F3 estimates in this setting included the rhoticised vowels in English (/ɚ ɝ/) and other rounded vowels in French (e.g., /y ø/). Both rhoticisation and rounding were known to depress F3 (Harrington, 2010) as corroborated by the other settings. The strongly bimodal distributions of /ɚ/ and /ɝ/ for 5-5500, with peaks near 3000 Hz, clearly could not be considered to be reliable. The heavy left tail in LTF4, extending to lower frequencies, from the setting 6-5000 was widespread across most vowels, as evident in Figures 4.11 and 4.12, but a bimodal distribution was noticeable in more fronted vowels (e.g., /e ɛ æ/), especially in French. For these vowels, a clear secondary peak around or below 2600 Hz could be found in the F4 distribution alongside another main peak above 3000 Hz.

4.6.3 Overall formant bandwidths

As illustrated in Figure 4.13, bandwidth estimates for all four formants showed strong positive skew and largely followed a log-normal distribution, with a small number of extremely high estimates. The setting 6-5000 presents an anomaly in BW2 and BW3, showing a somewhat bimodal distribution with a secondary peak at around 1800 Hz in both languages. Across different settings, bandwidth estimates were lowest F1 and monotonously increased for highest formants.

The effect of settings on bandwidth estimates is also apparent in Figure 4.13. The

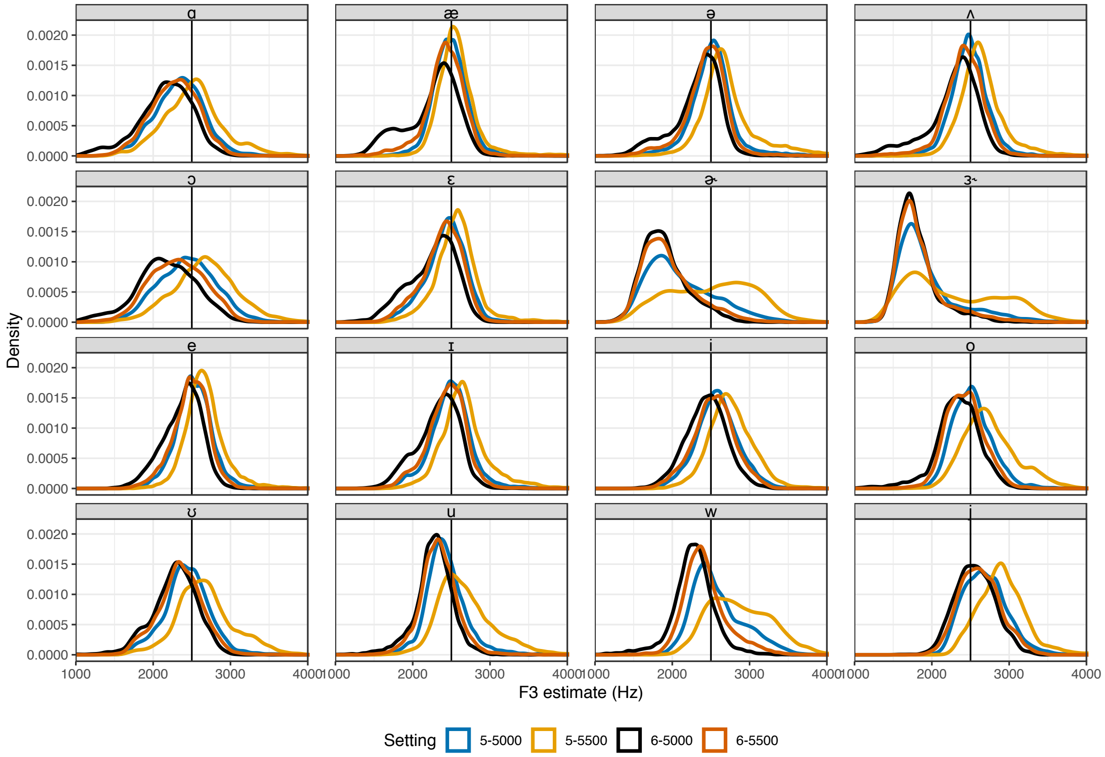

Figure 4.9: F3 formant frequency estimates by setting for each vowel in English.

distribution of BW1 had its peak at the lowest frequencies for the two settings with six formants, closely followed by 5-5000. The setting 5-5500 had its peak at a considerably higher frequency than the other settings. A similar, albeit diminished, trend can be observed for BW2 and BW3, whereas in the case of BW4, all settings produced very similar distributions of bandwidth estimates.

In summary, the effect of formant settings on bandwidths patterned similarly to their effect on formant centre frequencies. The settings 6-5000 and 6-5500 produced the lowest estimates here, followed by 5-5000 and finally 5-5500. However, the extent to which each formant was influenced was the other way round: Differences between settings here were greatest for BW1 and smallest for BW4. Concern may once again be raised for the setting 5-5500, which produced unusually high BW1, when compared with theoretically and empirically derived bandwidth measurements (Kent & Vorperian, 2018; K. N. Stevens, 1998). The setting 6-5000, with a relatively high proportion of BW2 and BW3 over 1000 Hz, has also emerged as another settings that may be considered unreliable and require further examination.

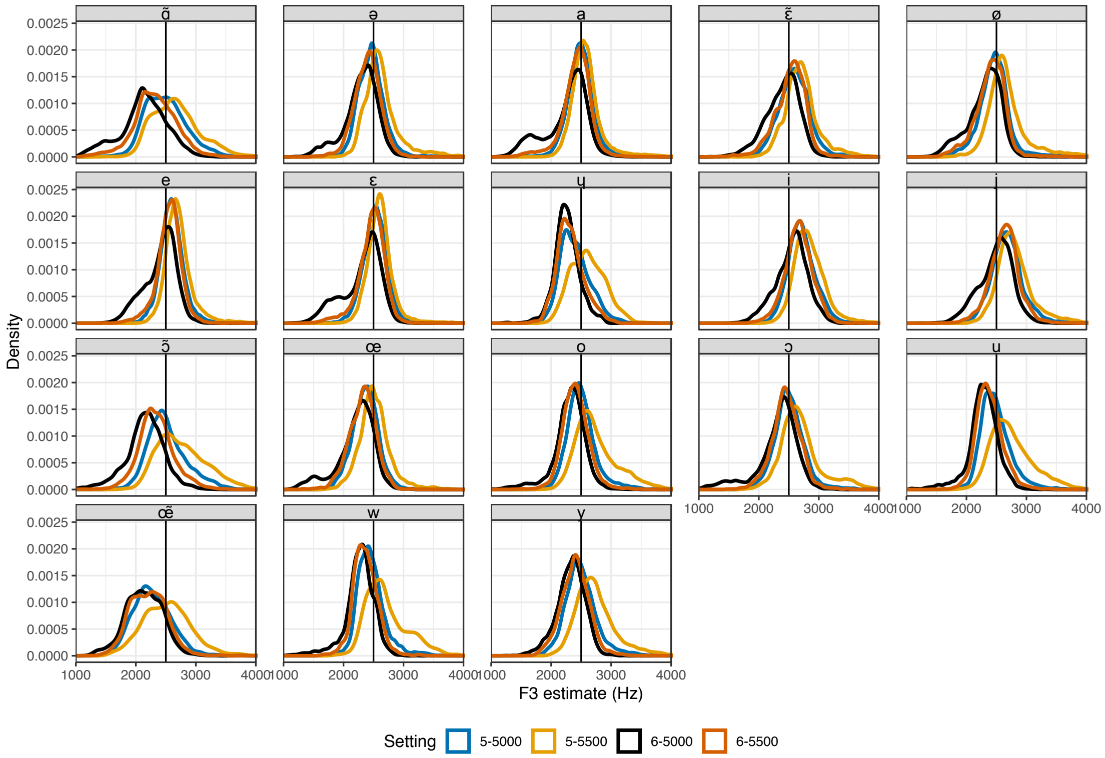

Figure 4.10: F3 formant frequency estimates by setting for each vowel in French.

4.6.4 Formant bandwidths by vowel

This section presents a more fine-grained analysis of formant bandwidths in individual vowels, focusing on BW1–BW3, which were shown to be most affected by formant settings above.

Figures 4.14 and 4.15 illustrate BW1 distributions collected from English and French vowels respectively. Formant setting had a clear effect on all vowels in English and all oral vowels in French. Whereas nasal vowels were recorded to have overall higher BW1, as would be expected due to the effects of nasal resonance, BW1 was generally stable across settings. The two six-formant settings yielded very similar distributions. While the distributions from the setting 5-5000 overlapped with them for the most part, there were also instances where a shift to higher frequencies could be observed. Such effects were particularly noticeable in back vowels (e.g., /o u w/) and rhoticised vowels (/ɚ ɜ˞/). The most prominent effects were found in the setting 5-5500, where an even greater upward shift was observable for all oral vowels.

In the case of BW2, illustrated in Figures 4.16 and 4.17, the secondary peak above 1000 Hz from the setting 6-5000 mentioned above would appear to be conditioned by vowel. Such peaks were prominent in almost all front vowels, regardless of height (e.g., /i

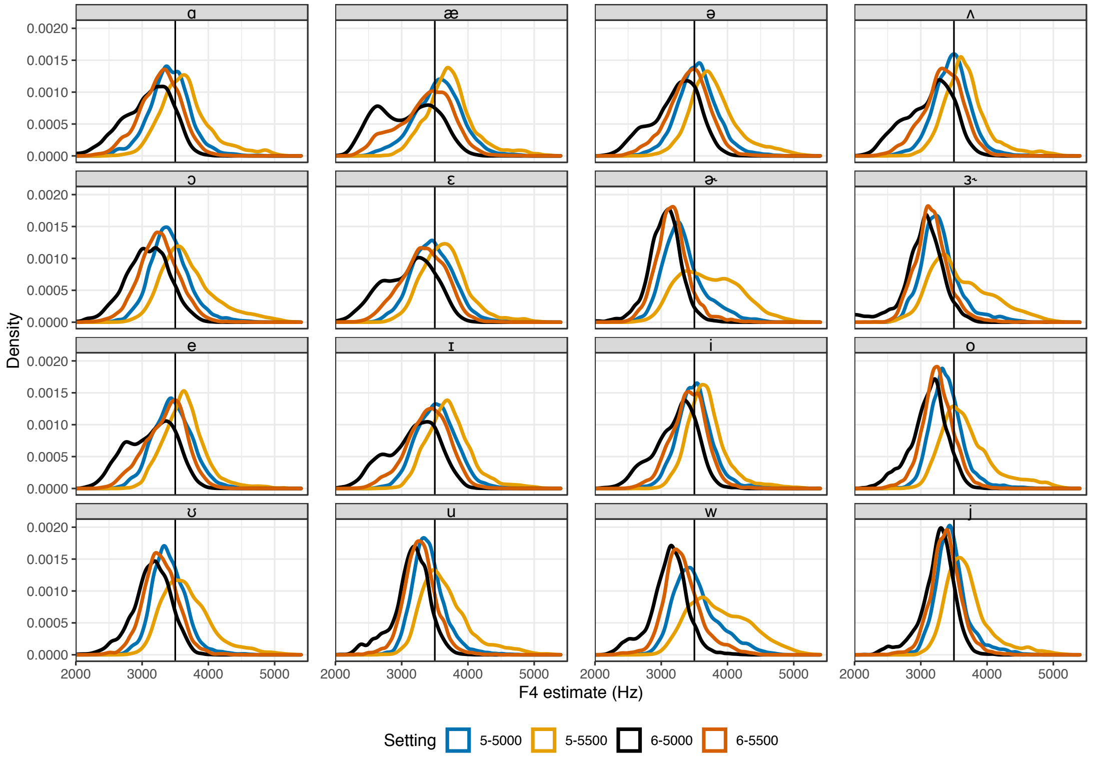

Figure 4.11: F4 formant frequency estimates by setting for each vowel in English.

e æ a/) or rounding (e.g., /y ø/), but not evidenced for the most part in back vowels (e.g., /u o/). Nevertheless, back vowels suffered from a different issue, as they were clearly much more subject to a wholesale upward shift in BW2 than front vowels in the settings 5-5500 and, to a lesser extent, 5-5000.

Unlike the distributions observed in BW2, Figures 4.18 and 4.19 show that the similar secondary peak in BW3 from the setting 6-5000 was clearly present to varying extents in all oral vowels, absent only from the French nasal vowels. More generally, formant setting appeared to have a gradient effect on the peak location of the distributions. The peak was lowest for the setting 6-5000 across the board, with a very similar or slightly higher peak for 6-5500. These were then followed by the setting 5-5000 and finally 5-5500. The discrepancy between the five- and six-formant settings was particularly pronounced for the rhoticised vowels in English (/ɚ ɝ/) and the high back /u/ and /w/.

4.6.5 Conclusions

The analysis above demonstrates the considerable impact that formant settings in Praat can have on the estimates of both formant centre frequencies and bandwidths. While the degree to which LTFDs are impacted varied depending on the formant in question

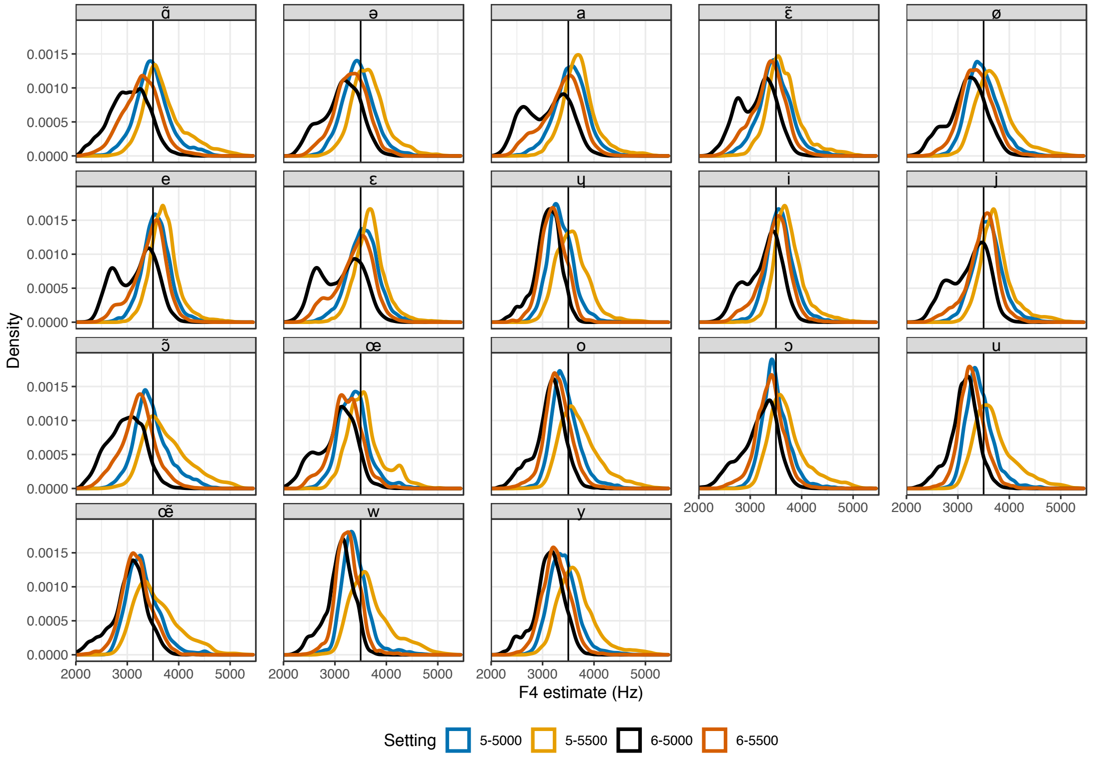

Figure 4.12: F4 formant frequency estimates by setting for each vowel in French.

and the identity of the vowel, the order of the settings was remarkably consistent across languages.

Differences in formant centre frequencies can mostly be attributed to formant overfitting, in the case of 6-5000, or underfitting, in the case of 5-5500. The remaining two settings, 5-5000 and 6-5500, were not immune from these issues when applied across the board to all vowels and speakers, but they were clearly affected to a much smaller extent. The patterning in the distributions of formant bandwidths similarly shows tendencies of misestimation in the settings 5-5500 and 6-5000, reinforcing their unsuitability to the current exercise.

Overall, in the present corpus, the settings 5-5000 and 6-5500 were considered more advantageous than 5-5500 and 6-5000. The setting 6-5500 was also considered to yield quantitatively more reliable estimates, particularly when it came to formant bandwidths, than 5-5000. On the basis of this preliminary testing, it was decided that, for the purposes of the current project, the setting 6-5500 would be applied to all formant extraction.

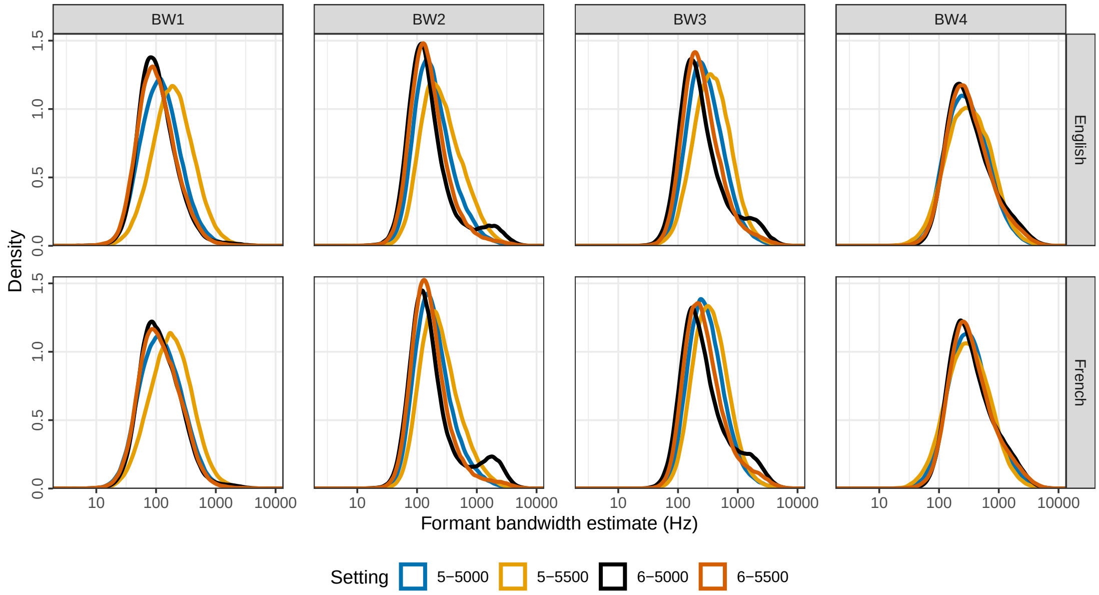

Figure 4.13: Formant bandwidth estimates in English and French by formant setting. Note that the horizontal axis is plotted on log scale.

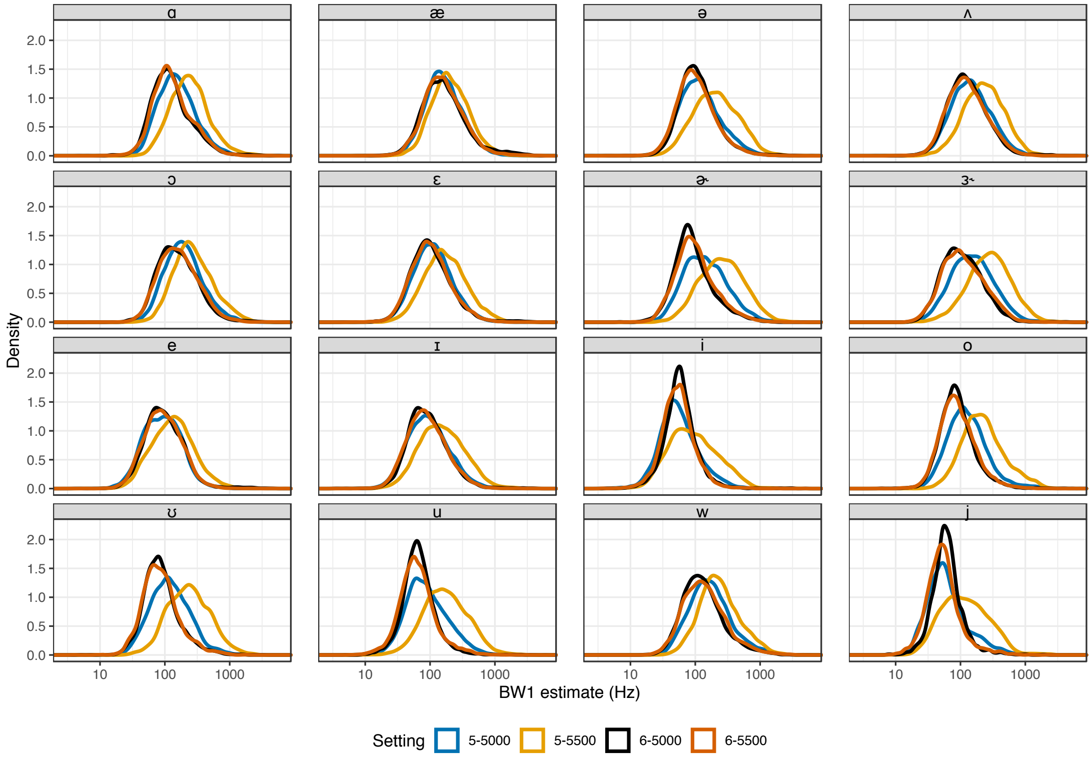

Figure 4.14: F1 bandwidth estimates by setting for each vowel in English.

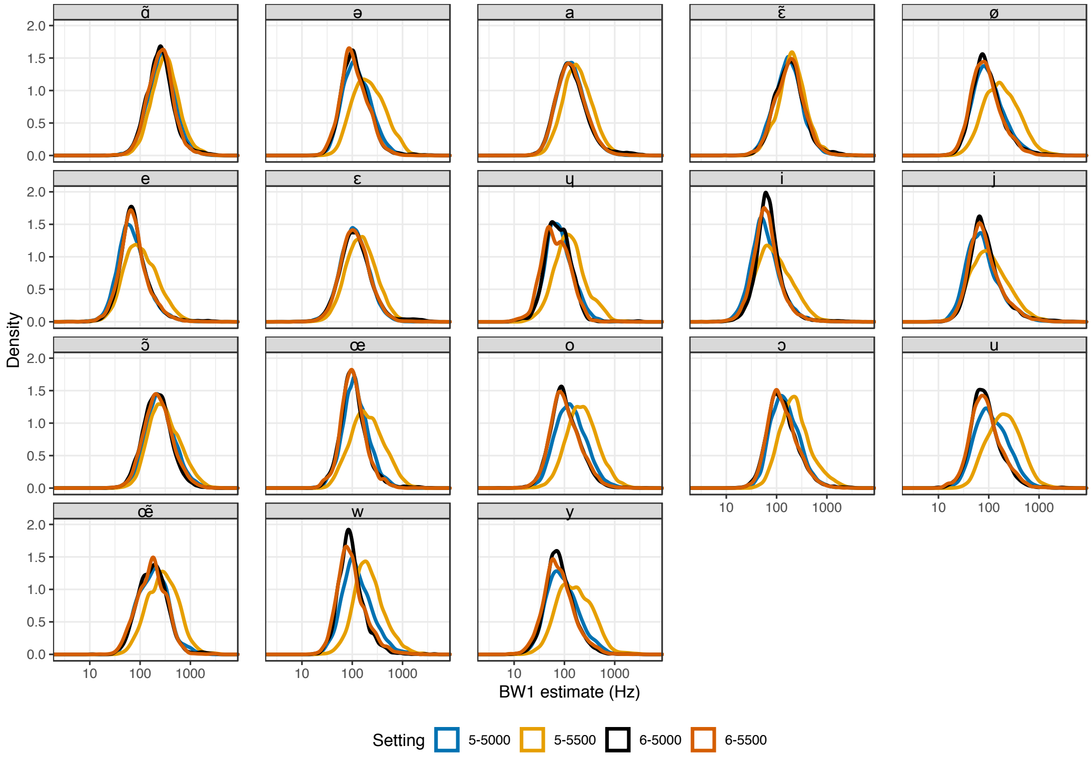

Figure 4.15: F1 bandwidth estimates by setting for each vowel in French.

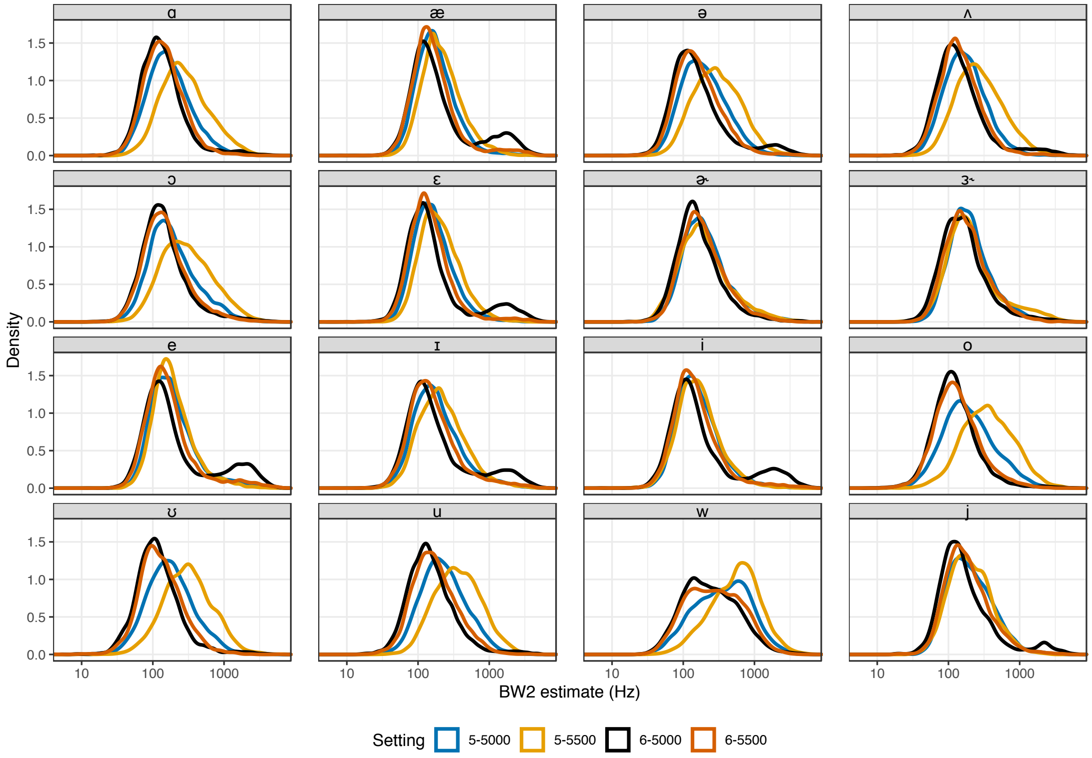

Figure 4.16: F2 bandwidth estimates by setting for each vowel in English.

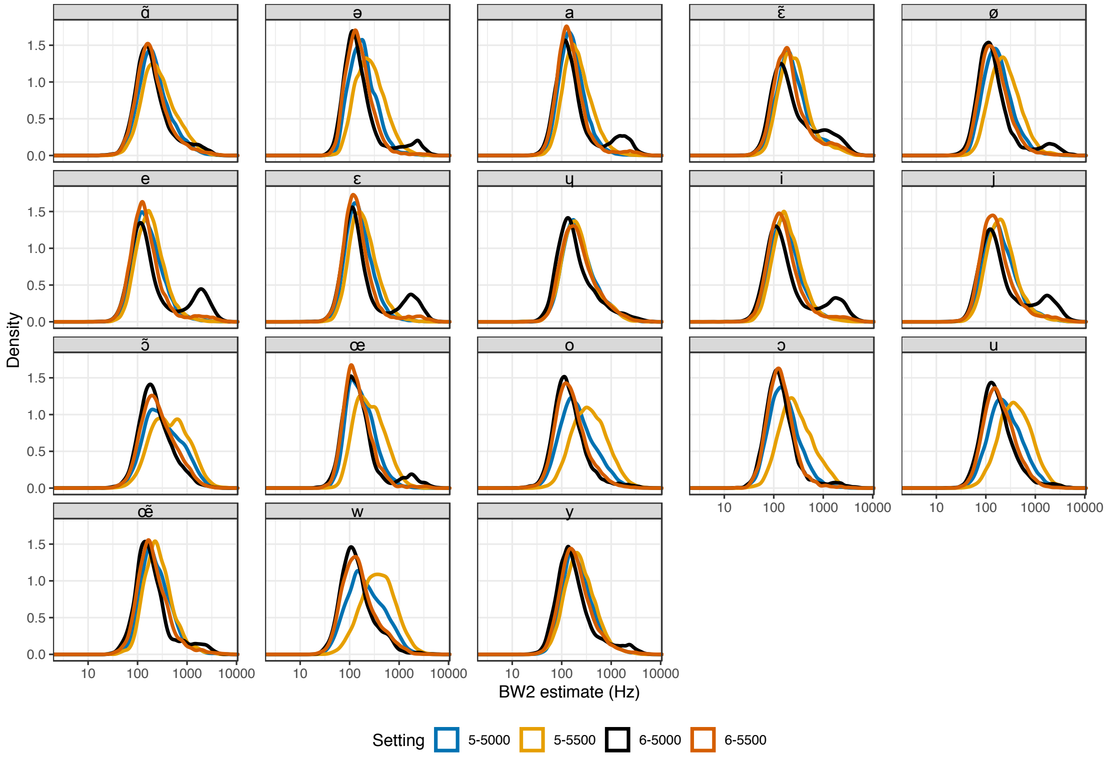

Figure 4.17: F2 bandwidth estimates by setting for each vowel in French.

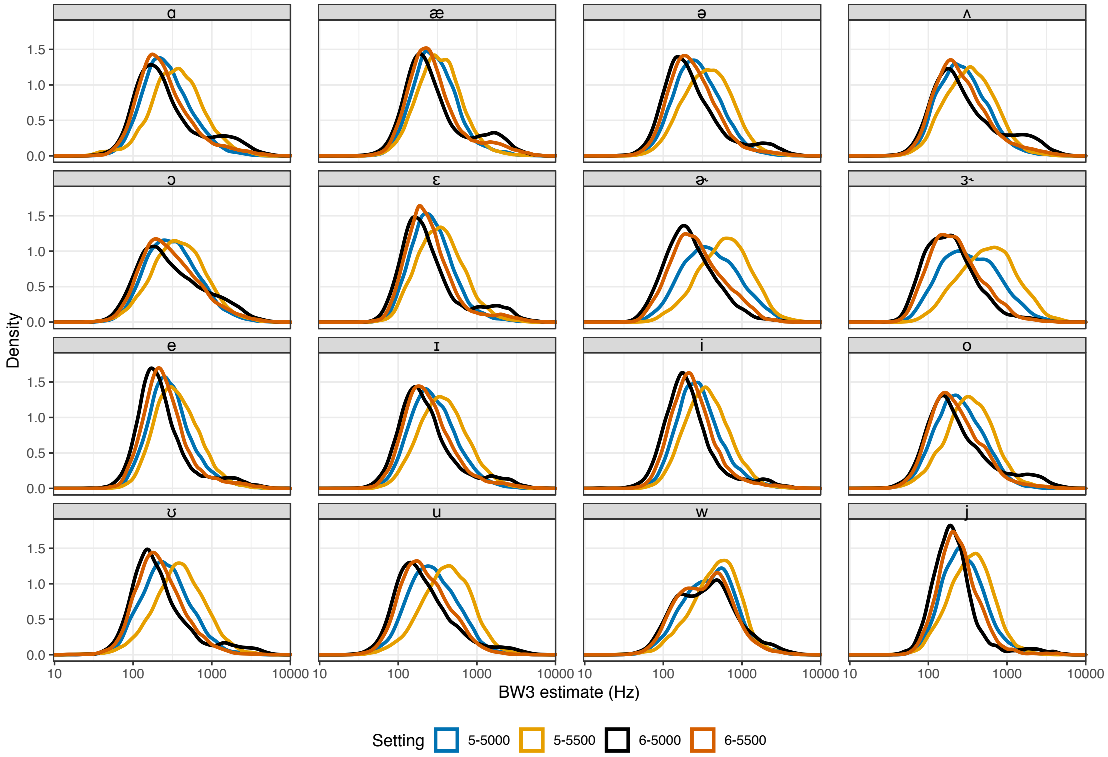

Figure 4.18: F3 bandwidth estimates by setting for each vowel in English.

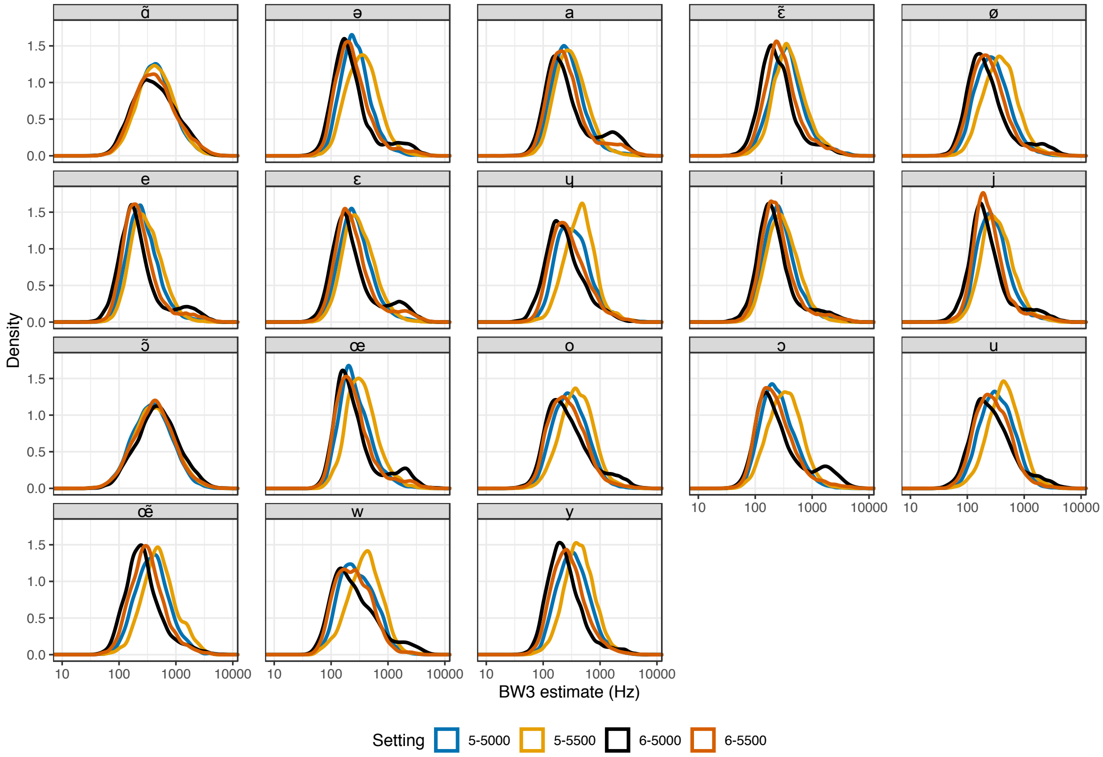

Figure 4.19: F3 bandwidth estimates by setting for each vowel in French.

Chapter 5

Results: Acoustic Phonetics

This chapter presents results from the acoustic analysis in two parts, first looking at the static and dynamic spectral analysis of /s/ in Section 5.1, before turning to the analysis of LTFDs in Section 5.2. The findings from each variable will be discussed within each part, before the chapter concludes in Section 5.3 with a general discussion of the implications that the current findings have for FVC, as well as specific predictions for the following forensic part of the current study.

5.1 /s/ acoustics

5.1.1 Static measurements

Figure 5.1 shows the overall distributions of each spectral moment in English and French, grouped by speaker L1. In this section, each spectral moment will be discussed in turn along with results from the corresponding statistical analysis.

Starting with CoG, Figure 5.1 shows broadly similar values across languages, although /s/ in French demonstrated a tendency to have slightly higher CoG than /s/ in English. In both languages, L1 French speakers produced higher CoG than L1 English speakers, while simultaneous bilinguals patterned closely with L1 speakers within each language. Such differences between languages and L1 groups were, however, not significant (Table 5.1). There was also no significant interaction between language and speaker L1. Of the included phonetic factors, only the environment following /s/ was shown to have a significant effect on CoG (Figure 5.2), with /s/ showing the highest predicted CoG when preceding an unrounded vowel, followed in descending order by rounded vowels, consonants other than /str/ clusters and pauses. The lowest predicted CoG came from the /str/ clusters, in line with previous research that /s/ shows signs of retraction in such

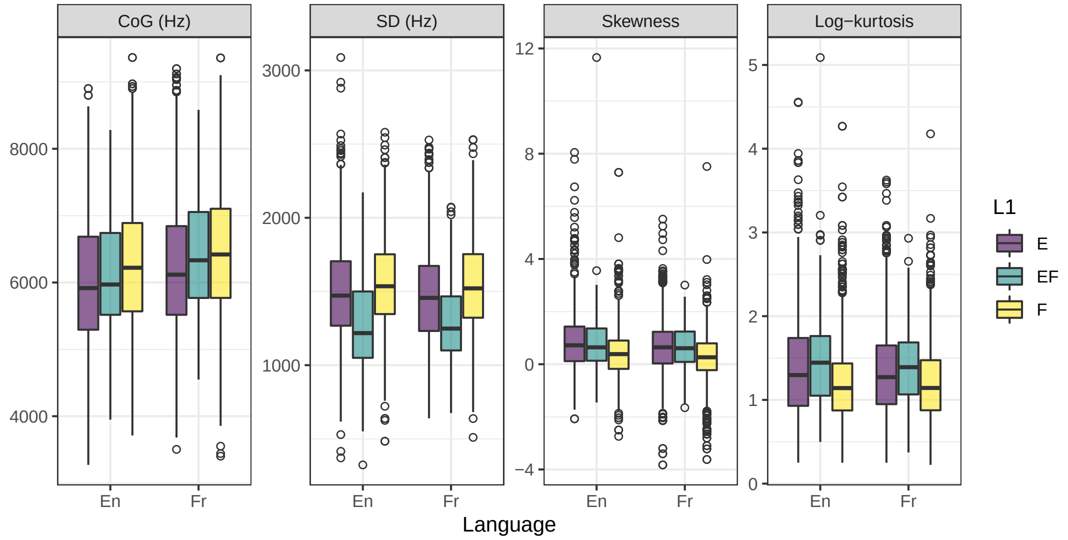

Figure 5.1: Boxplot of spectral moments in English and French for each L1 group (E: English; F: French; EF: simultaneous bilingual).

a context, though CoG was still high in /str/ (5862 Hz), and the distance between /s/ in /str/ and before unrounded vowels here was relatively small (439 Hz).

Table 5.1: Summary of fixed effects in best-fitting mixed-effects model for CoG. Effects (estimates) in Hz. Intercept refers to grand mean. p values based on model comparisons with LRTs.

	Estimate	SE	t	$p(\chi^2)$
(Intercept)	6164.37	98.87	62.35	–
Language				(.1041)
Speaker L1				(.4065)
Language × L1				(.4315)
Previous context				(.2469)
Following context (s#)	−178.54	41.96	−4.26	<.0001
Following context (sC)	−118.29	30.23	−3.91	
Following context (str)	−302.02	89.11	−3.39	
Following context (sRV)	1.37	42.14	0.03	
Duration				(.3129)

Like CoG, SD in English and French was also highly similar in Figure 5.1. Between different L1 groups, L1 English and L1 French speakers showed very similar /s/ SD values in both languages, whereas simultaneous bilinguals could be observed to produce /s/ with lower SD in both English and French. These trends were confirmed in the best-fitting model for SD (Table 5.2), where neither language nor its interaction with speaker L1 was returned as significant, but there was a significant main effect of speaker L1. Post-hoc Tukey-adjusted pairwise comparisons, conducted using the *emmeans* package

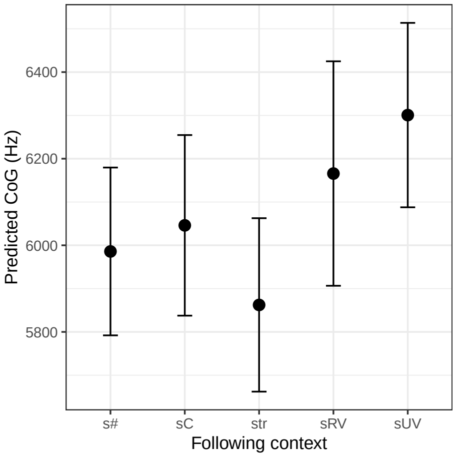

Figure 5.2: Predicted CoG (with 95% confidence intervals) for each following context.

(Lenth, 2020), indicated that there were significant differences between the simultaneous bilinguals and both groups of sequential bilinguals (L1 English: $p = .0476$; L1 French: $p = .0046$), but no significant difference between L1 English and L1 French speakers ($p = .3950$). In addition to speaker L1, both previous and following phonetic contexts, but not duration, were found to have a significant effect on SD (Figure 5.3).

Table 5.2: Summary of fixed effects in best-fitting mixed-effects model for SD. Effects (estimates) in Hz. Intercept refers to grand mean. p values based on model comparisons with LRTs.

	Estimate	SE	t	$p(\chi^2)$
(Intercept)	1523.19	24.67	61.74	–
Language				(.3830)
Speaker L1 (En)	−16.68	28.08	−0.59	.0041
Speaker L1 (Simul.)	−212.85	68.25	−3.11	
Language × L1				(.7844)
Previous context (#s)	16.55	15.46	1.07	.0080
Previous context (Cs)	3.97	8.47	0.47	
Previous context (RVs)	36.71	13.53	2.71	
Following context (s#)	−97.94	17.93	−5.46	<.0001
Following context (sC)	32.34	12.33	2.62	
Following context (str)	−100.55	40.01	−2.51	
Following context (sRV)	85.50	16.80	5.09	
Duration				.2674

Turning to the third spectral moment, all L1 groups produced small positive values of skewness that were similar in both languages. Figure 5.1 shows very little difference across L1 groups, although there seemed to be a tendency for L1 French speakers to show slightly lower skewness in their /s/. Model testing, summarised in Table 5.3, indicated that language and its interaction with speaker L1 did not have any significant

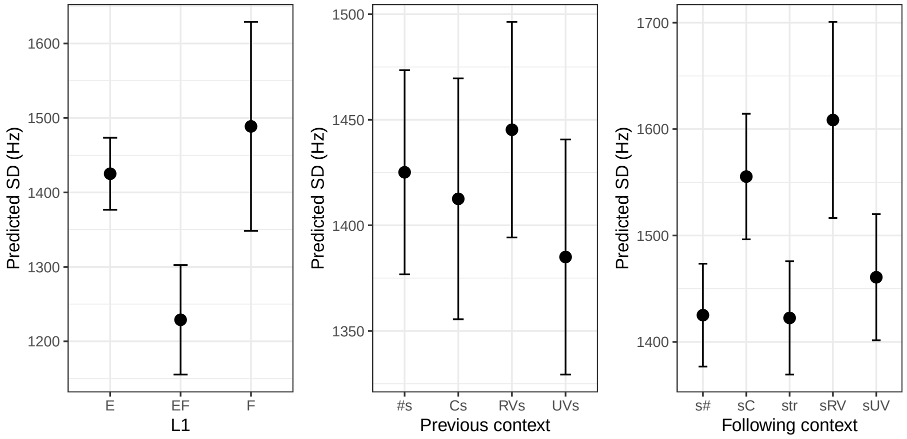

Figure 5.3: Predicted SD (with 95% confidence intervals) for each L1 group (E: English; F: French; EF: simultaneous bilingual), previous context and following context.

effect on skewness, while the main effect of speaker L1 was marginally insignificant. The following context was the only significant predictor in the best-fitting model (Figure 5.4). Post-hoc Tukey-adjusted pairwise comparisons between different L1 groups showed the difference between L1 English and L1 French speakers to be reaching significance ($p = .0569$).

Table 5.3: Summary of fixed effects in best-fitting mixed-effects model for skewness. Intercept refers to grand mean. p values based on model comparisons with LRTs.

	Estimate	SE	t	$p(\chi^2)$
(Intercept)	0.52	0.08	6.27	–
Language				$(.2553)$
Speaker L1 (En)	0.23	0.10	2.25	$(.0566)$
Speaker L1 (Simul.)	0.08	0.25	0.34	
Language × L1				$(.4880)$
Previous context				$(.5758)$
Following context (s#)	0.35	0.05	7.47	$<.0001$
Following context (sC)	0.05	0.03	1.89	
Following context (str)	0.24	0.10	2.33	
Following context (sRV)	-0.15	0.04	-4.10	
Duration				$(.2628)$

The last spectral moment, log-kurtosis, also displayed no discernible shifts between English and French. With regard to speaker L1, the patterning of log-kurtosis closely mirrored that of SD, but in the opposite direction: Simultaneous bilinguals showed higher values of log-kurtosis than L1 English and L1 French speakers, whose /s/ demonstrated generally similar log-kurtosis. As shown in Table 5.4, the factor of language, alongside

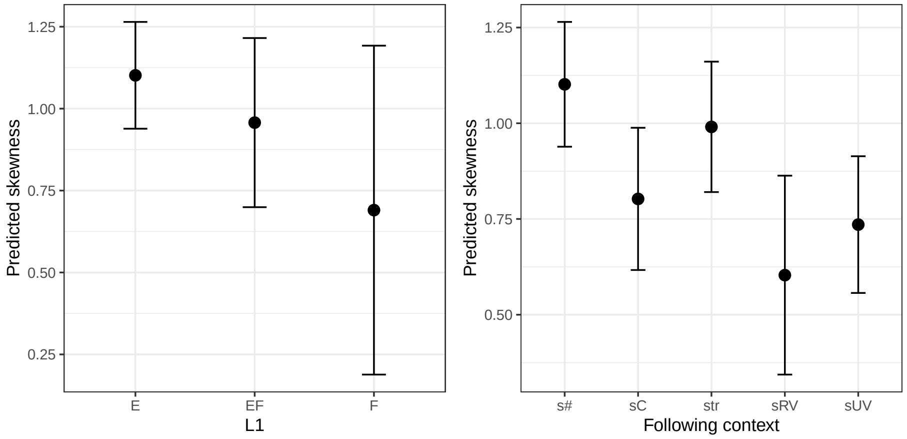

Figure 5.4: Predicted skewness (with 95% confidence intervals) for each L1 group (E: English; F: French; EF: simultaneous bilingual) and following context.

its interaction with speaker L1, was once again insignificant, while the effect of speaker L1 was marginally significant. However, post-hoc Tukey-adjusted pairwise comparisons found no significant differences between each pair of L1 groups (L1 English vs. simultaneous bilinguals: $p = .9171$; L1 French vs. simultaneous bilinguals: $p = .1980$; L1 English vs. L1 French: $p = .0856$). As in the case of SD, both previous and following contexts had a significant effect on log-kurtosis (Figure 5.5).

Table 5.4: Summary of fixed effects in best-fitting mixed-effects model for log-kurtosis. Intercept refers to grand mean. p values based on model comparisons with LRTs.

	Estimate	SE	t	$p(\chi^2)$
(Intercept)	1.25	0.04	32.65	–
Language				(.7856)
Speaker L1 (En)	0.08	0.04	1.87	.0486
Speaker L1 (Simul.)	0.13	0.11	1.24	
Language × L1				(.1784)
Previous context (#s)	−0.04	0.03	−1.57	.0232
Previous context (Cs)	0.01	0.01	0.89	
Previous context (RVs)	−0.05	0.02	2.41	
Following context (s#)	0.17	0.03	5.79	<.0001
Following context (sC)	−0.02	0.02	−1.28	
Following context (str)	0.01	0.07	0.16	
Following context (sRV)	−0.06	0.02	−2.50	
Duration				(.9128)

To compare the realisation of /s/ within and between individual speakers across languages, Figures 5.6 to 5.9 show, for each spectral moment, the correlation between English and French of each speaker's mean and spread. Strong correlations were found be-

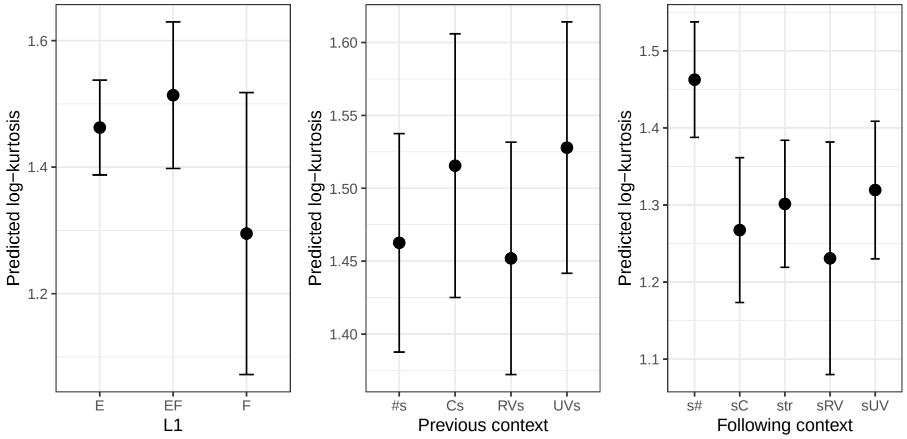

Figure 5.5: Predicted log-kurtosis (with 95% confidence intervals) for each L1 group (E: English; F: French; EF: simultaneous bilingual), previous context and following context.

tween English and French for all moments ($r > 0.80$), with CoG ($r = 0.92$) and skewness ($r = 0.90$) showing the strongest correlations.

In terms of variation between speakers, both CoG and SD display a high level of interspeaker variability in both languages (Figures 5.6, 5.7). Mean CoG ranges from 4282 Hz to 7993 Hz in English and further extends up to 8584 Hz in French. For mean SD, while most speakers occupy a narrow range between 1350 and 1600 Hz, a substantial minority of speakers report much more extreme values in either direction, such that it covers a range of 1044 to 2072 Hz in English and a slightly smaller range of 1128 to 2020 Hz in French. In comparison, Figures 5.8 and 5.9 show relatively low between-speaker variability for skewness and log-kurtosis. Mean skewness clusters within the range of 0 to 1 for most speakers, although a number of clearly defined outliers can be found in both directions. The dense concentration of speakers is even more apparent in the case of log-kurtosis, where only a small number of speakers report exceptionally high mean kurtosis.

Relevant too for forensic consideration is the within-speaker variability of spectral moments. In the left panels of Figures 5.6 to 5.9, this is represented by the whiskers extended from each speaker's mean. While a high degree of overlap between speakers is evidenced in all spectral moments, it can be seen from the long whiskers in Figures 5.7 to 5.9 that within-speaker variability, with respect to the whole distribution, is especially high for SD, skewness and kurtosis. Some speakers similarly exhibit much within-speaker variability in CoG, as can be seen from the right panel in Figure 5.6, but the standard deviation of CoG itself has a wide range of 323 to 1009 Hz, suggesting that speakers differ substantially in their ability to maintain a consistent acoustic target for CoG. Furthermore, for all spectral moments, a majority of speakers tend to have a greater spread in English than in French, as evident from the horizontal whiskers that are longer than the vertical whiskers in the left panels of Figures 5.6 to 5.9. This is explicitly illustrated in the corresponding right panels, which visualise the spread of spectral moments produced by each speaker in both languages. 34 (57%) speakers lie to the right the line of equality in the case of SD and kurtosis. The asymmetry is even more pronounced for skewness, where 38 (63%) speakers have a greater spread in English, and CoG, where the number of such speakers increases to 42 (70%). The absolute range of the spread across all speakers is also of greater magnitude in English than in French. As such, these results are suggestive of the existence of cross-linguistic patterns in the within-speaker variability of /s/, where each spectral moment is more variable in English than in French.

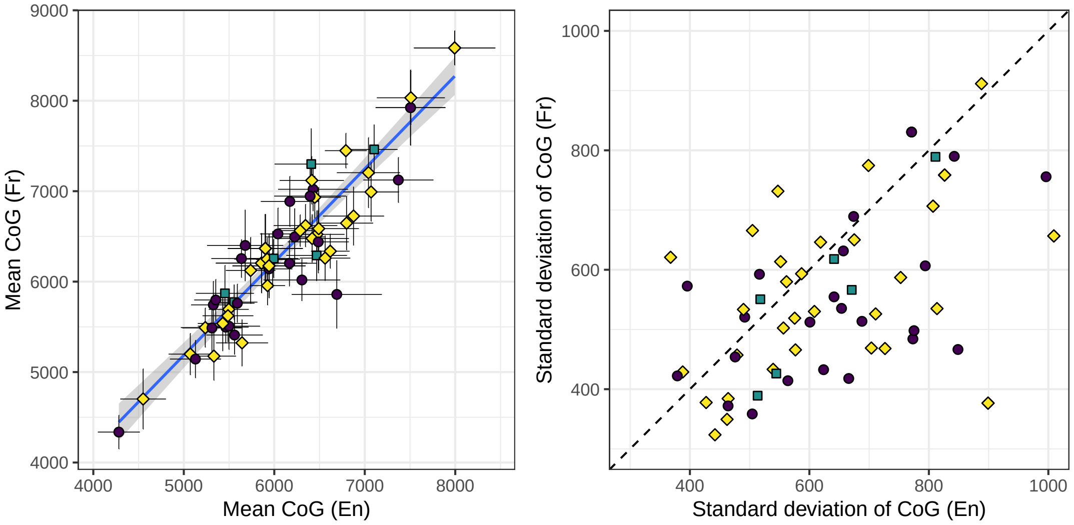

Figure 5.6: (Left) Scatterplot of speaker mean CoG (in Hz) in English and French with best-fitting line (95% confidence interval shaded). Whiskers extend to ±0.5× standard deviation. (Right) Scatterplot of standard deviation of CoG by speaker in English and French with line of equality (dashed). (Both) Purple circles: L1 English; green squares: simultaneous bilinguals; yellow diamonds: L1 French.

5.1.2 Dynamic measurements

This section presents the results from GAMM modelling of /s/. Table 5.5 summarises the outcomes from all model comparisons testing the effects of language and speaker L1. The discussion in this section will focus on these two factors, as well as consider the degree of variability in the trajectories exhibited between speakers, but full model summaries are presented (Tables 5.10–5.13) at the end of the chapter.

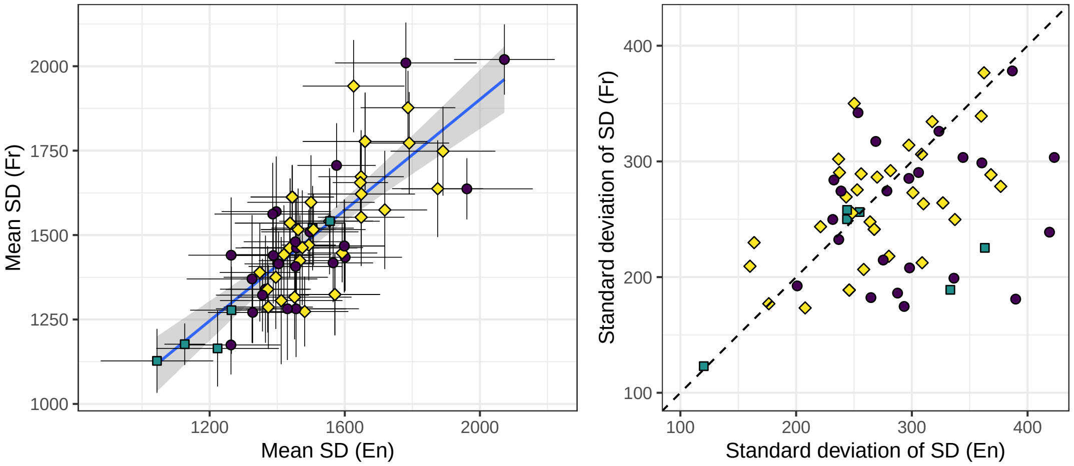

Figure 5.7: (Left) Scatterplot of speaker mean SD (in Hz) in English and French with best-fitting line (95% confidence interval shaded). Whiskers extend to ±0.5× standard deviation. (Right) Scatterplot of standard deviation of SD by speaker in English and French with line of equality (dashed). (Both) Purple circles: L1 English; green squares: simultaneous bilinguals; yellow diamonds: L1 French.

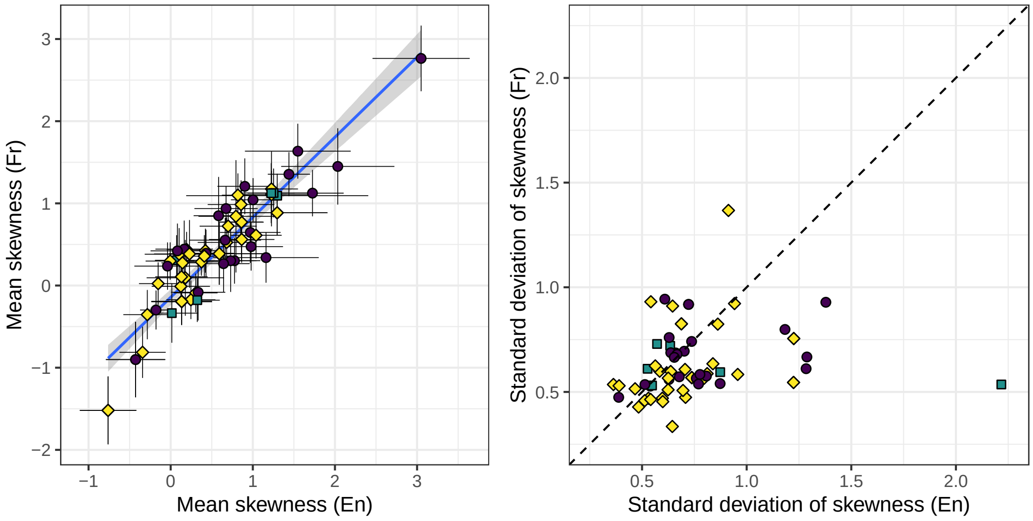

Figure 5.8: (Left) Scatterplot of speaker mean skewness in English and French with best-fitting line (95% confidence interval shaded). Whiskers extend to ±0.5× standard deviation. (Right) Scatterplot of standard deviation of skewness by speaker in English and French with line of equality (dashed). (Both) Purple circles: L1 English; green squares: simultaneous bilinguals; yellow diamonds: L1 French.

Focusing first on the trajectory of CoG over the duration of /s/, model comparisons (Table 5.5) showed that language had a significant effect on the shape of the trajectory. Including speaker L1, however, did not lead to better model fit, meaning that there were

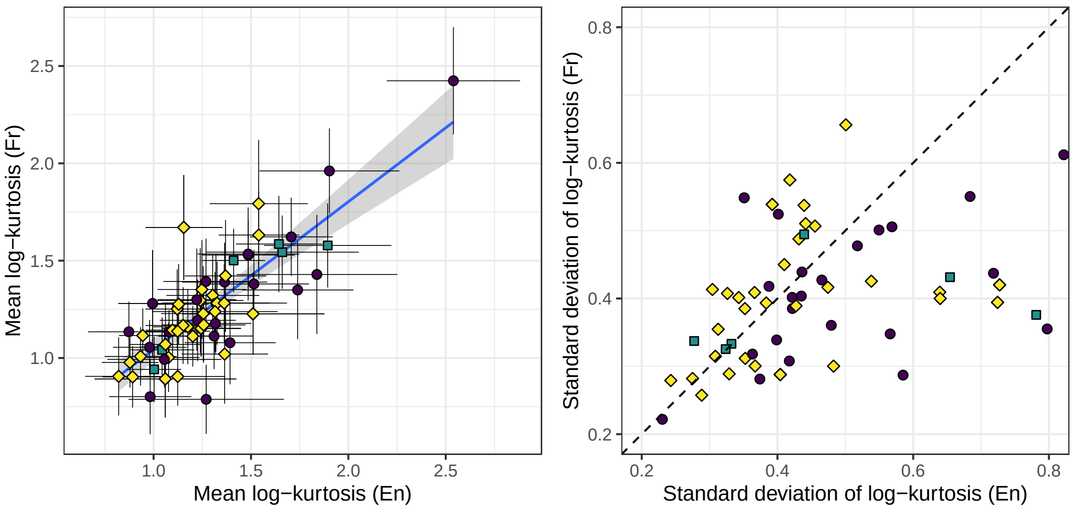

Figure 5.9: (Left) Scatterplot of speaker mean log-kurtosis in English and French with best-fitting line (95% confidence interval shaded). Whiskers extend to ±0.5× standard deviation. (Right) Scatterplot of standard deviation of log-kurtosis by speaker in English and French with line of equality (dashed). (Both) Purple circles: L1 English; green squares: simultaneous bilinguals; yellow diamonds: L1 French.

no significant differences between speakers with different L1 backgrounds. The difference between English and French is illustrated by the L1 English speakers in Figure 5.10. In English, /s/ CoG rises from the onset, past the midpoint, and reaches its peak at around 70% of the duration before falling sharply to the same level as the onset. While CoG in French follows a similarly parabolic trajectory, it begins at a higher frequency than English, reaches its peak earlier, at just after the midpoint, and then falls more sharply to reach a lower frequency at offset.

In the case of SD, including language did not result in better model fit. As such, no difference was found between English and French for the trajectory of /s/ SD. Table 5.5 further shows speaker L1 to have a significant effect on SD overall, but not on the shape of the trajectory. Figure 5.11 shows that, in both languages, SD moves in an opposite direction to CoG, first falling before rising to reach its highest level at the offset. Although SD in French in Figure 5.11 shows a tendency to rise later than in English, this difference was not significant. While little difference could be observed between L1 English and L1 French speakers, simultaneous bilinguals reported overall lower SD than the two groups of sequential bilinguals across the duration of /s/. These trajectories corroborate findings from Kavanagh (2012) and, to a lesser extent, Jongman et al. (2000), where SD traces a similar curve across the duration of /s/. The trough in SD indicates that the acoustic energy is concentrated around the narrowest range of frequencies near the middle of the

Table 5.5: Summary of GAMM model comparisons. Comparisons of shape effects only conducted for a predictor if overall comparison was significant. (* = No p-value as nested model had lower ML score than full model.)

Variable	Comparison	χ^2	df	$p(\chi^2)$
CoG	Overall: Language	125.05	3	$<.0001$
	Shape: Language	125.94	2	$<.0001$
	Overall: Speaker L1	4.40	6	(.1850)
SD	Overall: Language	-7.96	3	(*)
	Overall: Speaker L1	35.84	6	$<.0001$
	Shape: Speaker L1	-14.08	4	(*)
Skewness	Overall: Language	20.43	3	$<.0001$
	Shape: Language	18.70	2	$<.0001$
	Overall: Speaker L1	9.27	6	.0050
	Shape: Speaker L1	7.52	4	.0050
Log-kurtosis	Overall: Language	3.76	3	(.0570)
	Overall: Speaker L1	10.61	6	.0020
	Shape: Speaker L1	2.42	4	(.3040)

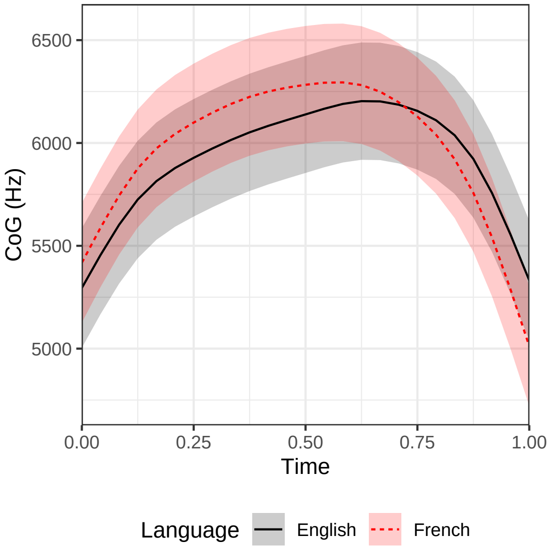

Figure 5.10: Predicted CoG trajectories for /s/ in word-initial context before unrounded vowels (#sUV) in English (black, solid) and French (red, dashed), with 95% confidence intervals shaded.

segment, when the jaw rises to its highest position (Iskarous et al., 2011), and more diffuse towards the onset and the offset. The findings for SD dynamics here closely align with those for static measurements, reaffirming the lack of cross-linguistic differences and the particular phonetic tendencies demonstrated by the small group of simultaneous bilinguals.

Turning to skewness, both language and speaker L1 were shown by model compar-

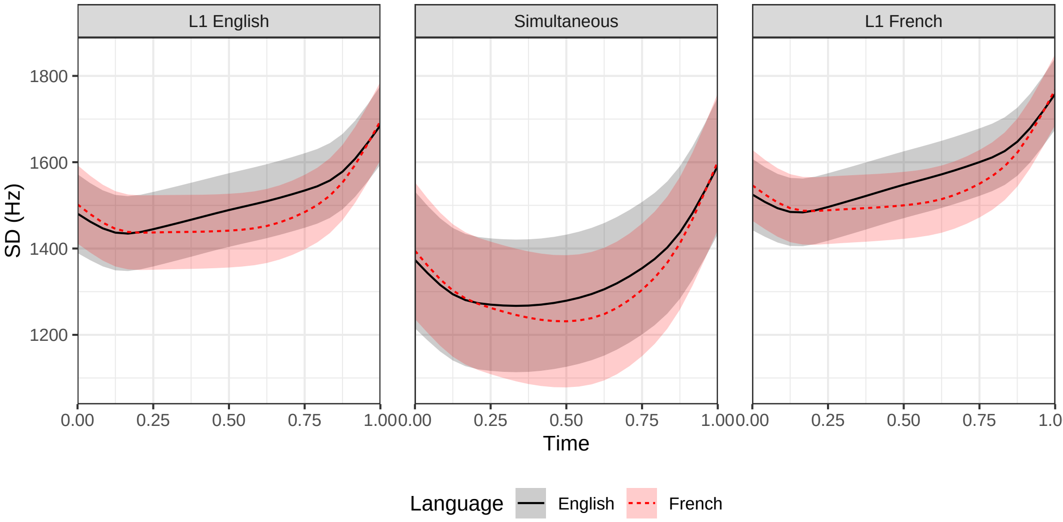

Figure 5.11: Predicted SD trajectories for /s/ in word-initial context before unrounded vowels (#sUV) in English (black, solid) and French (red, dashed), with 95% confidence intervals shaded.

isons to be significant predictors of the trajectory shape (Table 5.5). As shown in Figure 5.12, the difference between English and French is most distinctive at the onset, with French having a lower skewness than English, but the gap closes over the course of the fricative and is neutralised at the offset. Among different L1 groups, the trajectory of skewness tends to be higher overall for L1 English speakers than for L1 French speakers, but remains generally flat for both groups, with only very small movements throughout the duration of /s/. For simultaneous bilinguals, skewness more clearly demonstrates an initial rise and subsequent fall.

Model comparisons for log-kurtosis found no significant effect of language, but did show that an overall effect of speaker L1 on the variable. Nevertheless, including smooth terms for speaker L1 did not lead to an improved model, suggesting that the factor had no effect on the shape of the trajectory. Indeed, Figure 5.13 shows that log-kurtosis traces a very similar path in both English and French, falling from the onset until just past the midpoint, after which there is only a small degree of movement. Differences between speaker groups similar to those found for static measurements can also be found in Figure 5.13, where the trajectory of log-kurtosis as produced by L1 French speakers is shifted downwards relative to L1 English and simultaneous bilingual speakers.

To examine the extent of variability between speakers, Figure 5.14 illustrates the random smooths for individual speakers and shows how each speaker deviates from the overall smooth (i.e., average trajectory). For CoG, the lines representing most speakers are concentrated around the region near 0 Hz, meaning that their trajectories closely fol-

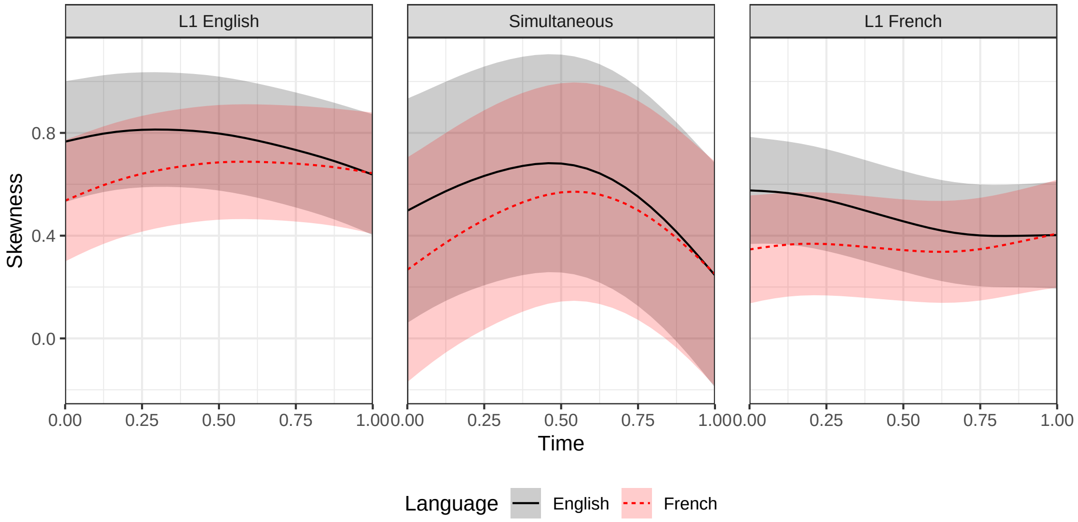

Figure 5.12: Predicted skewness trajectories for /s/ in word-initial context before unrounded vowels (#sUV) in English (black, solid) and French (red, dashed), with 95% confidence intervals shaded.

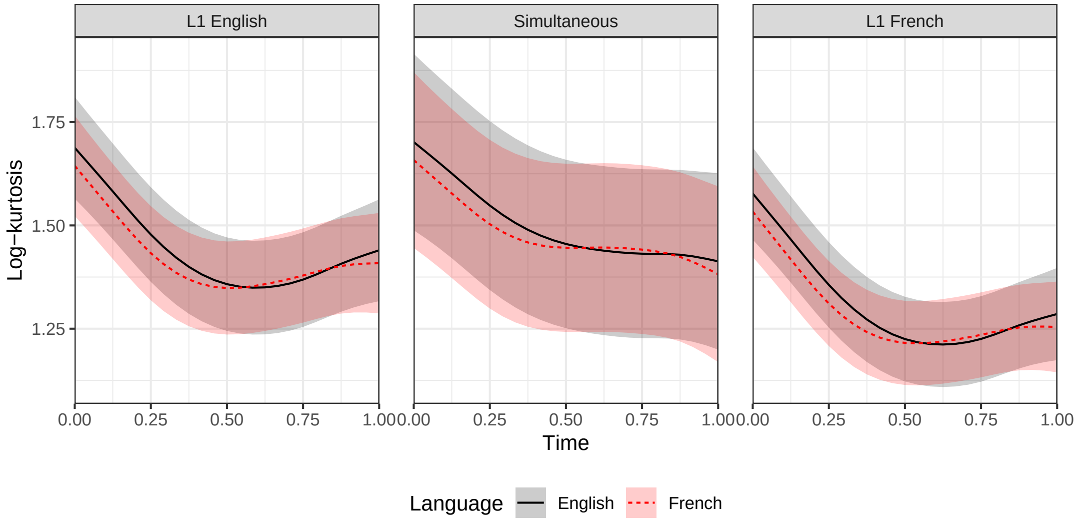

Figure 5.13: Predicted log-kurtosis trajectories for /s/ in word-initial context before unrounded vowels (#sUV) in English (black, solid) and French (red, dashed), with 95% confidence intervals shaded.

low the overall fit of the model. Moreover, most lines remain flat over the duration of the fricative, with only a small number of lines (mostly those most deviant from the overall fit) showing significant movements over time. CoG dynamics for the vast majority of speakers thus varies mostly in height, but not in the shape of the trajectory. Between-speaker variability of skewness presents a very similar picture, where most speakers do not deviate much from the overall group norm, either in height or in trajectory shape. These results suggest that the speaker-specific information which can be offered by dynamic measurements of CoG and skewness, on top of the static measurements, may be limited. While many speakers similarly cluster closely around 0 for log-kurtosis, they appear to demonstrate more movement over time with respect to the overall smooth. The incorporation of dynamic information for log-kurtosis is thus expected to carry stronger speaker-specificity than the use of only static measurements. By contrast, Figure 5.14 shows much greater variability in the dynamics of SD, with individual smooths spread over a wide range in both directions and displaying an array of movement patterns, thereby indicating a greater potential of increased speaker-specificity from SD dynamics.

5.1.3 Discussion

To summarise, the static acoustic analysis in Section 5.1.1 found that bilingual speakers of Canadian English and French in this study did not produce /s/ in the two languages with significant differences in any of the static spectral moments. Nevertheless, except for CoG, spectral moments were shown to be dependent on the L1 background of the speaker. /s/ as realised by simultaneous bilinguals reported lower SD and higher kurtosis than /s/ by L1 English and L1 French speakers, while skewness for L1 French speakers showed a tendency to be lower when compared to L1 English speakers. These findings were largely corroborated in the following dynamic analysis in 5.1.2, which found no effects of speaker L1 on CoG, and of language on SD and log-kurtosis. Dynamic modelling further showed that the effects of speaker L1 on SD and log-kurtosis were limited to the overall size but not the shape of the trajectory. Whereas /s/ in the two languages did not differ in terms of midpoint CoG and skewness, GAMM modelling revealed cross-linguistic differences in their trajectory over the duration of the fricative. CoG in French followed a trajectory with a sharper curvature than CoG in English. Skewness was lower in French /s/, but this difference was limited to the beginning portion. In addition, skewness dynamics were also significantly impacted by speaker L1, such that its trajectory was considerably flatter for sequential bilinguals than for simultaneous bilinguals.

The findings above show that, in terms of midpoint (static) measurements, bilingual

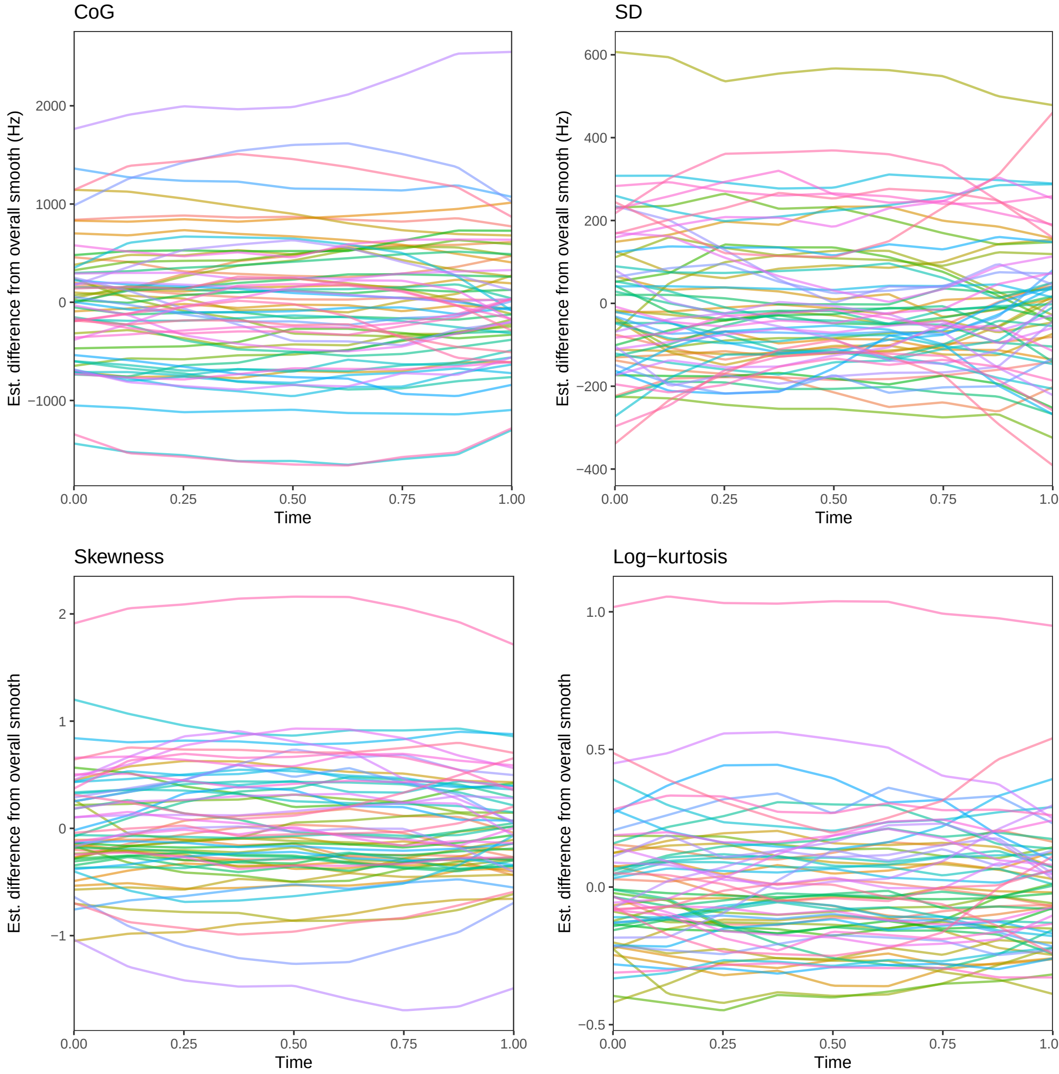

Figure 5.14: Random smooths for individual speakers in GAMMs fitted to CoG (top left), SD (top right), skewness (bottom left) and log-kurtosis (bottom right).

speakers did not contrast /s/ in Canadian English with /s/ in Canadian French acoustically. The lack of differences, particularly in the CoG measurements, suggests that, regardless of their L1 backgrounds, these speakers were not aiming for different targets when speaking either language. These results stand in contrast with those from Kitikanan et al. (2015) and Quené et al. (2017), where speakers distinguish the realisation of /s/ in their L1 and L2, and are instead more in line with the bilingual communities in Schertz et al. (2019) who make no distinction between their two languages. The specific language pair at work may be a relevant factor when comparing with previous studies, as Dutch, the L1 of the investigated speakers in Quené et al. (2017), is known to have a retracted /s/ relative to English. That is not the case for French, but neither is it for Thai in Kitikanan et al. (2015), which is found to have a more fronted articulation of /s/ than English. An alternative explanation for the different findings may lie in the background of the speakers. Both Kitikanan et al. (2015) and Quené et al. (2017) focused on late bilinguals: Whereas the Thai speakers in the former had only learnt English for several years in their home country and newly arrived in the UK, the Dutch speakers in the latter were considered to be highly proficient. The current study did not control for potential differences between early and late learners, as well as between speakers with different levels of proficiency, among speakers with the same L1. As such, the speakers here came from a much more heterogeneous background than the two mentioned studies and were relatively more akin to the L1 Korean–L2 Mandarin bilinguals in Schertz et al. (2019), who displayed considerable variation in self-rated proficiency and relative dominance of each language. This point will be returned to in more detail later in the discussion.

Nevertheless, cross-linguistic differences were indeed found for CoG and skewness. Such differences reside not in the target but within the spectral dynamics over the duration of /s/. The trajectory for English /s/ CoG observed here is broadly in line with findings from Iskarous et al. (2011) and Reidy (2016). In the word-initial context, which forms the focus of the earlier studies, /s/ exhibits a similar rise for most of its duration before falling to the offset. While there are discrepancies between the studies in how much CoG rises and falls over the course of the segment, they may be due to the methods of elicitation and precise phonetic environments, which are only coarsely categorised in the present study. Between English and French, the findings here suggest that, in terms of gestural timing, speakers seemed to reach the acoustic peak relatively earlier in their production of /s/ in French, mirroring the pattern found in Reidy (2016) between English and Japanese. The trajectories also differed in how much CoG dropped after reaching its peak, whereby French /s/ demonstrated a sharper drop especially toward the offset of the fricative.

To account for differences in CoG dynamics, Reidy (2016) offers an explanation for the cross-linguistic variation between English and Japanese based on a task-dynamic model of speech production (Fowler & Saltzman, 1993), attributing the earlier fall of peak ERB_N in Japanese to a greater extent of temporal overlap in gestural co-production between the fricative and the following vowel. The implication is that language-specific patterns of gestural co-production, or coarticulation, are something to be learned in the course of language acquisition. It could be the case that similar language-specific distinctions in coarticulation apply to the case of English and French, as a result of French /s/ favouring anticipatory coarticulation more than its English counterpart (Hoole et al., 1993; Niebuhr et al., 2011).

Cross-linguistic differences of /s/ also seem to manifest in the degree of variability of the segment, as all spectral moments showed a tendency to be more variable in English than in French. A particular contributing factor to this phenomenon may be the distribution of phonetic environments. As shown in Figures 5.2 to 5.5, the following context of /s/ has a consistent effect on spectral moments, while previous context similarly influences SD and kurtosis. To take CoG as an example, /s/ in the cluster /str/ has been found to show retraction in Canadian English (Stuart-Smith et al., 2019). This finding is replicated in the current analysis, where /str/ is estimated to have the lowest CoG among all following contexts. The occurrence of this cluster and the ensuing retraction are, however, limited to English (albeit in low counts) and not present in French. The scope of variability in French is thus considered to be more limited.

The patterns here also do not show much support for an alternative account based on the claim in the literature that non-native speakers show greater phonetic variability in speech production than native speakers of the same language (e.g., Wade et al., 2007; Witteman et al., 2014). In such a scenario, the L1 French–L2 English speakers, which account for more than half of the group, could have heavily contributed to the asymmetry in the data. While it is not the aim of the current analysis to address the question of whether non-native speech is more variable than native speech, it is worth considering briefly whether this factor may be acting as a confound for the effect of language. An inspection of Figures 5.6 to 5.9, however, shows clearly that, in all spectral moments, L1 English speakers did not appear to be any more likely than L1 French speakers to display higher variability in their L2 French relative to their L1 English, and vice versa. The results here suggest that the overall picture of higher variability in English is unlikely due to a bias induced by the imbalance in the population, and further align with recent research in Vaughn et al. (2019) and Xie and Jaeger (2020), who found no empirical evidence for a general claim of higher variability in non-native speech, but instead proposed that such a phenomenon is feature- and cue-specific.

Looking beyond systematic language effects, there is evidence of much between-speaker variation, both on the group level based on L1 background and on a more fine-grained, individual level. The effects of speaker L1 were most consistently found in SD and log-kurtosis, where simultaneous bilinguals patterned differently from both groups of sequential bilinguals. The lower SD and higher kurtosis from simultaneous bilinguals would imply that the magnitude of such effects was small. Whilst insignificant, Figure 5.1 does depict a tendency of L1 English speakers producing /s/ with lower CoG and higher skewness than L1 French speakers. Simultaneous bilinguals, for CoG in particular, patterned more closely with L1 English speakers when speaking English and more closely with L1 French speakers when speaking French. The general direction of the data is therefore trending in the same direction as the often-claimed distinction between English and French, in which L1 French speakers employ a more fronted articulation of /s/ than L1 English speakers, resulting in a higher CoG due to the decrease in the size of the front cavity. It is therefore possible that, on top of differences of timing in spectral moments, there are indeed differences in the targets of /s/ for Canadian English and French, but the acoustic consequences of such distinctions are not well captured by spectral moments. Indeed, Koenig et al. (2013) advocate the use of spectral measures such as the amplitude and sound level in different frequency regions, which are more theoretically driven to characterise fricatives but remain to be tested whether they could capture the posited differences. Alternatively, Jannedy and Weirich (2017) argue that DCT coefficients are better placed than spectral moments to characterise the entire spectrum of /s/ and capture more minute spectral differences. Another possible alternative to spectral measures that future investigations may pursue is the use of cepstral coefficients, which have been said to outperform spectral measures in classifying fricatives within the same language (Spinu et al., 2018; Spinu & Lilley, 2016). In that case, however, the issue of relating differences in coefficients to articulatory correlates remains, and if anything becomes more complicated.

Individual differences may nonetheless play a role in the lack of significant crosslinguistic differences. As Dart (1998) and subsequent studies have found, even within the same language speakers exhibit a great deal of variability in the specific target and gesture they adopt. Furthermore, as mentioned above, the heterogeneity within each group of speakers with different L1 backgrounds may have contributed to the apparent similarity between groups. If Canadian English and French do pursue different articulatory or acoustic targets, speakers who acquire their L2 earlier and receive more input may be more likely to form separate phonetic categories and consequently differentiate the two sounds (Flege & Bohn, 2021). The acquisition of /s/ is not a focus of the present study, and the data here do not in fact allow for in-depth explorations in this regard, but if the age of exposure to L2 is tentatively taken as a proxy for amount of L2 input, then a weak, insignificant trend can be observed in the current data suggesting that sequential bilinguals exposed to their L2 earlier producing /s/ in the L2 with CoG closer to the L1 speaker norms, away from the norms in their own L1 (Figure 5.15). Future studies with a design aimed at teasing apart different factors in the acquisition of /s/ will be able to address this issue more fully.

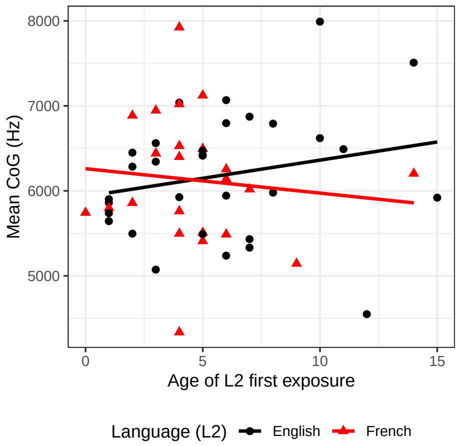

Figure 5.15: Scatterplot of mean CoG (in Hz) for L2 speech of self-reported sequential bilinguals (black circles: L1 French–L2 English bilinguals speaking English; red triangles: L1 English–L2 French speakers speaking French) and their age of L2 first exposure (to the nearest integer).

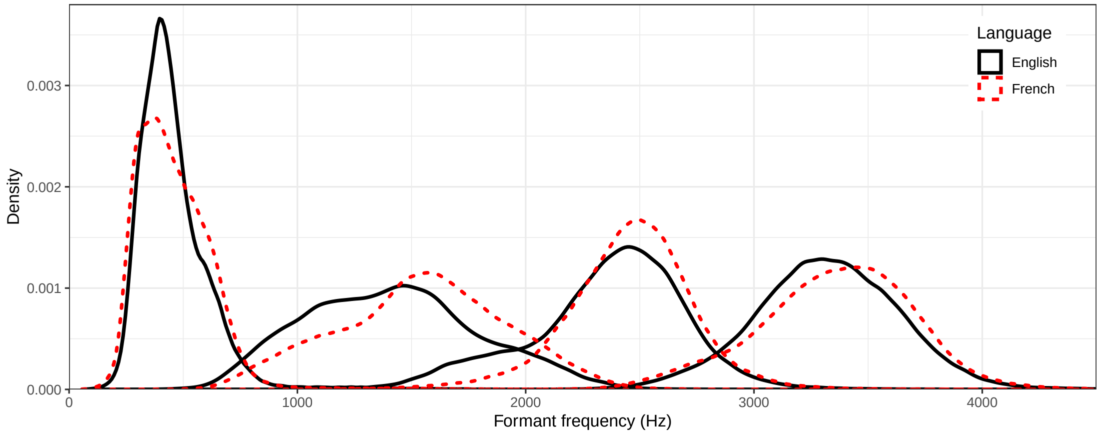

Figure 5.16: LTF1–4 distributions in English and French, pooled across all 60 speakers.

5.2 LTFDs

Figure 5.16 shows the overall distribution of LTFDs in each language. Compared to LTFDs produced in English, the distributions in French showed a similar peak frequency in LTF1 but higher peak frequencies in LTF2–4. Some differences in the shapes of the distributions could also be observed between the two languages. LTF1 was more sharply peaked in English, while LTF2 showed a heavier tail in lower frequencies in English. In the case of LTF3, both a flatter peak and heavier lower tail could be found.

Mixed-effects models fitted to the formant data confirmed cross-linguistic differences in LTFDs (Table 5.6). The effect of language was not significant for LTF1, but was significant for LTF2–4, indicating that speakers produced higher LTF2–4 means in French than in English. The estimated difference between the two languages was the largest for

LTF2 (133 Hz), but smaller for LTF3 (92 Hz) and LTF4 (34 Hz). Neither speaker L1 nor its interaction with language was significant for any of the LTFDs.

Table 5.6: Summary of fixed effects in best-fitting mixed-effects model for LTF1–4 (LTF1 log-transformed to correct for skew in distribution). Intercept refers to grand mean. *p* values based on model comparisons with LRTs.

	Estimate	SE	*t*	$p(\chi^2)$
LTF1				
(Intercept) (log10-transformed)	2.64	0.0032	830.90	–
Language				(.6472)
Speaker L1				(.1061)
Language × L1				(.6812)
LTF2				
(Intercept)	1475.35	6.59	223.97	–
Language	133.28	5.98	22.29	<.0001
Speaker L1				(.4606)
Language × L1				(.1801)
LTF3				
(Intercept)	2413.08	10.69	225.84	–
Language	92.27	7.03	13.12	<.0001
Speaker L1				(.9644)
Language × L1				(.8302)
LTF4				
(Intercept)	3330.81	14.18	234.89	–
Language	33.98	9.26	3.67	.0004
Speaker L1				(.6415)
Language × L1				(.5791)

To explore the variability of LTFDs between speakers, Figure 5.17 displays the individual LTF1–4 distributions in both languages. It is at once obvious that there is considerable between-speaker variation in terms of peak location for the higher formants, LTF3 and LTF4, in both English and French. Peaks in individual LTF3 distributions range from just above 2000 Hz to around 3000 Hz, while peaks in LTF4 distributions range from below 3000 Hz to around 3600 Hz. The shape of the distributions similarly demonstrates considerable variability, along a spectrum of flat to sharply peaked distributions. While the general pattern in both languages shows no strong differences, it can be observed that LTF3 distributions in French are generally more sharply peaked than those in English, in line with the overall observation above. These patterns can be contrasted with those displayed by LTF2, which has a relatively diffuse distribution in both languages. Speakers cluster close together, with little variation in shape or location of peak frequency. For LTF1, patterns of individual variability appear to differ between the two languages. Whereas in English the peaks are concentrated within a narrow range and there is only limited variation in shape (with some cases of bimodal distribution), in French LTF1 peaks are spread over a much wider range, accompanied by a greater variety in the shape of the distributions.

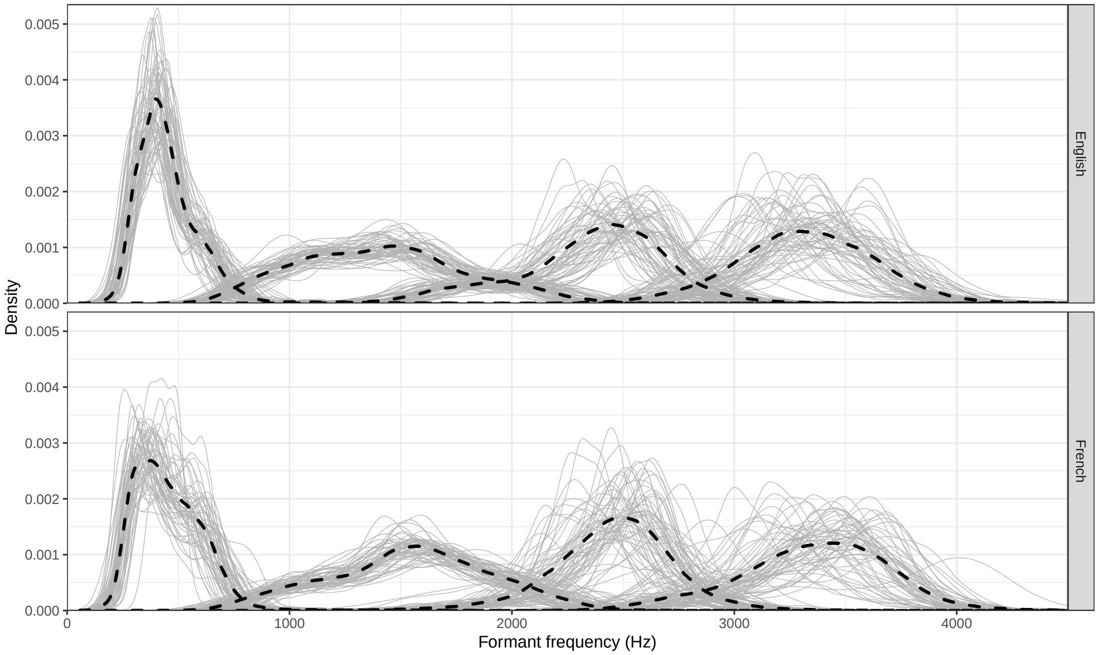

Figure 5.17: Individual (grey, solid) and overall (black, dashed) LTF1–4 distributions in English and French.

To further explore the cross-linguistic consistency within speakers, Figure 5.18 illustrates LTFD means for each speaker in English and French. LTFD SDs for each speaker are illustrated in Figure 5.19. The cross-linguistic differences in the means and shapes of LTFDs described above are clearly evidenced here. The vast majority of speakers produced higher LTF2–4 means in French in Figure 5.18, whereas in Figure 5.19 all speakers produced lower SDs for LTF3 in French than in English, indicating a lower variability that is consistent with the sharper peak in Figure 5.16. While considerable variability between speakers could be found, individuals remained consistent across languages. As summarised in Table 5.7, strong correlations of speaker means between English and French were found for LTF1, LTF3 and LTF4, while LTF2 means showed a weaker correlation. Similarly, SDs for each LTFD were strongly correlated, most strongly in the case of LTF1 and LTF4. Taken together, these results show that, in spite of the effect of language, the speaker-specificity of LTFDs is largely maintained across languages.

As for formant bandwidths, models fitted yielded results largely similar to those for LTF1–4. As summarised in Table 5.8, language was found to significantly improve model

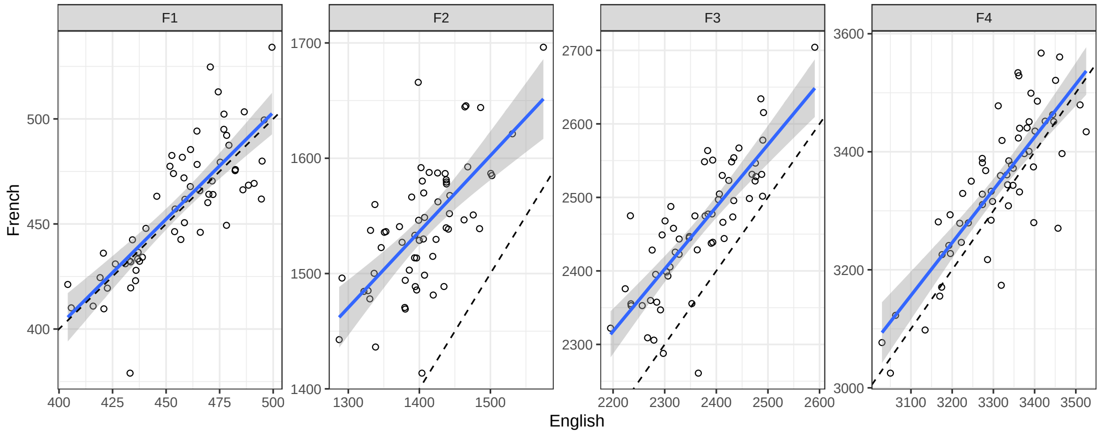

Figure 5.18: LTFD1–4 means by speaker with best-fitting line (solid; 95% confidence intervals shaded) and line of equality (dashed).

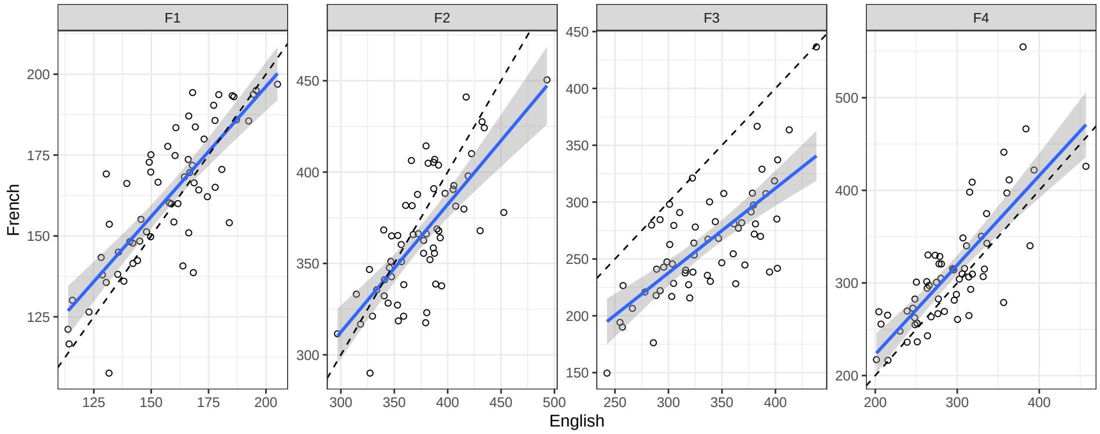

Figure 5.19: Standard deviations of LTFD1–4 by speaker with best-fitting line (solid; 95% confidence intervals shaded) and line of equality (dashed).

fit for BW1–4, with speakers producing overall higher formant bandwidths in French than in English. In all cases, speaker L1 did not have a significant effect, and there was no significant interaction between language and speaker L1. Table 5.8 further shows that, while significant, the effect of language was very small for any of the bandwidths. This is also illustrated in Figure 5.20, where the distributions in English and French overlapped to a high degree.

The distributions of BW1–4 from individual speakers are shown in Figure 5.21. Contrary to LTFDs depicted in 5.17, the distribution of bandwidths shows relatively little variation across all formants. In either language, the location of the peak frequency of individual speakers rarely deviates much from that of the overall distribution, except for a small number of speakers in the case of BW1 and BW4 in English and BW1 in French. Variability in shape is also generally low and largely confined to the sharpness of the

Table 5.7: Correlations between English and French (Pearson's r) for LTFD speaker means and SDs ($p < .0001$ in all cases).

	LTF1	LTF2	LTF3	LTF4
Mean	0.80	0.66	0.81	0.81
SD	0.80	0.73	0.71	0.78

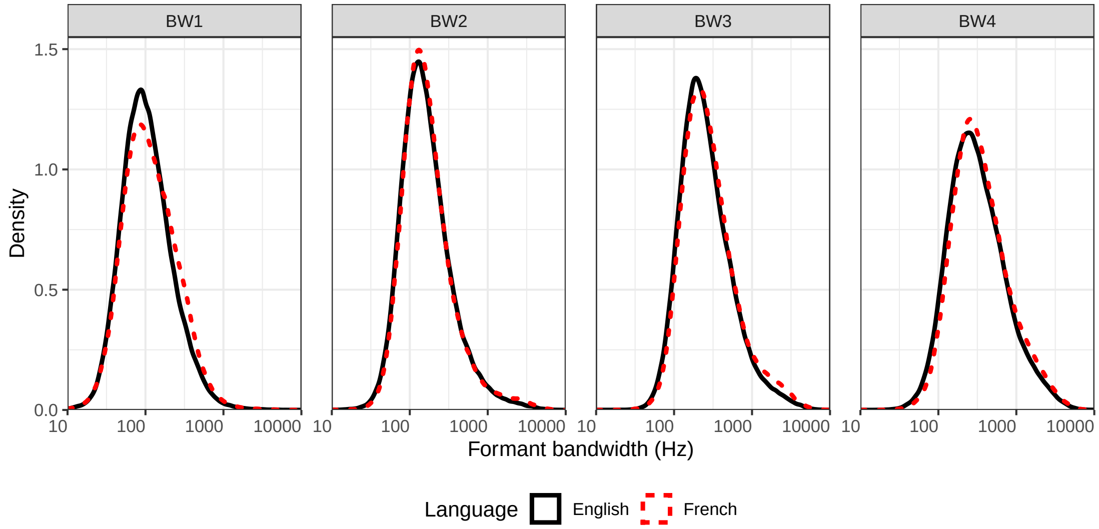

Figure 5.20: BW1–4 distributions in English and French, pooled across all 60 speakers.

peak, especially in the higher formants.

Figure 5.22 displays the means of log-transformed BW for each speaker in English and French, and the corresponding SDs by speaker are shown in Figure 5.23. Consistent with the overall patterns in Figure 5.21, the means of individual speakers vary within a relatively limited range for all formants. Although the cross-linguitic patterns in the BW distributions found above are subtle, the upward shift in French turns out to be remarkably consistent across speakers, as evident from the predominance of speakers situated above the line of equality in Figure 5.22. The degree of the shift is nonetheless subject to much individual variation, though considerably more so for BW2–4 than for BW1, which shows the strongest correlation between English and French out of all formants (Table 5.9). SDs of BWs similarly show moderate to strong correlations across languages, but unlike the means, SDs show no clear sign of systematic shifts, with the exception of BW1, where most speakers reported higher values in French, suggesting that BW1 distributions do not only shift to higher frequencies but also become more diffuse in French.

Table 5.8: Summary of fixed effects in best-fitting mixed-effects model for log10-transformed BW1–4. Intercept refers to grand mean. p values based on model comparisons with likelihood ratio tests.

	Estimate	SE	t	$p(\chi^2)$
BW1				
(Intercept)	2.06	0.012	175.72	–
Language	0.04	0.007	5.87	$<$.0001
Speaker L1				(.3189)
Language × L1				(.2484)
BW2				
(Intercept)	2.29	0.010	222.16	–
Language	0.03	0.009	3.08	.0028
Speaker L1				(.6504)
Language × L1				(.2293)
BW3				
(Intercept)	2.43	0.010	249.53	–
Language	0.04	0.009	5.00	$<$.0001
Speaker L1				(.1958)
Language × L1				(.4006)
BW4				
(Intercept)	2.52	0.010	253.19	–
Language	0.04	0.008	5.37	$<$.0001
Speaker L1				(.3211)
Language × L1				(.2836)

Table 5.9: Correlations between English and French (Pearson's r) for log-transformed BW speaker means and SDs ($p < .0001$ in all cases).

	BW1	BW2	BW3	BW4
Mean	0.85	0.65	0.67	0.71
SD	0.75	0.82	0.76	0.70

5.2.1 Discussion

Results from this part reveal systematic effects of language on LTFDs in bilingual speakers. In contrast to the analysis of /s/ in Section 5.1, speaker L1 was not found to have any effect on LTFDs. When compared to English, LTFDs in French exhibited a general shift towards higher frequencies regardless of the linguistic background of the speaker. This shift was found most prominently in LTF2, which may be attributed to a more crowded front vowel space in French, where pairs of unrounded and rounded front vowels can be found (see Section 4.7). Differences in vowel inventory may also be responsible for the higher formant bandwidths in French, especially for BW1 and BW2, as a consequence of the widening effect of nasal vowels on bandwidths (K. N. Stevens, 1998). It should be

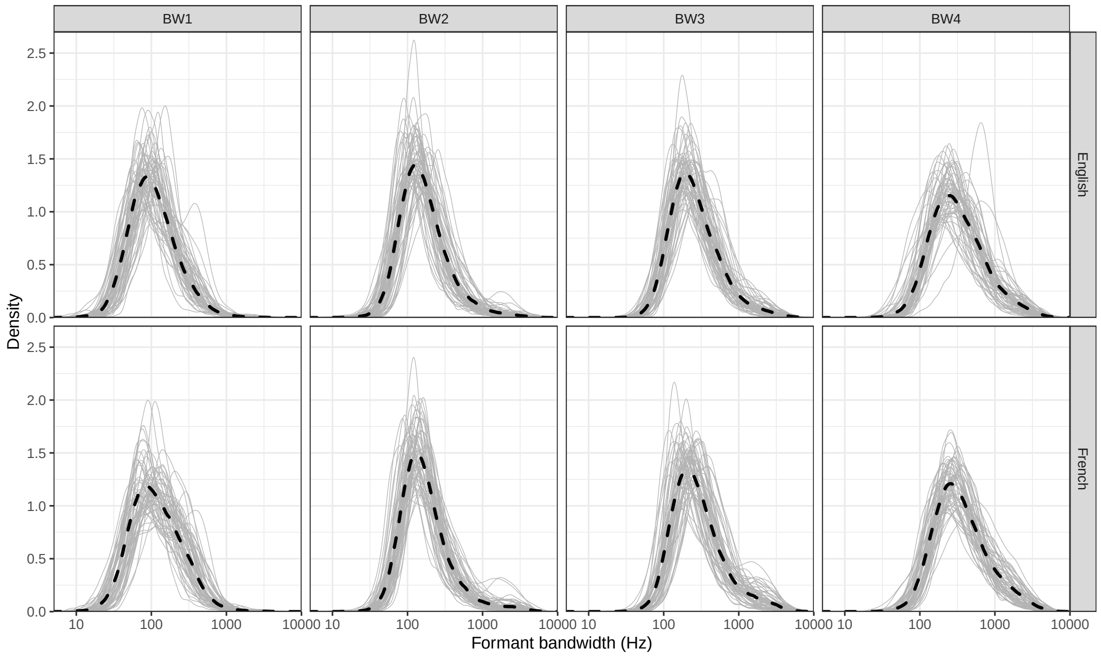

Figure 5.21: Individual (grey, solid) and overall (black, dashed) BW1–4 distributions in English and French.

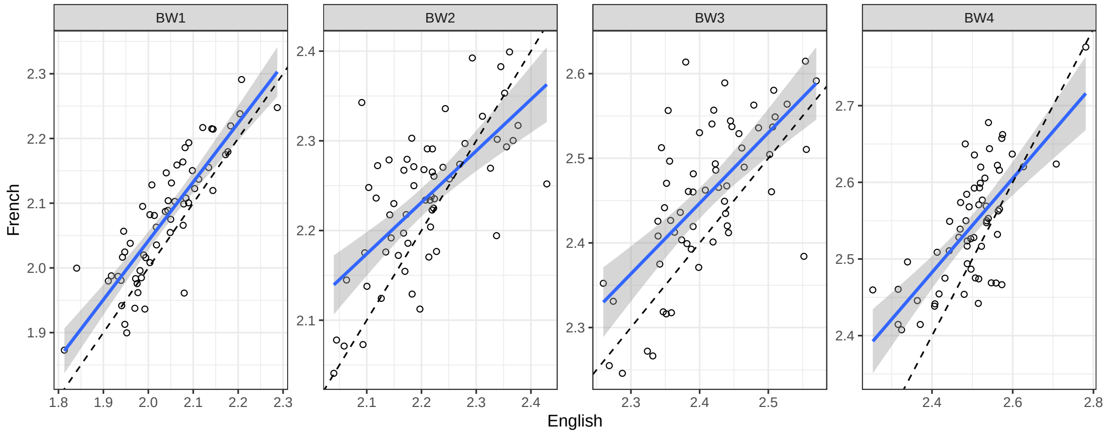

Figure 5.22: Means of log-transformed BW1–4 by speaker with best-fitting line (solid; 95% confidence intervals shaded) and line of equality (dashed).

noted that caution must be exercised when interpreting bandwidth measurements from LPC in Praat (and indeed other software for acoustic analysis), as they are much more susceptible to erroneous measurements than formant frequencies themselves (Burris et al., 2014).

Nevertheless, the factor of vowel inventory alone cannot fully account for the current findings. The upward shift in LTFDs was not limited to LTF2, but also extended to the higher formants, albeit in smaller magnitudes. Higher formants are generally considered

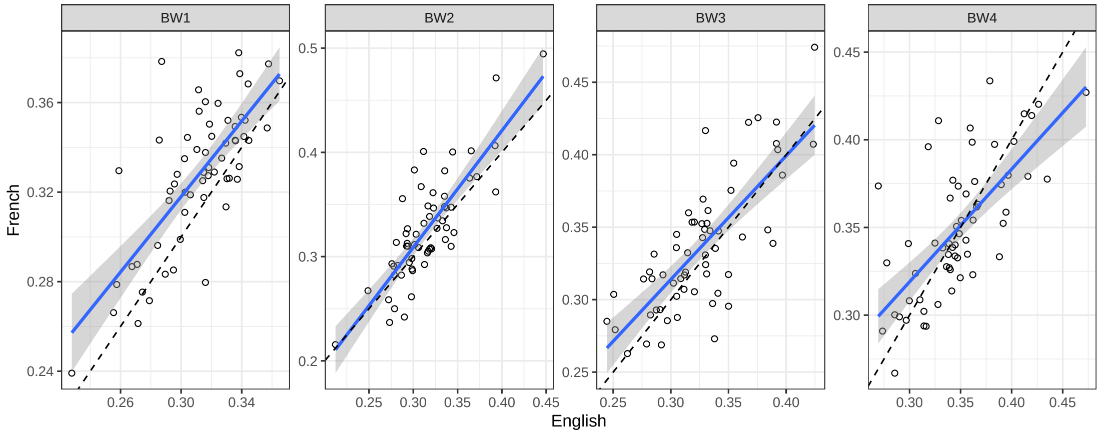

Figure 5.23: Standard deviations of log-transformed BW1–4 by speaker with best-fitting line (solid; 95% confidence intervals shaded) and line of equality (dashed).

to be less constrained to encode linguistic information than lower formants, but instead more indicative of a speaker's individual voice quality (Ladefoged & Johnson, 2015; McDougall, 2004; P. Rose, 2002). Therefore, it may be the case that bilingual speakers do make subtle changes in their setting of the supralaryngeal vocal tract across languages which are reflected in LTFDs here. This idea is supported in part by findings from Wilson and Gick (2014), who found distinct articulatory settings used by bilingual speakers of Quebecois French and Canadian English. Their findings, however, were limited to four speakers who were perceived as native-sounding in both languages, and the distinction was not manifested in the other four speakers who were perceived to be non-native in at least one of the languages in their study. It is unclear to what extent the same factor may have an effect on LTFDs in the current study, as a more direct comparison is prevented by the absence of accent ratings, but the remarkable consistency of the cross-linguistic patterns across languages, regardless of L1 background, would suggest that any role that being native-sounding has on the current findings is secondary to the effect of language itself. One reason for such a potential difference with Wilson and Gick (2014) may lie fundamentally in how the broad notion of articulatory setting is being measured, as Wilson and Gick (2014) operationalised articulatory setting as interspeech postures (ISPs), which refers to "the position of the articulators when they are motionless during interutterance pauses" (p. 361). Even though both ISPs and LTFDs aim to represent the general setting of articulators in a language, the former by definition focuses on their configuration in the absence of a sound signal, whereas the latter approaches this concept through the use of formants. While an underlying distinction in the articulatory settings will likely influence both ISPs and LTFDs, the formant-based nature of LTFDs means that it will also reflect the distribution of different sounds in the signal and any speaker-idiosyncratic aspects associated with such sounds.

When it comes to the speaker-specificity of LTFDs, the present findings indicate that it may depend on the particular formant in question. Whereas LTF3 and LTF4 are highly variable across speakers, in terms of both their central tendencies (peaks and means) and shapes, the same cannot be said of LTF1 and LTF2. This is unsurprising considering the differential distribution of linguistic and indexical information in lower versus higher formants mentioned above, especially since the current data controlled for the speech materials. Individual variation of formant bandwidths is highly constrained across the board.

5.3 Implications for FVC

The findings above offer much useful information about the speaker-discriminatory potentials of /s/ and LTFDs, within the specific context of cross-language comparisons as well as in FVC more generally.

Among static measurements for /s/, CoG exhibits both high between-speaker variability and (comparatively) low within-speaker variability in both languages. It is thus expected to perform relatively well as a parameter for speaker discrimination, at least in same-language comparisons. Between-speaker variability is also high for SD, but at the same time there is considerable variation within speakers, even within the same language. Skewness and kurtosis both show relatively high within-speaker variability and low between-speaker variability. As such, the discriminatory performance of SD is predicted to be not as good as CoG, and the higher spectral moments are expected to have poor discriminatory power. Cross-linguistically, the current findings indicate that the effectiveness of these parameters is unlikely to remain stable. As bilinguals commonly show greater within-speaker variability in English than in French for all spectral moments, it is suggested here that the same set of parameters will offer stronger performance in French than in English. When placed in the context of cross-language comparisons, the key finding that none of the midpoint measurements of spectral moments were found to be subject to cross-linguistic contrasts suggests *prima facie* that they would likely retain their usefulness in such cases. This is buttressed by the strong correlations between English and French demonstrated by the speakers, which indicate a high degree of within-speaker consistency for their acoustic targets across languages and further raise the possibility that individuals who are well discriminated in one language are likely well discriminated in the other language, meaning that individual performance may carry over to cross-language comparisons. However, the higher within-speaker variability in English than in French suggests that the overall or individual performance of spectral moments cannot be immune to the effects of language mismatch and will likely be adversely affected to some extent.

As for the acoustic dynamics of /s/, the analysis above shows that taking into account the trajectory of spectral moments can characterise the fricative more fully and provide more information about it as a whole. This is especially the case for CoG and SD, both of which follow a roughly parabolic trajectory, but less so for skewness, whose trajectory remains generally flat for most speakers. It is not evident, however, that the additional temporal dimension carries much more speaker-specific information on top of what static measurements can provide. The flatness of the random smooths in Figure 5.14, in particular for CoG and skewness, suggests that individual variation in shape is limited for these two spectral moments. Within the context of same-language comparisons, additional parameters relating to the curvature of the trajectory may only provide limited improvement in performance. These predictions also follow from Kavanagh (2012) and Smorenburg and Heeren (2020), both of which found that the inclusion of dynamic information resulted in only marginal improvement of the discriminatory performance of /s/, despite using different dynamic parameters. Kavanagh (2012) adopted a three-point approach, taking spectral moments from the onset, midpoint and offset of /s/, whereas Smorenburg and Heeren (2020) fitted quadratic polynomials to CoG trajectories and used the quadratic coefficients as input. Due to differences in what can be encoded in each of these approaches, they are not expected to contribute the same amount of information in cross-linguistic contexts. As the modelling above demonstrates, trajectories of spectral moments can differ in shape without an accompanying difference of the midpoint, and cross-linguistic dynamic differences can manifest over other parts of the trajectory. As the three-point approach is ill equipped to capture the acoustic detail between the onset and the midpoint, or between the midpoint and the offset, it is expected to be less sensitive to cross-linguistic differences than quadratic coefficients and consequently undergo less deterioration in performance. That said, systematic differences in the dynamics of CoG, SD and skewness indicate that their discriminatory performance, regardless of modelling technique, would likely be impacted to a greater extent than static measurements.

Turning finally to LTFDs, it is clear from the acoustic analysis above that their potentials for speaker-specificity differ considerably across formants. On the one hand, distributions of higher formants showed greater between-speaker variability than those of lower formants. On the other hand, LTF2 had the most diffuse distributions and hence high within-speaker variability. As such, while higher formants are expected to be stronger discriminants than lower formants (in line with findings from, e.g., Gold et al., 2013b), LTF2 in particular is predicted to perform poorly as a feature in FVC. As formant bandwidths display limited variation, their inclusion is not expected to lead to substantial improvement in the performance of LTFDs. Further, while speakers tend to show a high degree of consistency across languages, the current analysis found systematic differences between English and French for LTF2–4, as well as larger variation within speakers in English for LTF3. All BW distributions were similarly susceptible to the effect of language. It follows that, in cross-language comparisons, LTF2–4 would experience a greater deterioration of discriminatory power than LTF1, and including corresponding bandwidths in such cases would compound, rather than ameliorate, the effects of language mismatch.

Table 5.10: Full model summary of GAMM fitted to CoG trajectories in English and French. Note that *p*-values here are not used for significance testing.

Parametric terms	Estimate	*SE*	*t*	*p*
(Intercept)	5769.54	140.10	41.18	<.0001
Language = Fr	35.43	64.11	0.55	.5805
SpeakerL1 = EF	351.23	283.60	1.24	.2155
SpeakerL1 = F	179.14	170.99	1.05	.2948
PrevContext = RVs	10.45	33.22	0.32	.7531
PrevContext = Cs	78.91	23.79	3.32	.0009
PrevContext = #s	93.34	32.35	2.89	.0039
FollContext = sRV	−93.34	60.47	−1.54	.1227
FollContext = sC	−104.36	46.73	−2.23	.0255
FollContext = str	−75.00	77.69	−0.97	.3344
FollContext = s#	−97.64	46.10	−2.12	.0342
Smooth terms	edf	Ref.df	*F*	*p*
s(Window)	7.79	7.87	137.89	<.0001
s(Window):Language = Fr	2.90	2.99	106.27	<.0001
s(Window):SpeakerL1 = EF	1.07	1.09	0.01	.9412
s(Window):SpeakerL1 = F	2.41	2.52	3.59	.0123
s(Window):PrevContext = RVs	1.01	1.01	72.73	<.0001
s(Window):PrevContext = Cs	2.95	3.00	117.23	<.0001
s(Window):PrevContext = #s	2.68	2.92	22.36	<.0001
s(Window):FollContext = sRV	3.29	3.75	6.16	.0002
s(Window):FollContext = sC	3.91	3.99	200.10	<.0001
s(Window):FollContext = str	3.16	3.66	10.72	<.0001
s(Window):FollContext = s#	1.00	1.00	4.73	.0298
s(Duration)	2.91	3.74	1.94	.1052
ti(Window, Duration)	10.20	14.02	7.64	<.0001
s(Word)	93.15	108.00	20.98	<.0001
s(Window, Speaker)	367.78	534.00	37.60	<.0001

Table 5.11: Full model summary of GAMM fitted to SD trajectories in English and French.

Parametric terms	Estimate	*SE*	*t*	*p*
(Intercept)	1501.47	40.87	36.74	<.0001
Language = Fr	−19.59	21.47	−0.91	.3614
SpeakerL1 = EF	−161.57	80.33	−2.01	.0443
SpeakerL1 = F	58.69	47.69	1.23	.2184
PrevContext = RVs	35.50	13.50	2.63	.0086
PrevContext = Cs	25.64	9.72	2.64	.0083
PrevContext = #s	12.97	13.30	0.98	.3295
FollContext = sRV	128.13	22.36	5.73	<.0001
FollContext = sC	53.54	17.80	3.01	.0026
FollContext = str	−107.66	31.28	−3.44	.0006
FollContext = s#	−40.65	18.14	−2.24	.0250
Smooth terms	edf	Ref.df	*F*	*p*
s(Window)	7.06	7.43	35.19	<.0001
s(Window):Language = Fr	2.86	2.98	18.83	<.0001
s(Window):SpeakerL1 = EF	2.49	2.61	4.94	.0024
s(Window):SpeakerL1 = F	1.03	1.04	0.51	.4810
s(Window):PrevContext = RVs	1.00	1.00	25.28	<.0001
s(Window):PrevContext = Cs	1.03	1.06	10.70	.0010
s(Window):PrevContext = #s	2.28	2.66	11.80	<.0001
s(Window):FollContext = sRV	3.68	3.94	23.79	<.0001
s(Window):FollContext = sC	1.00	1.00	113.83	<.0001
s(Window):FollContext = str	2.81	3.36	4.48	.0029
s(Window):FollContext = s#	1.01	1.01	15.04	.0001
s(Duration)	4.91	6.12	6.77	<.0001
ti(Window, Duration)	7.46	11.48	6.58	<.0001
s(Word)	88.18	108.00	12.80	<.0001
s(Window, Speaker)	329.66	534.00	12.02	<.0001

Table 5.12: Full model summary of GAMM fitted to skewness trajectories in English and French.

Parametric terms	Estimate	*SE*	*t*	*p*
(Intercept)	0.66	0.11	6.05	<.0001
Language = Fr	−0.07	0.04	−1.74	.0819
SpeakerL1 = EF	−0.17	0.23	−0.75	.4552
SpeakerL1 = F	−0.26	0.14	−1.88	.0600
PrevContext = RVs	0.005	0.03	0.16	.8755
PrevContext = Cs	−0.03	0.02	−1.31	.1905
PrevContext = #s	0.005	0.03	0.14	.8906
FollContext = sRV	0.002	0.05	0.04	.9673
FollContext = sC	0.10	0.04	2.47	.0134
FollContext = str	0.24	0.08	3.24	.0012
FollContext = s#	0.26	0.04	6.26	<.0001
Smooth terms	edf	Ref.df	*F*	*p*
s(Window)	1.00	1.00	1.67	.1971
s(Window):Language = Fr	2.15	2.51	19.30	<.0001
s(Window):SpeakerL1 = EF	2.61	2.73	6.24	.0003
s(Window):SpeakerL1 = F	2.07	2.26	0.93	.2765
s(Window):PrevContext = RVs	2.59	2.87	20.49	<.0001
s(Window):PrevContext = Cs	1.21	1.38	0.29	.7437
s(Window):PrevContext = #s	2.29	2.66	18.11	<.0001
s(Window):FollContext = sRV	3.71	3.93	36.73	<.0001
s(Window):FollContext = sC	1.00	1.01	39.60	<.0001
s(Window):FollContext = str	1.00	1.01	2.90	.0887
s(Window):FollContext = s#	2.39	2.91	5.84	.0005
s(Duration)	3.75	4.74	3.09	.0098
ti(Window, Duration)	3.85	5.16	2.49	.0284
s(Word)	79.73	108.00	6.88	<.0001
s(Window, Speaker)	319.67	534.00	20.96	<.0001

Table 5.13: Full model summary of GAMM fitted to log-kurtosis trajectories in English and French.

Parametric terms	Estimate	*SE*	*t*	*p*
(Intercept)	1.48	0.05	27.67	<.0001
Language = Fr	−0.02	0.02	−0.89	.3738
SpeakerL1 = EF	0.05	0.11	0.48	.6336
SpeakerL1 = F	−0.13	0.06	−2.08	.0379
PrevContext = RVs	−0.03	0.02	−1.77	.0766
PrevContext = Cs	−0.03	0.01	−2.47	.0137
PrevContext = #s	−0.03	0.02	−1.71	.0875
FollContext = sRV	−0.11	0.03	−4.07	<.0001
FollContext = sC	−0.07	0.02	−3.08	.0021
FollContext = str	−0.05	0.04	−1.16	.2475
FollContext = s#	0.08	0.02	3.08	.0021
Smooth terms	edf	Ref.df	*F*	*p*
s(Window)	4.82	5.62	9.37	<.0001
s(Window):Language = Fr	2.73	2.93	3.86	.0058
s(Window):SpeakerL1 = EF	2.23	2.40	2.74	.0399
s(Window):SpeakerL1 = F	1.01	1.02	0.65	.4163
s(Window):PrevContext = RVs	1.00	1.01	3.10	.0785
s(Window):PrevContext = Cs	2.46	2.79	6.49	.0002
s(Window):PrevContext = #s	2.19	2.57	19.57	<.0001
s(Window):FollContext = sRV	3.70	3.94	10.77	<.0001
s(Window):FollContext = sC	3.35	3.75	4.07	.0079
s(Window):FollContext = str	1.00	1.00	8.52	.0035
s(Window):FollContext = s#	1.00	1.00	5.86	.0155
s(Duration)	1.00	1.00	0.42	.5166
ti(Window, Duration)	1.01	1.02	9.16	.0024
s(Word)	81.15	108.00	7.43	<.0001
s(Window, Speaker)	290.68	534.00	13.33	<.0001

Chapter 6

Methodology: LR-based System Testing

Building on the findings from the previous chapter, two sets of experiments were carried out on each variable to address the main research questions of the current study. In this chapter, the framework and general procedure for LR-based system testing and evaluation are first outlined in Sections 6.1. Details of the set-ups and specific methods used in each experiment are then described in Sections 6.2 and 6.3. The materials used are the same as those used in the phonetic part of the study and will not be repeated here. A summary of the materials and description of the speakers can be found in Sections 4.1 and 4.2.

6.1 Framework for system testing

In general terms, system testing refers to the process of using one or more corpora of existing recordings, either taken off the shelf or specifically collected, where the identity of the speaker in each recording is already known, and conducting comparisons between pairs of speakers in the corpora, using a particular set of features of interest, so that the outcomes of such comparisons can be assessed against the "ground truth" to evaluate the validity of the chosen features in the specific conditions of the recordings in question. To carry out system testing within the LR framework, the current study followed the two-stage process set out in Morrison (2013; see also Morrison et al., 2021).

The whole set of 60 speakers were first randomly partitioned into three sets: *test*, *training* and *background*, each made up of 20 speakers. Dividing speakers into separate sets ensured that the background set, which was then used to model the relevant population, did not contain the known speaker in any comparison, as the nature of the defence's hypothesis (that the QS is not produced by the known speaker) is such that it would be fallacious to include the known speaker in the background set. This practice would not be possible when the number of speakers available in the database is too low, as test, training and background sets that are too small result in highly unstable LRs and, consequently, lower system reliability (Hughes, 2017). In such cases, cross-validation may be applied to the calculation and calibration of LRs, which allows the same speakers to simultaneously act as test, training and background data. While larger sample sizes can further increase precision (Hughes, 2017), the current set of 60 speakers, common in FVC research, is considered to be sufficiently large to be divided into independent test, training and background sets to yield relatively reliable results.

In the *feature-to-score* stage, comparisons were first conducted for all speaker-pairs in the test set, with each speaker acting in turn as the questioned and known speaker for all speakers, resulting in 20 same-speaker (SS) and 380 different-speaker (DS) comparisons. In each comparison, data from the KS and from the data in the background set were modelled using appropriate statistical procedures, such that data from the QS could be evaluated against each of them. Two standard procedures used in acoustic-phonetic FVC (Morrison, 2011) are the multivariate kernel density (MVKD) approach (C. G. G. Aitken & Lucy, 2004), which was originally applied to forensic evidence collected from glass fragments, and the Gaussian mixture model–Universal background model (GMM–UBM) approach (Reynolds et al., 2000). Following standard practice for each set of features, in the current study, MVKD was used to analyse data from the fricative /s/, whereas the GMM–UBM approach was adopted for LTFDs (see Sections 6.2.1 and 6.2.2 for further rationale behind and details of each approach). These comparisons resulted in a total of $20 + 380 = 400$ LR-like scores. The same process was then applied to the training set in order to generate a set of training scores.

In the *score-to-LR* stage, the test scores were calibrated, or converted, to produce interpretable LRs, or more commonly log-transformed LRs (LLRs), for the strength of evidence to be assessed. LLR $>$ 0 in a comparison indicates support for the same-speaker hypothesis, while LLR $<$ 0 indicates support for the different-speaker hypothesis. As such, negative LLRs in SS comparisons and positive LLRs in DS comparisons are considered to be contrary-to-fact errors. Different techniques for converting scores to LRs have been put forward, of which the logistic regression procedure is most commonly adopted in LR-based FVC. Other proposals for score conversion include linear discriminant analysis (LDA), Bayesian modelling and the use of an empirical lower and upper bound (Vergeer et al., 2016), all of which have been evaluated in Morrison and Poh (2018) for their effectiveness in avoiding overstating the strength of the evidence and in Wang and Hughes (2021) for their sensitivity to sample size and sampling variability. Following the recommendations in Morrison (2011) and Morrison and Poh (2018), the current study applied a regularised version of the logistic regression procedure implemented in Morrison (2011), in order to avoid numerical issues that may arise from complete separation of SS and DS score distributions in the training set.

In statistics more generally, logistic regression is commonly used to model outcome variables with binary outcomes. In the context of FVC, this form of regression is thus suited to model the relationship between scores and the type of comparison that gives rise to them, which is either between the same speaker or between different speakers. A probability curve is fitted to scores calculated from the training data, with parameters estimated via maximum likelihood estimation, to model the odds of any score value as a same-speaker or different-speaker conclusion, such that the probability of a score being classified as from an SS comparison monotonically increases. In a way, this method is equivalent to the LDA procedure, in the sense that the two can be mathematically mapped to the other, but it is more robust than the latter to violations of the assumption of equal variance in the scores. While LDA seeks to directly model the same-speaker and different-speaker score distributions using single Gaussian curves with equal variances, so that calibrated LRs can be calculated from the ratio of the probabilities of scores on each distribution, logistic regression does not seek to model the score distributions directly but the boundary between the two categories of same-speaker and different-speaker conclusions (Morrison, 2013). In the regularised version of the procedure implemented here, an extra copy of each score is included, each of these copies being assigned the opposite type of comparison and given an extremely small weight (0.001), in order to overcome potential computational issues without inducing substantial shrinkage. Larger weights may be attached to the extra copies if the aim is to induce score shrinkage such that the (quantitative) strength of evidence is not overstated (Morrison & Poh, 2018). Coefficients from the model obtained are then applied to transform the test scores into LRs via translation (linear shifting) and scaling, with the goal of minimising the log likelihood ratio cost function (see Section 6.1.1 below).

As system validity is subject to variation due to effects of speaker sampling, especially from the membership of the test set (Wang et al., 2019b), the whole sampling and testing procedure was repeated 100 times to minimise the effects of random speaker sampling, as well as to ensure that all 60 speakers were compared with one another. It should be noted that, as all replications drew from the same set of 60 speakers, they were not independent of each other, and as such would likely underestimate the range of variability when compared to fully independent samples. In the current study, the composition of the test, training and background sets in each of the 100 replications was randomly pre-generated, and the same make-up was applied to all system testing in Experiments 1 and 2, in order to eliminate variability between systems due to effects of speaker sampling and to ensure maximal comparability of LRs and validity metrics across the board.

In this study, system testing was carried out in R via a custom-built package (Lo, 2021), which implements the framework and different modelling approaches outlined above. Specific implementations of MVKD and GMM–UBM are described in Sections 6.2.1 and 6.2.2.

6.1.1 Performance evaluation

In numerical LR-based FVC, system evaluation is predominantly conducted on the global level, where the strength of performance is indicated by means of a single metric score. The most commonly used metrics are the equal error rate (EER) and the log likelihood ratio cost function (C_{llr}; Brümmer & du Preez, 2006). The EER identifies the percentage of comparisons with contrary-to-fact outcomes, at a score threshold where the rate of misses (same speaker identified as different) is equal to the rate of false matches (different speakers identified as the same). On the other hand, C_{llr}, calculated from Equation 6.1, takes into account the magnitude of errors, such that contrary-to-fact LRs of a larger magnitude incur higher penalty and, in turn, higher C_{llr}. A C_{llr} that is greater than 1 indicates that the system is poorly calibrated and provides no useful speaker-discriminatory information. For both of these metrics, stronger system performance is indicated by smaller values. The closer they are to 0, the better the system is judged to be performing.

$$C_{\text{llr}} = \frac{1}{2}\left[\frac{1}{N_{i=j}}\sum_{i=j} log_2(1+\frac{1}{LR_{ij}}) + \frac{1}{N_{i\neq j}}\sum_{i\neq j} log_2(1+LR_{ij})\right] \tag{6.1}$$

where $i, j = i, j$-th speaker,

$LR_{ij} =$ LR from speaker comparison with i as suspect and j as offender,

$N_{i=j}, N_{i\neq j} =$ number of same-speaker $(i = j)$ and different-speaker$(i \neq j)$ comparisons.

Other graphical means of evaluation, such as receiver operating characteristic (ROC) curves, detection error tradeoff (DET) curves and Tippett plots, are also often employed to provide more information about the overall performance of the system (see Morrison & Enzinger, 2016). Commonly used in the assessment of ASR systems, both ROC and DET curves track the rate of false matches against the rate of misses across a range of acceptance thresholds in a single curve, whereas Tippett plots trace performance of SS and DS comparisons in separate curves to visualise the cumulative distribution of scores (or LRs) in each type of comparison.

As much as the above metrics and graphs can illustrate the global level of system performance, their diagnostic value can be limited. A more microscopic view of system performance, beyond rates and sizes of error, can be obtained by examining the performance of the system on individual speakers. By analysing in detail the speakers for which the system performs exceptionally well or disproportionately contribute to errors, researchers can gain insights into the nature of the errors in the system and work towards improving system design in a targeted manner. Further, identifying individual voices who are difficult to match against any speakers in the same database may be helpful for optimising homogeneity within forensic databases (San Segundo et al., 2017).

To date, analysis of individual performance is rarely performed in the context of FVC, and the causes behind speakers being identified as outliers within any tested system are still underexplored. In addition to technical properties of the audio samples, such as non-uniform duration and acoustic quality (Nash, 2019), variation of individual performance in any given system can arise due to physiological and behavioural reasons, as well as the impact of these factors on the quality of data capture (Dunstone & Yager, 2009). f_0, for example, is susceptible to variation due to physiological factors in both the short and the long term (Braun, 1995; Rhodes, 2012). In terms of behaviour, speakers may seek to disguise their voice. They may also shout at such a volume that causes clipping in the recording, thus impacting data capture. A close analysis of LRs from individual speakers and the corresponding speech data used to generate those LRs may thus not only be useful in diagnosing the sources of variation in individual performance, but also more generally help understand the relationship between input and output in numerical LR-based testing.

Within the context of ASR, Alexander et al. (2014) has proposed preliminary links between individual performance and aspects of voice quality. However, subjective judgments of voice quality are far removed from cepstral coefficients, the highly abstract acoustic features used in ASR systems, and do not encode the same speaker-specific information (French et al., 2015). The connection between individual LR performance and the input data thus warrants further investigation.

In order to examine how the performance of bilingual speakers varies in different languages and in cross-language comparisons, as well as the connection between their performance and the underlying speech data, the current study evaluated performance not only on the global level, using EER and C_{llr}, but also on the individual level, with the aid of zooplots.

Zooplots are a diagnostic tool used in the field of biometrics to visualise individual user performance. The zooplot is built upon the idea of a “biometric menagerie”, developed by Doddington et al. (1998) and later on expanded by Dunstone and Yager (2009), where speakers are classified into user groups or animals based on their individual performance. In a zooplot, each speaker’s performance in SS comparisons (or genuine performance) is plotted against their performance in DS comparisons (or imposter performance). The use of zooplots thus facilitates the identification of problematic speakers in the database for further analysis and diagnosis. Additionally, the distribution of speakers in a zooplot can be indicative of systematic weaknesses in the algorithm or the database used. A predominance of speakers who perform well in DS comparisons but poorly in SS comparisons, for example, may be an outcome of poor-quality enrolment in the database (Dunstone & Yager, 2009).

In their original formulation, Doddington et al. (1998) distinguish a default group of speakers, *sheep*, from other speakers who tend to contribute disproportionately to system errors. These animal groups include *goats*, whose voices are particularly difficult to match and hence likely produce errors in SS comparisons; *lambs*, who may disproportionately account for false matches due to their voices being easily imitable; and *wolves*, whose voices may easily imitate others’ and thus also contribute to false matches. Dunstone and Yager (2009) introduces a set of relational animals, which are defined by the relationship between a speaker’s performance in SS comparisons and in DS comparisons, rather than their performance in a single type of comparisons. Described in FVC terms, these groups include:

- *doves*, who perform relatively well in both types of comparisons;
- *worms*, who perform relatively poorly in both types of comparisons;
- *phantoms*, whose voice characteristics are difficult to match against any speaker and so who perform well in DS comparisons but poorly in SS comparisons; and
- *chameleons*, who can be easily matched with (or camouflage as) any speaker and thus perform well in SS comparisons but not in DS comparisons.

Other researchers have since sought to refine the zooplot to enrich the information related to speaker performance that can be presented. Instead of representing speakers as uniformly sized points, Alexander et al. (2014) proposes to depict each speaker on the zooplot as an ellipse, with the length of each axis denoting the score variability for the speaker in each type of comparison. Speakers can then be dubbed “tall” or “short”, and "fat" or "thin", according to the shape of their ellipse. As this additional dimension highlights issues concerning the variability of scores for a speaker, it is claimed to provide diagnostic information independent of the speaker's location in the zooplot.

To enable the visualisation of specific speaker-pairs whose comparisons give rise to a speaker's individual performance, Giot et al. (2016) incorporates a graph-theoretic element to their implementation of the zooplot, known as a "zoo graph". In this design, speakers are not merely isolated points on a scatterplot, but instead function as nodes that can be linked with other nodes by directed edges. Given any score threshold, the presence of an edge from speaker A to speaker B indicates that, on average, speaker A is recognised by the system as speaker B at that threshold. In a zoo graph, the optimal configuration would be where speakers are only connected with themselves but not with any other speakers. As such, the identification of problematic speakers in the system can be accompanied by further analysis of the individuals who are the sources of errors for them. The inclusion of such relationships in the graph, however, means that there may be a large number of edges, rendering it highly cluttered and unwieldy to use.

As the present exploration focuses on the stability of the performance of individual speakers in different languages and different types of comparisons, as well as the relationship between individual performance and the underlying acoustic data, the high level of complexity introduced by the zoo graph was considered to be undesirable for this study. Similarly, while the additional information conveyed by tallness and thinness as proposed by Alexander et al. (2014) is no doubt valuable, the focus here is on the speakers' relative position, rather than the overall variability of their LRs. As such, the "original" zooplot by Dunstone and Yager (2009) was chosen as the form of visualisation for the individual-level analysis.

6.1.1.1 Individual-level analysis

The zooplot for any particular system was constructed by plotting a speaker's average performance in DS comparisons against their performance in SS comparisons, where speakers with stronger performance were positioned towards the top and the right of the plot. In this study, average performance of any speaker is defined as the arithmetic mean of LLRs from all SS or DS comparisons across all 100 replications involving that speaker.

Individual-level analysis of LR performance was conducted in three ways. First, the overall distribution of speakers on the zooplots was analysed, in terms of both the absolute values of LLRs, as they have been calibrated and can be directly interpreted, and the relative positions of speakers. Second, speakers with outlying performance in each system were identified and categorised into one of the four relational animal groups defined above. Following Dunstone and Yager (2009), *doves* are defined as speakers whose average performance is within the best 25% of all speakers in both types of comparisons. This group can therefore be located in the top right corner of the zooplot, as they comprise speakers with the *highest*, most positive mean LLR in SS comparisons (SS-LLR) and the *lowest*, most negative LLR in DS comparisons (DS-LLR). Worms, phantoms and phantoms are analogously defined, as outlined in Table 6.1. Speakers whose mean LLR lies between the top (or bottom) 25% to 30% were additionally identified near-members of animal groups to mitigate the cliff-edge effect of borderline cases documented in O'Connor et al. (2015). Membership of relational groups was then analysed in each system and compared across different systems. In any given system, it is naturally expected that some speakers would fall within each relational group. A particularly high or low number of speakers in these groups, however, not only reflect performance issues of those individuals, but can also point to an overall lack of independence between a speaker's performance in SS and DS comparisons, and thus be indicative of more systemic issues, such as the choice of parameters or modelling techniques. As Dunstone and Yager (2009) describe, under the null hypothesis that performance in different types of comparisons is independent, the probability that a speaker falls within a particular relational group is $(\frac{1}{4})^2 = \frac{1}{16}$. The probability that there are at least x speakers in a relational group within a population of n speakers is given by $\sum_{i=x}^{n} C_i^n(\frac{1}{16})^i(\frac{15}{16})^{n-i}$, while the probability that there are at most x speakers in a group is $\sum_{i=0}^{x} C_i^n(\frac{1}{16})^i(\frac{15}{16})^{n-i}$. The null hypothesis is rejected if the resultant probability is below $\frac{\alpha}{2}$ (as the hypothesis is two-tailed and non-directional). In the current study, the background population consists of 60 speakers, which means that a minimum of nine members indicates the significant presence of a particular group ($p = .0118$), while a significant absence is only supported when no speakers fall into it ($p = .0208$). The correlation between speakers' performance in SS and DS comparisons was further evaluated by calculating the Pearson's correlation coefficient of their SS-LLR and DS-LLR in each system. As stronger performance is indexed by more positive SS-LLR and more negative DS-LLR, stronger positive correlation of SS and DS performance is indicated by a more negative correlation coefficient. Third, to explore the relationship between LR performance and speech production, the corresponding acoustic data of speakers classified as members of relational groups were compared with those of the other speakers.

	Mean SS-LLR	Mean DS-LLR	Location
Doves	Highest 25%	Lowest 25%	Top right
Worms	Lowest 25%	Highest 25%	Bottom left
Phantoms	Lowest 25%	Lowest 25%	Top left
Chameleons	Highest 25%	Highest 25%	Bottom right

Table 6.1: Inclusion criteria for doves, worms, phantoms and chameleons and their locations in zooplots.

6.2 Experiment 1: Same language comparisons

The first set of experiments addresses the question of whether the discriminatory potential of the same features is maintained across English and French when there is no language mismatch. System testing was conducted following the procedure described in Section 6.1, where both QS and KS were in the same language. As only a single recording is available for each speaker, the first half of the recording was taken to be the KS, and the second half was taken to be the QS. The use of contemporaneous samples here is likely to lead to more optimistic performance than the use of non-contemporaneous samples (which is necessarily the case in forensic materials) obtained from separate recording sessions, due to a more limited range of within-speaker variation in the former scenario. Given that this bias applies to all systems tested in this experiment, it is not considered to be a major limitation, as the comparison here is focused on any differences between the two languages.

6.2.1 /s/

For the fricative /s/, system configuration was varied along three dimensions: (1) the combination of acoustic parameters; (2) the representation of the acoustic parameters, or the mode of data input; and (3) language.

The acoustic parameters used for testing included all four spectral moments: CoG, SD, skewness and kurtosis. As in the acoustic analysis in Chapter 4, kurtosis was log-transformed to correct for the positive skew in its distribution, in order for its use in the MVKD formula (see below) to be appropriate. Within each language, system testing permuted through all 15 possible combinations of spectral moments by including or excluding each of them.

Additionally, both static and dynamic representations were tested to explore whether a dynamic input would improve the performance of the relevant acoustic parameters over static (midpoint) input. Using the full set of measurements for each token from all nine windows as input for the MVKD formula would provide the highest level of detail in representing /s/ acoustics, but the MVKD formula may encounter computational issues and provide suboptimal performance when it is fed a large number of parameters, resulting in significant misestimation of LRs. Additionally, as spectral moments, like vowel formants, within the same token are interdependent, a maximal representation of spectral dynamics using all nine sets of measurements does not necessarily add useful speaker-specific information beyond what a reduced representation can achieve (McDougall, 2004). Instead, following Kavanagh (2012) and Smorenburg and Heeren (2020), two different representations of dynamic input were tested: a three-point approach, using the onset (window 1), midpoint (window 5) and offset (window 9); and a polynomial approach, in which the nine sets of measurements from each token were fitted with a polynomial curve to model each spectral moment. The latter approach is also commonly used in FVC to model vowel formant dynamics (McDougall, 2006). Informed by the GAMM models in Section 5.1.2, which suggest that spectral moments, especially CoG and SD, follow a roughly parabolic trajectory over time, each parameter was fitted with a quadratic polynomial, in the form of $M(t) = c_0+c_1t+c_2t^2+\varepsilon$, where $M =$ measured spectral moment, $t =$ centred time-step, and ε = random residual error. The coefficients c_0, c_1 and c_2 were then used in place of the original measurements as input (McDougall, 2006). In both approaches, the number of parameters required to represent each spectral moment was reduced to only three, and the maximum number of parameters used as input was reduced from $9 \times 4 = 36$ to $3 \times 4 = 12$.

In the *feature-to-score* stage, the MVKD formula (C. G. G. Aitken & Lucy, 2004), given in Equation 6.2, was used to model the acoustic data and compute an uncalibrated score for each comparison. MVKD assumes a normal distribution for the suspect speaker (hence the need to log-transform kurtosis), but models the background data with a Gaussian kernel density model. As such, MVKD is capable of accounting for correlation between input parameters, such as that between different spectral moments in /s/. The formula was implemented in R via an adaptation of Morrison (2007)'s MATLAB implementation.

$$Score = \frac{p(E|H_p)}{p(E|H_d)} \tag{6.2}$$

$$\begin{aligned} p(E|H_p) = {} & (2\pi)^{-p}|D_Q|^{-\frac{1}{2}}|D_K|^{-\frac{1}{2}}|C|^{-\frac{1}{2}}(mh^p)^{-1}|{D_Q}^{-1} + {D_K}^{-1} + (h^2C)^{-1}|^{-\frac{1}{2}} \\ & \times e^{-\frac{1}{2}(\mu_Q-\mu_K)^T(D_Q+D_K)^{-1}(\mu_Q-\mu_K)} \\ & \times \sum_{i=1}^{m} e^{-\frac{1}{2}(\mu^*-\mu_{B_i})^T[({D_Q}^{-1}+{D_K}^{-1})^{-1}+h^2C]^{-1}(\mu^*-\mu_{B_i})} \end{aligned}$$

$$
\begin{aligned}
p(E|H_d) = {} & (2\pi)^{-p}|C|^{-1}(mh^p)^{-2} \\
& \times |D_Q|^{-\frac{1}{2}}|D_Q{}^{-1} + (h^2C)^{-1}|^{-\frac{1}{2}} \sum_{i=1}^{m} e^{-\frac{1}{2}(\mu_Q-\mu_{B_i})^T(D_Q+h^2C)^{-1}(\mu_Q-\mu_{B_i})} \\
& \times |D_K|^{-\frac{1}{2}}|D_K{}^{-1} + (h^2C)^{-1}|^{-\frac{1}{2}} \sum_{i=1}^{m} e^{-\frac{1}{2}(\mu_K-\mu_{B_i})^T(D_K+h^2C)^{-1}(\mu_K-\mu_{B_i})}
\end{aligned}
$$

$$
\begin{aligned}
\text{where } Q, K &= \text{questioned speaker and known speaker,} \\
D &= \text{speaker variance-covariance matrix,} \\
C &= \text{between-speaker variance-covariance matrix,} \\
B_i &= i\text{-th speaker in the reference (background) set,} \\
m &= \text{number of reference (background) speakers,} \\
p &= \text{number of parameters,} \\
h &= \text{optimal smoothing parameter} = \left[\frac{4}{m(2p+1)}\right]^{\frac{1}{p+4}}, \\
\mu &= \text{speaker mean vector,} \\
\mu^* &= (D_Q{}^{-1} + D_K{}^{-1})^{-1}(D_Q{}^{-1}\mu_Q + D_K{}^{-1}\mu_K).
\end{aligned}
$$

Two separate sets of LMEMs were fitted to C_{llr} and EER obtained from all 100 replications of the tested systems to evaluate the effect of language on the discriminatory power of individual spectral moments and of combinations of multiple spectral moments. In the first set, LMEMs were fitted to systems using individual spectral moments with C_{llr} or EER as the outcome variable. The effect-coded factors of language, spectral moment and mode of data input, as well as all of their interactions were included in the initial full model as fixed factors, while random intercepts were included for each replication. In the second set, LMEMs were fitted to C_{llr} and EEM from all tested systems, grouped by number of spectral moments included (1–4). The number of spectral moments (treated as a categorical rather than numeric variable), mode of data input and language were included as fixed predictors, with random by-replication intercepts once again included. As in the acoustic-phonetic analysis (Section 4.4.4), a step-down approach was used to evaluate the statistical significance of the fixed predictors, where a reduced model was formed by excluding each predictor in turn from the full model, starting with the highest-order interaction, and compared with the full model using LRTs.

Individual-level analysis was mainly limited to the subset of systems which only used one of the spectral moments as input, in order to facilitate one-to-one comparison with the acoustic data. Systems using a combination of all four spectral moments were also subjected to individual analysis to explore the effects of combining acoustic parameters on individual performance.

6.2.2 LTFDs

Testing for LTFDs followed the same procedure, where system configuration was similarly varied along three dimensions: (1) the combination of acoustic parameters; (2) the mode of input, namely whether formant bandwidths were included; and (3) language. In this case, the acoustic parameters included F1–4 and BW1–4 estimates. In a similar vein to the testing of /s/, all formant combinations were tested. Each combination comprised two modes of input: F-only, where only formant centre frequencies were used, and F+BW, where corresponding bandwidth estimates were also included.

In the *feature-to-score* stage, previous research has applied two different approaches to modelling LTFD data. Following Moos (2010), Gold et al. (2013b) and Tomić and French (2019) both package raw formant measurements into "local LTFDs" by averaging consecutive formant measurements over short periods of time and subjected the resultant "local LTFDs" to an MVKD-based analysis. By modelling "local LTFDs" instead of the raw formant measurements, this method substantially reduces the number of data points and is indeed intended to reduce the variability within the data. As each "local LTFD" is averaged over multiple vowel tokens, it incurs loss of substantial amount of information, as formant measurements pertaining to individual vowels become unavailable for use. A more standard approach, adopted in the current study, is to apply the GMM–UBM approach as described in Reynolds et al. (2000; e.g., Becker et al., 2008, 2009; Hughes et al., 2017), in which the relatively large amount of raw measurements available from LTFDs are taken advantage of to model the shape of the whole distribution. Data were modelled and compared using the GMM–UBM approach, implemented by means of the *mclust* package in R (Scrucca et al., 2016). The reference population was modelled with a UBM, a GMM composed of 12 Gaussians calculated using data pooled from all 20 speakers in the background set. In each comparison, a GMM for the known speaker was derived by using data from the KS to adapt the UBM by means of maximum a posteriori estimation, and the LLR-like output score of the comparison was calculated by Equation 6.3. These scores were then converted to LRs using the procedure set out in Section 6.1 above.

$$Score = \frac{1}{N}\sum_{i=1}^{N} log_{10}\frac{p(x_i|S)}{p(x_i|B)} \tag{6.3}$$

$$\begin{aligned} \text{where } x_1, x_2, \ldots, x_N &= \text{each set of input data from the QS,} \\ S &= \text{suspect GMM,} \\ B &= \text{background GMM.} \end{aligned}$$

Statistical evaluation of the effect of language, among other factors, on validity metrics was carried out in a manner analogous to that described in the previous section, with one set of models fitted to C_{llr} and EER from systems using only individual LTFs (including the predictors language, LTF and mode of data input) and another set fitted to C_{llr} and EER from all systems (including the predictors language, number of LTFs and mode of data input).

6.2.3 ASR

ASR testing was conducted using the software *Phonexia Voice Inspector v4.0* (henceforth "Phonexia"). In addition to a general voice comparison mode, Phonexia offers an evaluative functionality that performs system validation using user-provided data, providing the metrics of EER and C_{llr}, as well as a set of visualisations, such as the DET plot and Tippett plot, to illustrate the results. However, a crucial requirement for using this functionality, namely that at least three recordings should be available for each speaker, is not met in the present case. As such, the current testing was performed within the general comparison interface.

After being low-pass filtered at 8 kHz, all samples were manually added through the case manager, with the first half of each speaker's recording submitted as the KS ("suspected reference speaker" in Phonexia), and the second half of the recorded materials as "Questioned recordings". Each KS was then manually selected in the graphical user interface to generate a score table consisting of the output scores from comparisons between the KS and each QS, resulting in a total of $60 \times 60 = 3{,}600$ scores. For reasons described below, the software-internal score-to-LR conversion was not used. The scores were instead taken directly from the software and calibrated to become LLRs using the same logistic regression procedure as in the case of other linguistic-phonetic variables.

The architecture of the system is described in detail in Jessen et al. (2019), but a summary is given here. In this version of Phonexia, samples under comparison are first processed through a four-stage process of voice activity detection. An x-vector approach is adopted for the *feature extraction* stage, which involves the use of a deep neural network (DNN) frontend to process the extracted MFCCs. 20 MFCCs are extracted every 1 ms, with 24 mel filters between 64 and 3800 Hz. Mean and variance are calculated from a window of 3 s around each frame for normalisation. Notably, unlike previous versions, no deltas, double deltas or any f_0-based features are used. The extracted MFCCs are fed directly to a first DNN, trained to perform speaker classification. Speaker representation is obtained from one of the internal layers of the DNN and summarised over the entire sample. Summary statistics, in the form of mean and variance, are then passed on to be a second DNN, and speaker representation is once again obtained from an internal layer to model each speaker for the purpose of comparison.

The core *comparison* stage is carried out by means of probabilistic linear discriminant analysis (PLDA), which does not directly model the speaker in the KS for comparison with the QS, but instead arrives at the output score by comparing the probability of obtaining the x-vectors from the KS and the QS, assuming the speakers in the samples were the same randomly selected person from the relevant population, with the probability of obtaining those feature vectors if the speakers were different randomly selected people from the relevant population (Morrison et al., 2020). At this stage, comparison is done using the software-internal pretrained models as reference. Any case-specific background set (known as "population set" in Phonexia) that formed the reference population in Sections 6.2.1 and 6.2.2 above is thus not yet brought into consideration. The output of the comparison is an uncalibrated log-transformed score (called "Evidence" in Phonexia).

Where a population set is supplied by the user, scores would automatically undergo *calibration*, such that Phonexia also reports a corresponding LLR for each comparison. Different modes are available for determining how the final LLR is derived, but the basic principle of the conversion remains the same, which involves an LDA procedure, using a Gaussian (normal) curve each to estimate the SS (target) and DS (non-target) score distributions, and then comparing the probability of the Evidence (score) in the two distributions. If the population set is chosen as the basis of the target score distribution, then all possible SS comparisons are conducted within the set to produce the target scores. In that case, Phonexia will also use the population set as the basis of the non-target score distribution. If a minimum of three recordings is available for the suspected reference speaker, then it is possible to generate the target scores from only the specific individual. The non-target score distribution can then be chosen to be based on comparing the population set with the QS or with the suspected reference speaker, depending on the presence of condition mismatch between the provided samples.

In the current study, it was not possible to make use of the option to base the target score distribution on the suspected speaker, as there was only one recording per language for each speaker. The only remaining software-internal calibration option was to set the target and non-target score distributions to be both derived from the population set, but its use would also be problematic for two reasons. First, it would mean that a different method of calibration would be used for this part of the experiment from that for other variables, thus reducing the comparability between the sets of results. Second, calibration should be based on comparison scores that are reflective of the specific case circumstances. While this is unproblematic in Experiment 1, in cases of language mismatch, to be tested in Experiment 2, this could not be possibly accomplished using this option, as both target and non-target score distributions used to calibrate the system output should themselves be derived from cross-language comparisons. On the one hand, if the samples making up the population set were all in the same language, the scores used for calibration would be derived from same-language rather than cross-language comparisons. On the other hand, if the population set comprised at least an English recording and a French recording from each speaker, exhaustive comparisons of all recordings within the set would necessarily incur a mixture of same-language and cross-language comparisons in the non-target score distributions.

Therefore, as mentioned above, the uncalibrated scores were separately processed outside Phonexia through logistic regression calibration. For each of the 100 replications, scores resulting from comparisons among the 20 speakers in the training set were used to generate the calibration coefficients, which were used to calibrate the scores from the 20 speakers in the test set. Due to the use of the pretrained models instead of case-specific reference data in the comparison stage, the background set was in fact rendered irrelevant in this procedure.

Statistical analysis of the effect of language on system performance was performed in the same fashion as previously described. In this case, however, only one set of models were fitted to validity metrics, with language as the sole predictor of fixed effects and random by-replication intercepts included, as systems did not vary in any other dimension.

6.3 Experiment 2: Cross-language comparisons

The second set of experiments addresses the validity of cross-language comparisons in a bilingual population and the question of what effects language mismatch has on system and individual performance of different variables. To focus on the effect of language mismatch between samples, the relevant population is assumed to be the English–French bilinguals, and the effect of a broad or narrow definition of the relevant population is not investigated in detail.

System testing was conducted following the same framework as Experiment 1, but the QS was drawn from one language while the KS was drawn from the other. Both language pairings of En–Fr systems, with an English QS and a French KS, and Fr–En systems, with a French QS and an English KS, were tested. To keep the amount of data tested the same as in Experiment 1, the English and French recordings for each speaker remained divided into two halves, so that the KS consisted of the first half of the speaker's recording in the KS language, rather than the full recording, and the QS consisted of the second half of the speaker's recording in the QS language. Performance on both global and individual levels was evaluated with 100 replications.

6.3.1 /s/ & LTFDs

Subject to differences in modelling approach (MVKD vs. GMM–UBM) and specific acoustic parameters (spectral moments vs. formants) described in Section 6.2, systems were set up in the same way for /s/ and LTFDs to assess the effects of language mismatch on their performance. These two variables are therefore discussed together here, while testing in Phonexia necessitated a slightly different approach that is described in Section 6.3.2.

For both /s/ and LTFDs, the aim of system testing in this experiment is not only to investigate the impact of language mismatch on global and individual performance for specific combinations of parameters, but also to examine any relationships between shifts in patterns of performance and underlying speech data of individual speakers. As such, system testing focused on five combinations of parameters for each of /s/ and LTFDs, including each individual spectral moment or LTF, as well as the combination of all four spectral moments or LTFs.

To address the main RQ2 and test the effect of language mismatch between samples, systems were set up in two primary conditions that differed in how the reference population was modelled. In Condition C1, the language of the background was matched with that of the KS, such that the mismatch between the background and the QS was the same as that between the KS and the QS. Condition C2, in which the language of the background data was matched instead with the QS language, was set up to assess the practical impact of such a decision by the analyst. Note that the definition of the relevant population remains unchanged in Conditions C1 and C2. In both conditions, the systems were trained and calibrated with training data that matched the case circumstances, meaning that the training data were cross-language comparisons with the same language pairing of QS and KS and same language of the background data. C_{llr} and EER obtained from these two conditions were compared with those obtained from same-language comparisons in Experiment 1.

Two secondary conditions (C1b and C2b) were set up to address RQ2.1, that is, the effects of mismatch between the conditions of test data and the conditions of the training data ("training mismatch") in relation to cross-language comparisons. These two conditions were identical to Conditions C1 and C2 respectively in the set-up of the languages, but the test data and the training data were mismatched here in the sense that calibration was performed using training scores from same-language comparisons, rather than from cross-language comparisons. In Condition C1b, where the language of the background data was matched with the KS language in the test data, calibration coefficients were generated from training scores that came from same-language comparisons in the KS language. Likewise, in Condition C2b, calibration was done using training scores taken from same-language comparisons in the QS language. In other words, the only difference in how LRs were derived between Condition C1b and same-language comparisons in the KS language is the language of the QS, and the only difference between Condition C2b and same-language comparisons in the QS language is the language of the KS.

Table 6.2 details each of the four conditions. Each condition corresponds with a different scenario of the type of reference data available to the analyst when faced with cross-language comparisons. Conditions C1 and C2 represent cases where the analyst does have access to a bilingual reference database, either off-the-shelf or specifically collected, such that cross-language comparisons can be conducted to carry out appropriate case-specific calibration, but different decisions are taken by the analyst as to the language of the background data. Conditions C1b and C2b, on the other hand, are akin to scenarios in which the analyst only has access to reference data in one of the languages from the samples and, as such, is restricted in terms of the choice of training data.

Table 6.2: Summary of languages used and conditions of training data in each condition.

Condition	QS	KS	Background	Training comparisons
C1	English	French	French	Cross-language
C1b			French	Same-language (French)
C2			English	Cross-language
C2b			English	Same-language (English)
C1	French	English	English	Cross-language
C1b			English	Same-language (English)
C2			French	Cross-language
C2b			French	Same-language (French)

To test the effects of language mismatch, LMEMs were fitted to C_{llr} and EER obtained from all 100 replications of Conditions C1 and C2, as well as those from same-language comparisons. For each combination of acoustic parameters, En–Fr systems and Fr–En systems were separately compared with corresponding systems of same-language comparisons. Condition, mode of data input (/s/: midpoint, quadratic and three-point; LTFDs: F-only, F+BW) and their interaction were included as fixed effects, and random intercepts were included for each replication. As per Section 6.2.1 above, significance of fixed factors was determined by LRTs in a step-down approach.

A separate set of LMEMs were fitted to C_{llr} from both primary and secondary conditions (C1, C2, C1b and C2b), so as to test the effects of training mismatch. EER was not analysed for this set of comparisons, as calibration based on different sets of training scores affects only C_{llr} but not EER. The reason EER is immune to calibration is inherent in how EER is determined (see Section 6.1.1 above) and how calibration is carried out. As calibration shifts and scales the entire set of scores in a uniform fashion, the EER score threshold shifts accordingly, such that the proportion of scores from SS and DS comparisons on each side of the threshold is fixed, regardless of shifting and scaling. The presence/absence of training mismatch, condition (C1/C1b vs. C2/C2b) and mode of data input were included as predictors of fixed effects, alongside all two-way and three-way interactions, but En–Fr and Fr–En comparisons were overall modelled separately.

6.3.2 ASR

The testing of cross-language comparisons in ASR systems was more limited, as the background model used in comparisons to derive output scores is preset and not tied to the analyst's decision. Whereas in the approaches above for /s/ and LTFDs, the background model is derived from data coming from a sample of the relevant population and as such different models for each language could be constructed from data in each language, the acoustic model is pre-built and fixed within the current ASR software, regardless of the language of the samples compared. The distinction between Conditions C1 and C2, as well as between C1b and C2b, is thus unavailable for testing. As such, only a single cross-language condition was set up for testing the effect of language mismatch for RQ2.

Following the same procedure as in Section 6.2.3, the first half of each speaker's French recording was submitted as the KS, and the second half of each speaker's English recording was submitted as the QS. A total of $60 \times 60 = 3{,}600$ scores from En–Fr comparisons were then generated. For each replication, scores from speakers in the test set were then calibrated to produce LRs using scores from speakers in the training set also derived from En–Fr comparisons. Fr–En scores were generated and calibrated in the same fashion by reversing the languages of the QS and the KS. Validity metrics were then compared with those from same-language comparisons using LMEMs to assess the effect of language mismatch, with condition included as the fixed factor and random by-replication intercepts included.

To test the effect of training mismatch (RQ2.1) on ASR systems in cross-language comparisons, test scores from cross-language comparisons were calibrated in a separate condition using scores from same-language comparisons, obtained from speakers in the training set. For /s/ and LTFDs, the language of the same-language comparisons used to calibrate these scores were matched with the language of the background data (the KS language in Condition 1b and QS language in Condition 2b). Here, En–Fr and Fr–En scores were each separately calibrated using scores from En–En and Fr–Fr comparisons to explore the potential effect of language. Even though the choice of language could not be specifically linked to the reference model used in the feature-to-score stage, testing the effect of language at this stage nonetheless remains important, as the availability of reference materials in one or both languages may well influence the analyst's decision of how LRs in such cases are derived and in turn impact the validity of the comparison outcomes. For each language pairing of comparisons (En–Fr and Fr–En), the effect of training mismatch was then tested by fitting LMEMs to the three sets of scores, with "training condition" (calibrated using cross-language, En–En or Fr–Fr scores) included as the fixed factor and random intercepts included for each replication.

Chapter 7

Results: /s/

This chapter presents results from Experiment 1 (Section 7.1) and Experiment 2 (Section 7.2) for the segmental variable /s/. Within each experiment, results from the global-level analysis are first presented and discussed. Zooplot analysis on the individual level then follows, accompanied by an acoustic analysis of speakers classified as outlying in each system. Each section concludes with a discussion of findings from each experiment with regard to /s/.

7.1 Experiment 1: Same-language comparisons

7.1.1 Global metrics

Figures 7.1 and 7.2 presents the distributions of C_{llr} and EER for all tested systems, grouped by the number of spectral moments included as input in the system, in English and French respectively. Figures 7.3 and 7.4 further highlight validity metrics obtained from systems of individual moments, illustrating the distributions for each spectral moment in each language.

Starting with individual moments, systems produced mean C_{llr} between 0.41 and 0.84 and EER between 11.2% and 32.9%. The lowest mean C_{llr} and EER were achieved using quadratic coefficients of CoG in French. C_{llr} was below 1 in almost all replications, with only a small number of exceptions for skewness and kurtosis. Thus, all spectral moments, regardless of the mode of input, resulted in systems that were not poorly calibrated and provided useful speaker-specific information in voice comparison, although validity measures suggest that, in isolation, their performance was only weak to moderate.

Model comparisons showed a significant three-way interaction of language, param-

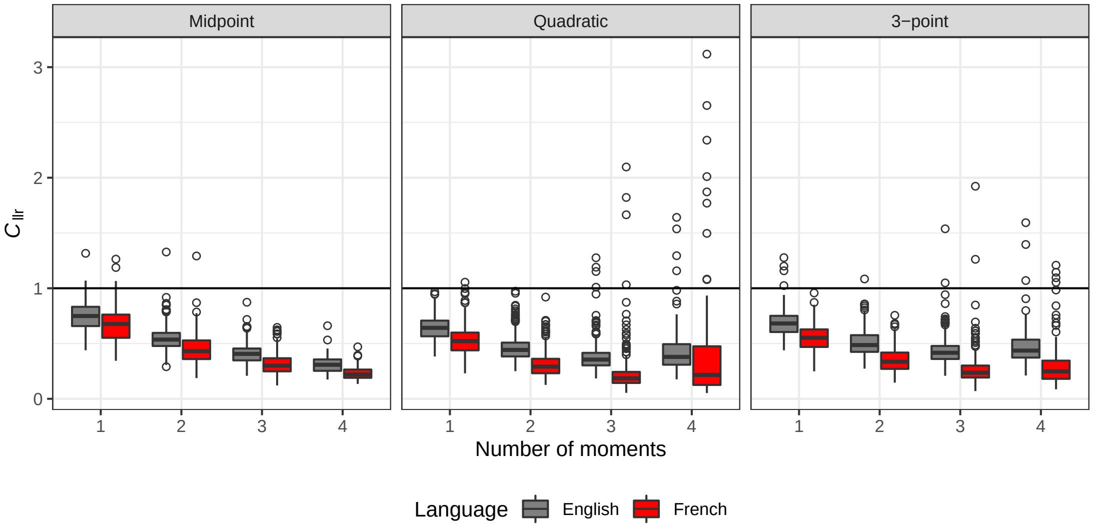

Figure 7.1: C_{llr} from system testing of /s/ in English (grey) and French (red), grouped by mode of input and number of spectral moments included. Horizontal rule reference at $C_{\text{llr}} = 1$.

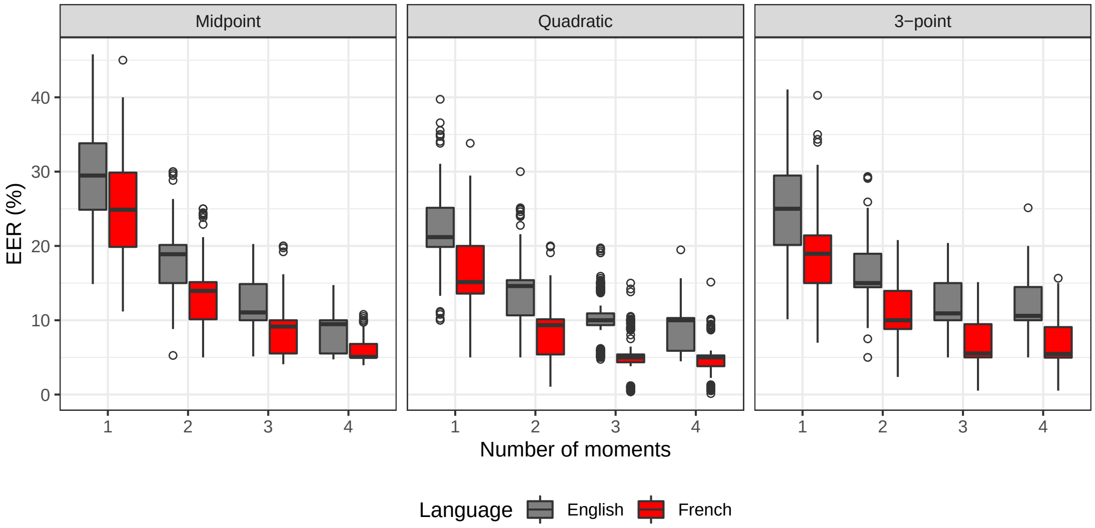

Figure 7.2: EER from system testing of /s/ in English (grey) and French (red), grouped by mode of input and number of spectral moments included.

eter and mode of input for both C_{llr} ($\chi^2(6) = 17.15,\ p = .0087$) and EER ($\chi^2(6) = 24.98,\ p = .0003$). While both C_{llr} and EER were consistently lower in French than in English, their patterning exhibited subtle differences across parameters and types of input. In the midpoint-only systems, C_{llr} was lower in French by 0.08 on average and EER decreased by 4.1%. The degree of downward shift between languages was marginally greater for the dynamic systems: The quadratic systems on average reported C_{llr} and EER that was lower by 0.12 and 6.3% in French, while the difference between English and

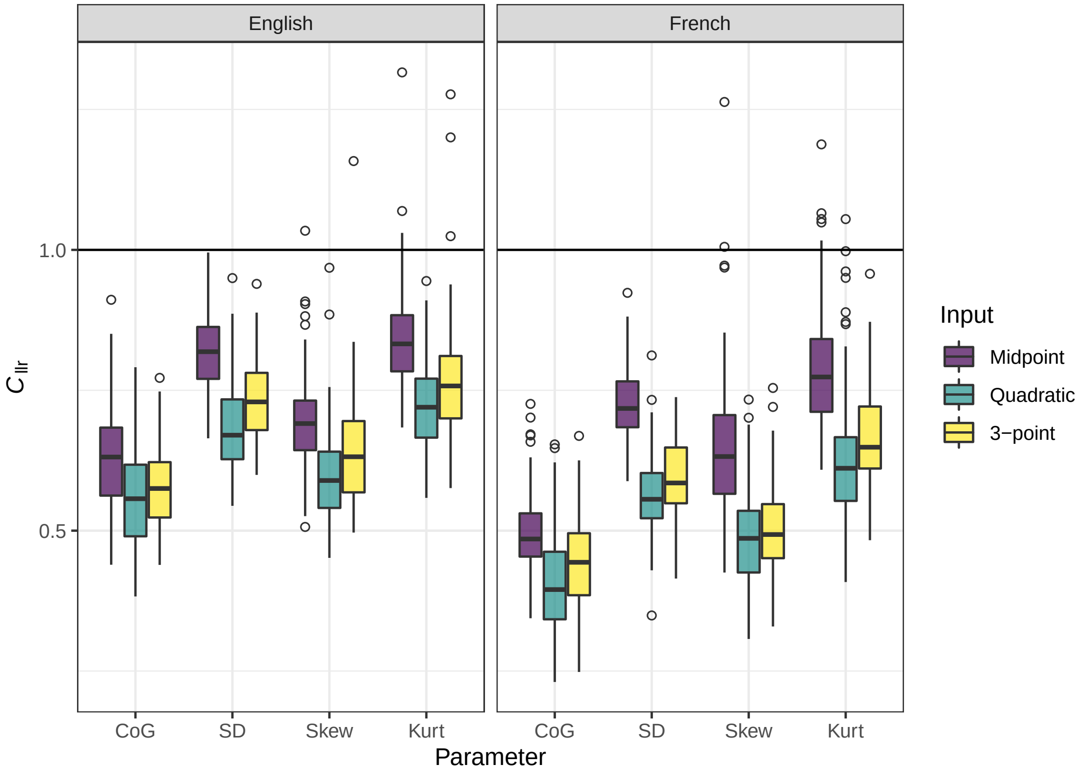

Figure 7.3: C_{llr} from system testing of all midpoint-only (purple), quadratic (green) and three-point (yellow) systems using individual spectral moments in English and French. Horizontal rule reference at $C_{\text{llr}} = 1$.

French in the three-point systems was 0.13 for C_{llr} and 6.1% for EER.

Out of the four spectral moments, CoG produced the lowest mean C_{llr} and EER across all types of input in both languages. Mean C_{llr} for CoG ranged from 0.41 to 0.63, and mean EER ranged from 11.2% to 22.6%. Skewness was the next best performing spectral moment, resulting in mean C_{llr} and EER of 0.49–0.69 and 14.6%–25.7%. SD and kurtosis reported the highest metric scores: Mean C_{llr} was 0.56–0.82 for SD and 0.63–0.84 for kurtosis, while mean EER was 19.4%–32.6% for SD and 19.7%–32.9% for kurtosis. Post-hoc Tukey-adjusted pairwise comparisons for C_{llr} within each language and input, carried out with the *emmeans* package in R (Lenth, 2020), showed a consistent, stepwise significant ranking of CoG, skewness, SD and kurtosis (all $p < .0050$), with the exception of midpoint-only systems in English, where kurtosis did not produce significantly higher C_{llr} than SD ($p = .1735$). An equivalent set of comparisons for EER found a similar hierarchy for CoG, skewness and SD (all $p < .0001$), but kurtosis was not significantly different from SD ($p > .08$) except for the three-point systems in French ($p < .0001$).

In terms of modes of input, it is clear from Figures 7.3 and 7.4 that dynamic input consistently outperformed static input, as evidenced by the overall lower C_{llr} and EER

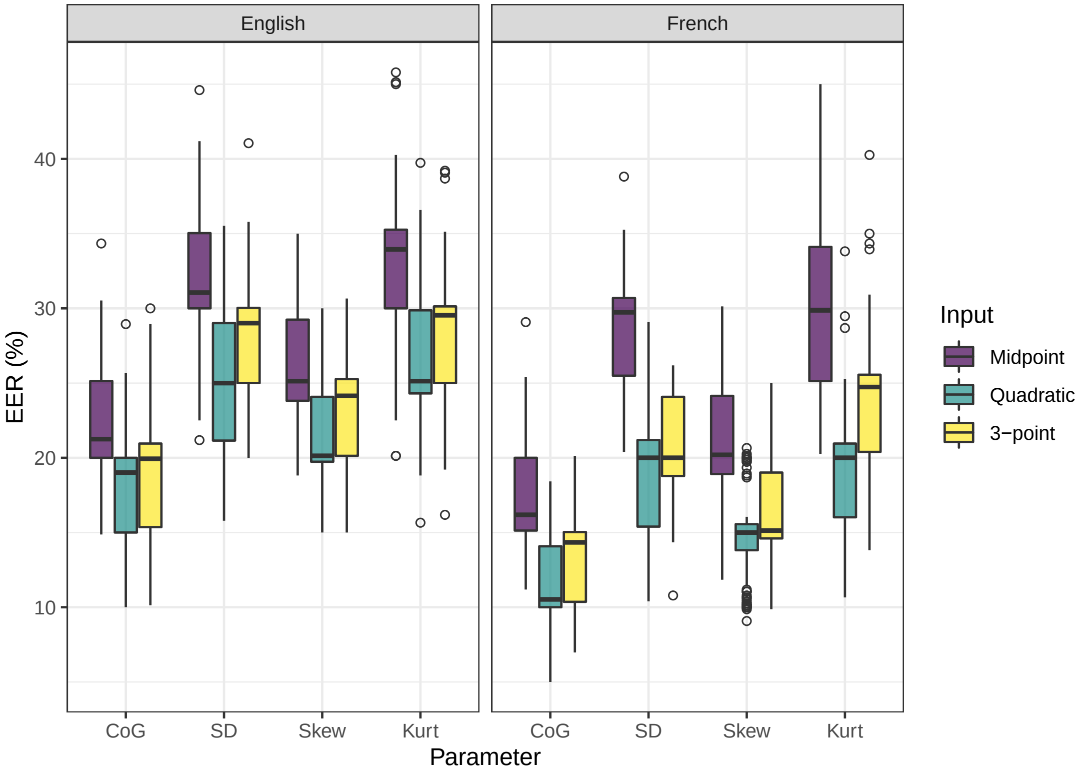

Figure 7.4: EER from system testing of all midpoint-only (purple), quadratic (green) and three-point (yellow) systems using individual spectral moments in English and French.

from the quadratic and three-point systems when compared to the midpoint-only systems. Between the two sets of dynamic input, the differences in C_{llr} and EER were relatively small, but quadratic systems consistently tended to produce lower metric scores than three-point systems. Post-hoc Tukey-adjusted pairwise comparisons by language and parameter confirmed that C_{llr} and EER from quadratic systems were significantly lower than those from three-point systems (all $p < .05$), with some exceptions: C_{llr} between the two was not significantly different for CoG in English ($p = .3771$) and skewness in French ($p = .6285$), while EER was not significantly for SD in French ($p = .3244$). C_{llr} and EER from both dynamic systems were consistently significantly lower than those from corresponding static systems (all $p < .0001$).

Figures 7.1 and 7.2 show that, as the number of spectral moments included in the systems increased, both C_{llr} and EER generally decreased, indicating stronger overall validity as more parameters were combined. Mean C_{llr} and EER for each number of moments are summarised in Table 7.1. In midpoint-only systems, mean C_{llr} decreased from 0.75 to 0.31 in English and from 0.67 to 0.23 in French. Significant main effects of number of moments included ($\chi^2(3) = 3186.8,\ p < .0001$) and language ($\chi^2(1) = 534.51,\ p < .0001$)

were found for C_{llr} in midpoint-only systems, confirming the trends observed, but there was no significant interaction between the two factors ($\chi^2(3) = 3.65$, $p = .3024$), suggesting that C_{llr} decreased at the same rate in both English and French as the number of moments increased. EER similarly decreased from 28.5% to 8.5% in English and from 24.4% to 6.1% in French, but stepwise improvements of EER in French became progressively smaller than those in English, as shown by a significant interaction between number of variants and language ($\chi^2(1) = 14.21,\ p = .0026$).

In both sets of dynamic systems, EER also decreased as the number of moments included increased, from 22.5–24.6% to 9.3–11.9% in English and from 16.2–18.5% to 4.8–6.5% in French, though the decrease resulting from the inclusion of a fourth moment was only marginal (0.2–1.1%). A significant interaction between the fixed factors in the models fitted to the quadratic systems ($\chi^2(3) = 16.12,\ p = .0011$) points to a narrowing gap between English and French as more spectral moments were included, similar to that in midpoint-only systems, though the interaction was insignificant in the models fitted to the three-point systems ($\chi^2(3) = 5.35,\ p = .1478$). Post-hoc Tukey-adjusted comparisons indicated that the differences in EER between systems using three and four moments were largely insignificant (French quadratic: $p = .7300$; three-point: $p = .7448$), with the exception of a slight but significant decrease in English quadratic systems ($p = .0264$).

Mean C_{llr} in dynamic systems, on the other hand, did not decrease monotonically. A significant interaction of number of moments and language in models fitted to both quadratic ($\chi^2(3) = 35.10,\ p < .0001$) and three-point ($\chi^2(3) = 12.57,\ p = .0057$) systems suggests that, even though systems in French generally yielded lower C_{llr} than systems in English, the effect of including more spectral moments in the system did not follow parallel trajectories in the two languages. Nevertheless, it can be seen that, in both languages and for both modes of input, inclusion of a second and third spectral moment led to lower C_{llr}, which rose at the inclusion of a fourth moment. Post-hoc Tukey-adjusted comparisons showed this uptick in C_{llr} to be significant (all $p < .0025$), rising to a level comparable to (or significantly higher than, in the case of the French quadratic systems: $p < .0001$) the corresponding systems with only two moments (all $p > .10$). In fact, when all four moments were included in the quadratic systems, C_{llr} in French was not significantly different from that in English ($p = .3877$).

The trends described above for systems of individual moments in relation to mode of input extend to systems combining multiple moments, but only to a limited degree. While dynamics systems including two spectral moments outperformed the corresponding midpoint-only systems in terms of both C_{llr} and EER, Table 7.1 shows that the difference across modes of input was largely levelled out when a third moment had been included. This was especially evident in English, where mean C_{llr} and EER for the midpoint-only systems (0.41, 12.1%) were virtually identical to those obtained for the 3-point systems (0.43, 12.3%). When all moments were included, the trend between midpoint-only and dynamic systems was mostly reversed: Midpoint-only systems yielded lower C_{llr} and similar to lower EER than dynamic systems.

Table 7.1: Mean C_{llr} and EER from midpoint-only, quadratic and three-point systems in English and French by number of spectral moments included.

		Midpoint		Quadratic		3-point	
	No. of moments	C_{llr}	EER	C_{llr}	EER	C_{llr}	EER
English	1	0.75	28.5%	0.64	22.5%	0.68	24.6%
	2	0.54	17.9%	0.45	14.1%	0.51	15.8%
	3	0.41	12.1%	0.38	10.4%	0.43	12.3%
	4	0.31	8.5%	0.44	9.3%	0.48	11.9%
French	1	0.67	24.4%	0.52	16.2%	0.55	18.5%
	2	0.47	13.3%	0.30	8.6%	0.35	10.4%
	3	0.32	8.5%	0.22	5.2%	0.26	6.7%
	4	0.23	6.1%	0.42	4.8%	0.32	6.5%

7.1.2 Interim discussion

In this experiment, the primary question sought to be addressed is whether the discriminatory potential of the same acoustic parameters of /s/ remains stable across languages. The findings above clearly indicate that this is not the case for /s/ in Canadian English and French. Regardless of the combination of parameters or mode of data input, stronger performance was found for /s/ in French than in English. As the experiment was conducted in both languages with the same bilingual population, and the composition of the test, training and background sets remained the same across all tested systems for each replication, the quantitative differences found here cannot be simply attributed to idiosyncratic physiological variation or sampling from different populations. Such cross-linguistic discrepancy in discriminatory power is in line with the predictions made earlier in Section 5.3 and is here proposed to reflect the greater within-speaker variability demonstrated in the English data.

What does remain stable across languages is the relative discriminatory power between spectral moments. In both English and French, CoG turned out to be the best-performing parameter, closely followed by skewness. By contrast, SD and kurtosis were both relatively poor speaker discriminants, resulting in high EER and C_{llr} that were close to 1. These results arguably reflect the acoustic relationships between spectral moments, as CoG and skewness often are correlated, as are SD and kurtosis (Harrington, 2010). As both CoG and skewness are related to the place of articulation of /s/, while both SD and kurtosis have been argued to correlate with the tongue gesture (whether /s/ is formed with an apical or laminal constriction; F. Li et al., 2009), the current findings provide support for the conclusion that, at least for /s/, parameters characterising place of articulation offer stronger speaker-specificity than parameters characterising tongue gesture.

Performance between different modes of input was also compared. Previous research by Kavanagh (2012) and Smorenburg and Heeren (2020) found that spectral dynamics provided little to no improvement in discriminatory power over static measurements. Current findings diverge from earlier results, showing instead that systems using dynamic input did demonstrate stronger measures of validity than those using only static input. In line with findings from vowel formant dynamics (Hughes et al., 2016; McDougall, 2004), A dynamic characterisation of spectral moments can therefore be said to capture a higher degree of speaker-specificity. Nevertheless, the level of additional idiosyncratic information available remains limited, as improvements in C_{llr} and EER were not high. Based on the acoustic findings in Chapter 5, this conclusion is in fact not unexpected, as much of the between-speaker variation in spectral dynamics lies not so much in the shapes of the trajectories themselves, but in the relative height of the trajectories. The prediction that quadratic coefficients are better equipped than the three-point approach to encode speaker-specific information in spectral dynamics has also been borne out in the current data. Even though there were only small differences in C_{llr} and EER between the two approaches, quadratic systems always outperformed three-point systems, indicating the importance of accounting for speaker-specific information outside onset, midpoint and offset as the sole predefined points of interest.

Findings from C_{llr} nevertheless point to the need for caution when using dynamic systems that combine input from multiple spectral moments. In spite of their relatively strong overall performance, when three or four moments were included, dynamic systems produced a much wider range of C_{llr} over the 100 replications than other tested systems, with a number of high outliers in excess of 1. Quadratic systems and systems in French suffered more strongly, but the same phenomenon can be found in three-point systems and systems in English to a lesser extent. In these circumstances, the level of performance of dynamic systems appeared to depend much on the specific combination of speakers in the test, training and background sets in each replication (Wang et al., 2019a). Improved performance of dynamic systems, especially in French, then, is countered by lower reliability than the corresponding static systems, whose C_{llr} occupied a much lower range and did not exceed 1 in any replication when three or four moments were combined together. Such deterioration in reliability is very likely due to limitations in the MVKD formula itself in modelling acoustic parameters. As Nair et al. (2014) point out, MVKD was originally designed to accommodate a small number of parameters, and its heavy reliance on kernel density estimation and matrix inversion could lead to computational issues when a large number of parameters are used, resulting in drastic misestimation of LRs. At 12 parameters (4 spectral moments $\times$ 3 coefficients), the application of MVKD likely suffers from the adverse effects of overly high dimensionality, when there were insufficient data to compensate for the issue. That EERs remained low for these combinations of parameters suggests that there were very few contrary-to-fact comparisons where LLR < 0. LRs that were contrary to fact, however, were of very high magnitude, confirmed by a close examination of individual LRs, which would give an exaggerated estimate of the strength of evidence. Further means of dimension reduction, such as Principal Component Analysis (Nair et al., 2014), or a different modelling technique capable of handling a higher number of dimensions, such as the GMM–UBM approach (Reynolds et al., 2000), may be necessary for the discriminatory potential of dynamic measurements to be more effectively harnessed.

7.1.3 Individual-level analysis

For systems using each individual spectral moment and the combination of all moments, zooplot analysis was carried out for each method of data input. An initial analysis suggests that the distribution and classification of speakers show good correspondence across methods of input. The current section therefore focuses only on midpoint-only systems, due to the aforementioned greater stability of individual LRs when all spectral moments are combined, as well as to facilitate the visualisation of acoustic comparisons in the next section. Zooplots and overall speaker classifications derived from the dynamic systems are included in Appendix B.

7.1.3.1 CoG

The zooplots for midpoint CoG systems in English and French are shown in Figure 7.5. In English, all speakers produced negative mean DS-LLR and most speakers produced positive mean SS-LLR, meaning that, on average, the vast majority of speakers can be identified with themselves and distinguished from other speakers. Seven speakers (12%), however, produced a negative mean SS-LLR of up to -0.87. The overall range of mean SS-LLR was very narrow, with a maximum of 1.88. The range of mean DS-LLR was much wider, with a number of speakers producing mean DS-LLR beyond -10 and the most extreme value being −17.3. Generally, speakers with higher SS-LLR also correspondingly produced higher DS-LLR, and a strong correlation can be found between performance in SS comparisons and that in DS comparisons ($r = -.73$, $p < .0001$). Indeed, very few speakers can be found in the upper left or the lower right region of the zooplot: Only one phantom speaker was present and no chameleon speakers could be identified in this system.

In general, the distribution of speakers for CoG in French resembles that described for English, and each relational group contained approximately the same number of speakers as it did in English (11 doves, four worms, three phantoms and no chameleons). SS and DS performance is also strongly correlated with each other ($r = -.77$, $p < .0001$). Nevertheless, between the two languages there is a notable distinction in the range of values obtained. As shown in the zooplot on the right in Figure 7.5, most speakers produced a small positive mean SS-LLR between 0 and 1, and negative mean DS-LLR between 0 and −10. The overall range of mean SS-LLR remains relatively limited, from −0.37 to 3.40, with only four speakers producing a small negative mean SS-LLR. In terms of DS-LLR, however, speakers with exceptionally good performance produced extreme values of up to −50.7.

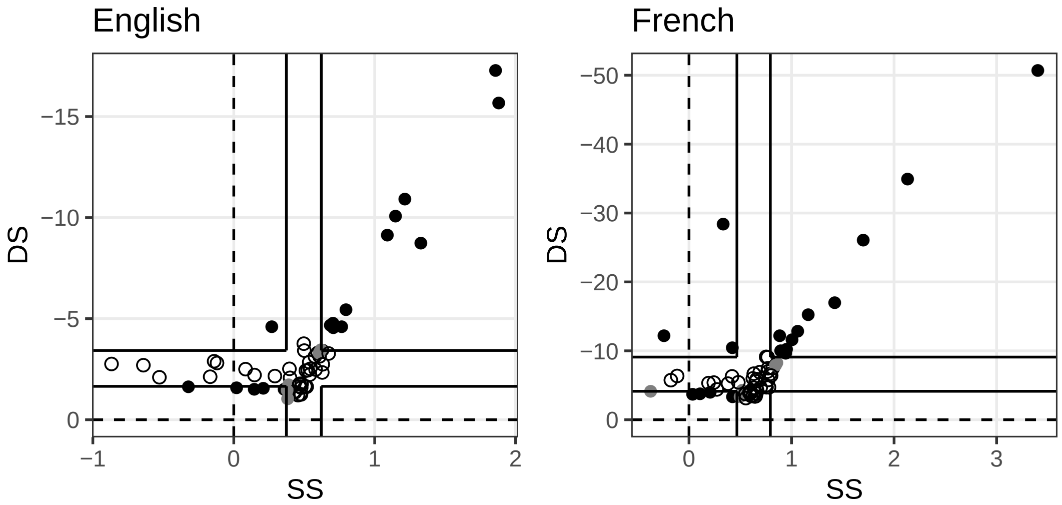

Figure 7.5: Zooplots for systems with midpoint CoG as input in English (left) and French (right). Abscissa and ordinate respectively show mean SS-LLR and DS-LLR. Solid line segments represent 25th and 75th percentiles; dotted lines indicate mean LLR = 0. Members and near-members of relational groups respectively in black and grey.

7.1.3.2 SD

The zooplots for systems using SD, shown in Figure 7.6, show broad similarities with those for CoG at first glance. Indeed, the number of speakers classified as members of each relational group are virtually indistinguishable from the zooplots in Figure 7.5. In both English and French, a significant dove population of 11 speakers could be found, alongside two worm speakers each and no chameleons. Both languages also had a small phantom group, consisting of a sole speaker in English and three in French.

A number of differences between the zooplots here and those for CoG above cannot be ignored. The ranges of SS-LLR and DS-LLR were both considerably smaller: In English mean SS-LLR ranged from −0.68 to 1.99 and in French from −0.81 to 1.33, while mean DS-LLR ranged from −0.13 to −4.09 in English and from −0.40 to −8.66 in French. At the same time, a greater proportion of speakers—12 in English and eight in French—were on average not well-matched with themselves, producing negative mean SS-LLR. While no speakers were identified as chameleons in either language, speakers formed a dense cluster close to the lower right corner of the zooplot. Accordingly, performance in SS and DS comparisons is only moderately correlated in both English ($r = -.54$, $p < .0001$) and French ($r = -.53$, $p < .0001$).

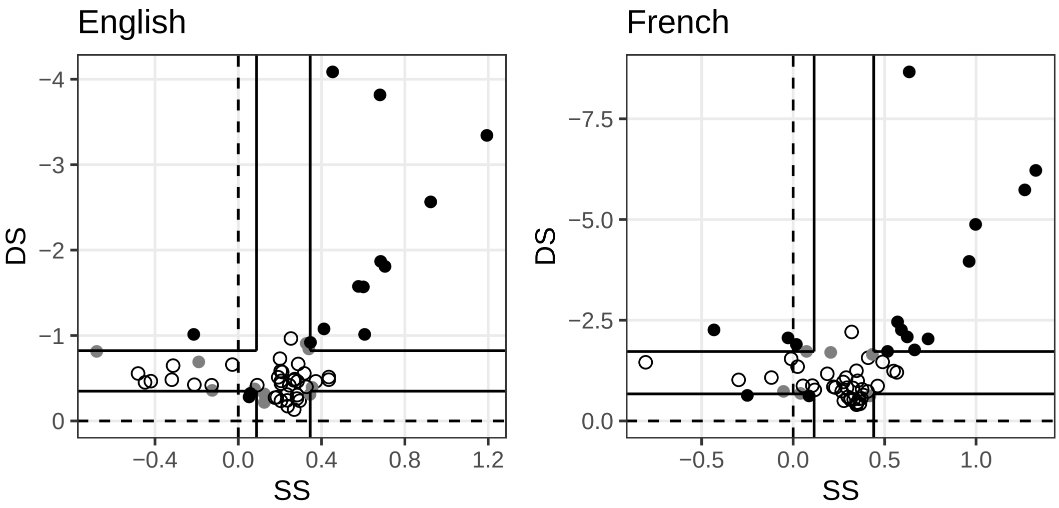

Figure 7.6: Zooplots for systems with midpoint SD as input in English (left) and French (right).

7.1.3.3 Skewness

Turning to the third spectral moment, skewness, the zooplots in Figure 7.7 show almost all speakers to be clustered within a narrow range of SS-LLR and DS-LLR values. In both

English and French, only a very small number of exceptional speakers produced a mean SS-LLR greater than 1 or a mean DS-LLR beyond -5. The isolated speaker with the most extreme performance produced a mean DS-LLR beyond -20 in both languages, as well as the highest mean SS-LLR (5.36 in English, 4.48 in French) across all individual spectral moments. As in the case of CoG, there were only a small proportion of speakers producing a negative mean SS-LLR (English: 13%; French: 8%), and performance in different types of comparison remains strongly correlated (English: $r = -.88$; French: $r = -.81$; $p < .0001$). The findings above of a significant presence of doves (11 in English, 12 in French) and a complete absence of chameleons are also maintained here. Additionally, six worms and one phantom were identified in English, whereas only four worms and no phantoms were identified in French.

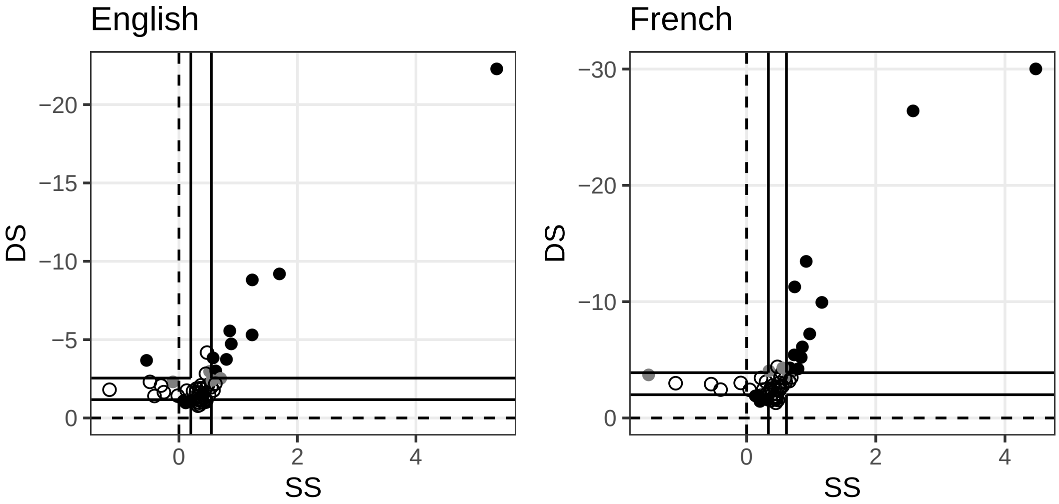

Figure 7.7: Zooplots for systems with midpoint skewness as input in English (left) and French (right).

7.1.3.4 Kurtosis

The zooplots for the kurtosis-based systems, displayed in Figure 7.8, show a striking degree of clustering, especially along the SS-LLR dimension. The distance between the 25th and 75th percentiles, which mark the boundaries of the relational groups, is the smallest across all spectral moments. In French, mean SS-LLR for half of the speakers lay between 0.15 and 0.41, while in English, the range further narrowed to 0.11 to 0.30, indicating that the strength of evidence offered by most speakers is very limited. While SS and DS performance remains strongly correlated (English: $r = -.75$; French: $r = -.80$; $p < .0001$), the tight clustering in the distribution meant that the population of each relational group was relatively small. In each language, 10 doves, who were clearly set apart from the rest of the speakers, were identified. Of the other groups, only three speakers were classified as phantoms and one as chameleon in English, where the worm group was wholly absent. In a similar vein, only two worms and one phantom, but no chameleons, were found in French.

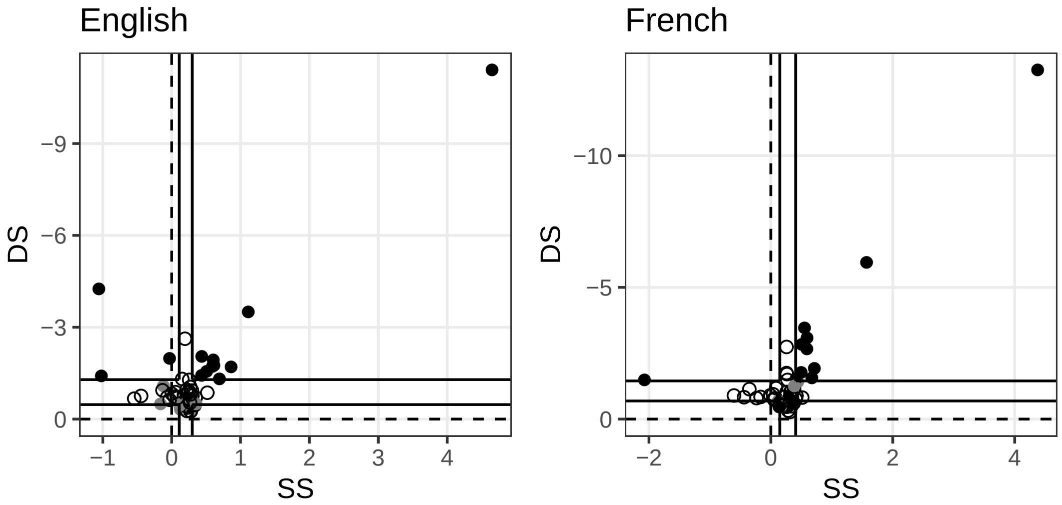

Figure 7.8: Zooplots for systems with midpoint log-kurtosis as input in English (left) and French (right).

7.1.3.5 All moments combined

When compared with the zooplots for individual spectral moments, the zooplots for the combined systems in Figure 7.9 show a noticeably less clustered distribution, with greater separation between individual speakers. In both languages, speakers generally showed stronger performance in SS comparisons when compared to systems of individual moments, with a number of speakers producing mean SS-LLR $>$ 2, which was a rare occurrence in the systems using individual moments. Only three speakers produced small negative mean SS-LLR in either language. There is also a shift towards more highly negative DS-LLR, not only for the particularly well-performing speakers, but also for the rest of the population. Doves remained the largest animal group in both English (eight speakers) and French (11 speakers), followed by the worms (three and five speakers in English and French respectively) and the phantoms (two each), whereas chameleons remained virtually absent (only one speaker in French).

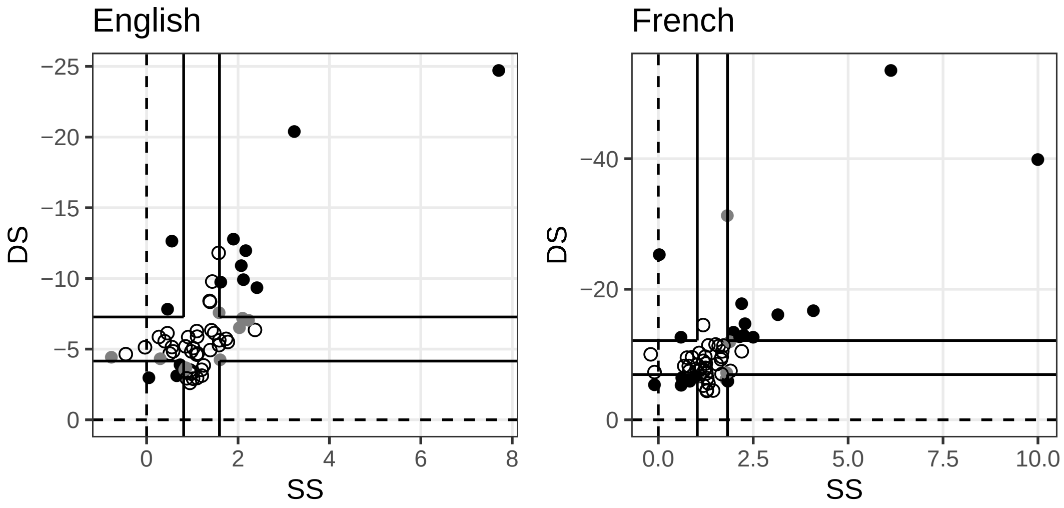

Figure 7.9: Zooplots for systems with midpoint of all four spectral moments as input in English (left) and French (right).

7.1.3.6 Speaker classification

In this section, the performance of individuals across different systems is considered, along with the stability of their performance across languages. Figure 7.10 summarises the animal group classification of all 60 speakers for each system tested in both English and French.

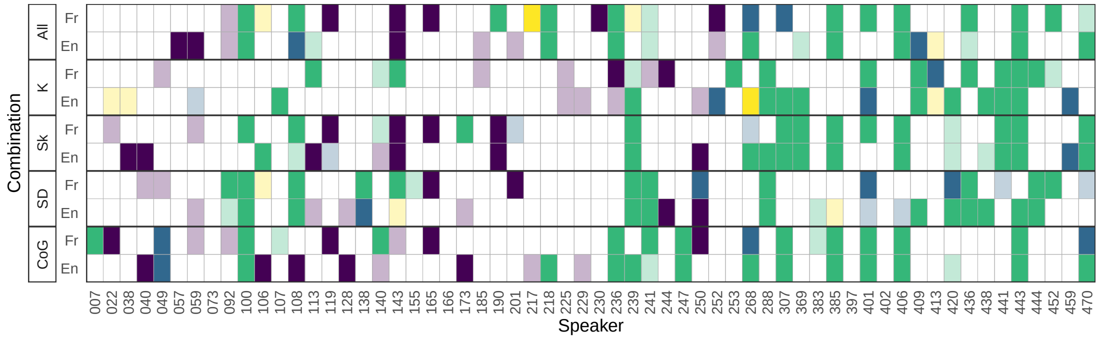

Figure 7.10: Summary of speakers classified as doves (green), worms (purple), phantoms (blue) or chameleons (yellow) in midpoint-only systems tested (near-members of each group in lighter shade).

In English, while 13 speakers (22%) were not classified as a member of any relational group across all four spectral moments, of the 47 who could be considered as having exceptional performance in one way or another, 25 were in or near the same group for multiple moments, and three (239, 420 and 443) were in or near the doves for all individual moments. On the other hand, it was not uncommon for speakers to fall into different animal groups for systems of different moments. This is applicable to 10 speakers, two of which were classified as best-performing doves for one moment and worst-performing worms for another. Overall, although each spectral moment largely captured a different group of outlying speakers, there is some evidence that the speaker-specific information contained in the spectral moments is not independent of one another. In particular, there is much overlap of dove membership between CoG and skewness, as well as between SD and kurtosis, whereas Figure 7.10 shows overlap of other relational groups or between other pairs of moments to be much rarer.

The performance of individuals in different spectral moments generally exhibits the same patterns in French. 14 out of 60 speakers (23%) did not produce outlying performance for any spectral moment, while 23 out of the remaining 46 were in or near the same group more than once. Two speakers (143 and 401) were members or near-members of some relational group for all four spectral moments, but in neither case did the speaker receive the same classification across the board. In fact, 143 was one of three speakers in French to be both doves and worms for different moments. The overlap of zoo membership between different moments is even stronger here than in English, with the systems for CoG and skewness sharing seven doves/near-doves and five worms/near-worms.

When the systems combining all spectral moments were compared with those of individual moments, only a moderate level of correspondence of zoo membership could be found. In both English and French, approximately one third of all speakers who were in or near an animal group (English: 8/22; French: 7/24) were not similarly classified for any individual moment. These speakers were either not placed in any group or placed in groups that were distinct from the one in the combined system. The remaining relational group members in the combined systems had the most common classification with the CoG systems, closely followed by the systems using midpoint skewness, while kurtosis accounted for the fewest members. Notably, there are also speakers (e.g., 239) who were consistently classified as a particular animal (mostly dove) for three or even four individual moments, but did not produce outlying performance in the combined system. These findings suggest a complex interplay of speaker-specificity when the contribution from multiple spectral moments is considered in tandem. That speakers with outlying performance in individual moments cannot well account for those in the combined systems may be due to the dominating influence of isolated speakers (doves) with extreme performance in one or more individual moments, which means that other speakers in the group, despite being identified as doves, performed at a relatively similar level to speakers outside the group.

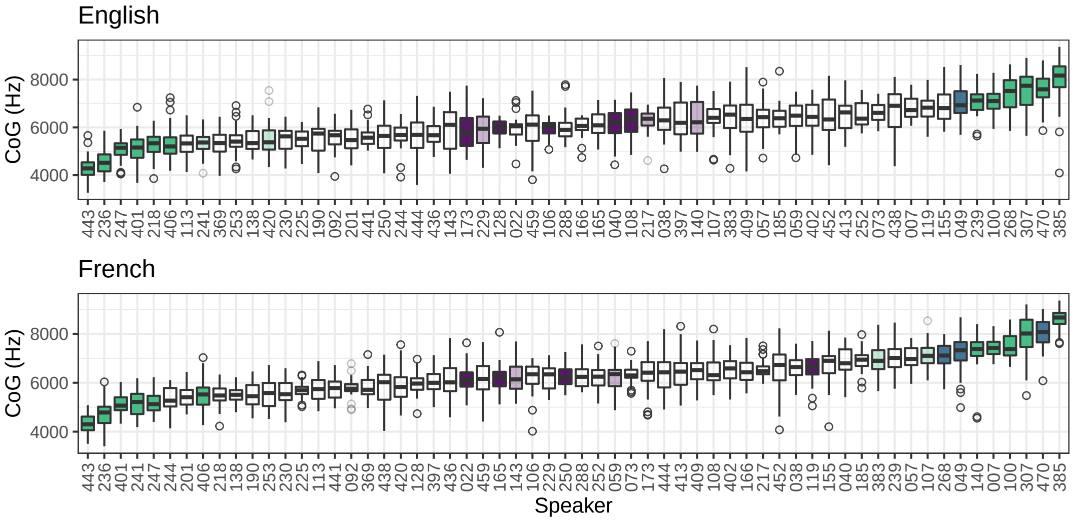

Figure 7.11: Boxplot of CoG (Hz) for all speakers in English and French, arranged in ascending order of speaker mean by language, with doves in green, worms in purple and phantoms in blue. Light colours denote near-members.

Although speakers produced stronger performance in French than in English, as evident from the results presented thus far, Figure 7.10 shows that the relative performance of individuals remained generally stable across languages. This is especially the case for speakers identified as performing exceptionally well, as dove membership in each set of systems matched up for a vast majority of speakers in the group. While speakers were only sporadically classified as members of other animal groups in both languages, actual mismatch in classification within systems using the same set of parameters was rare, amounting to only two to three speakers per system. All such cases of mismatch involved speakers being classified as doves or worms in one language but phantoms in the other (with the exception of 140 for CoG, who was a dove in French but a near-worm in English), suggesting that performance remained stable in at least one type of comparison for these speakers.

7.1.4 Acoustic analysis

Figures 7.11 to 7.14 present a breakdown of each speaker's individual distribution for each spectral moment, arranged in ascending order of speaker mean, where speakers with outlying LR performance as identified above are indicated accordingly.

In the CoG distributions in Figure 7.11, a clear pattern that emerges in both English and French is the collection of doves and phantoms towards the margins of the group. As both doves and phantoms are speakers with the strongest performance in DS compar-

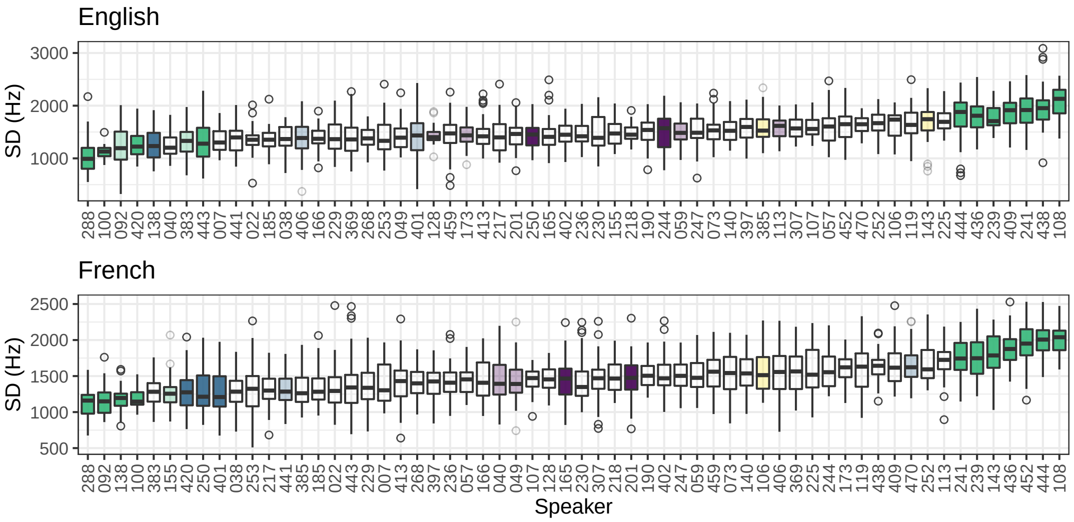

Figure 7.12: Boxplot of SD (Hz) for all speakers in English and French, arranged in ascending order of speaker mean by language, with doves in green, worms in purple, phantoms in blue and chameleons in yellow. Light colours denote near-members.

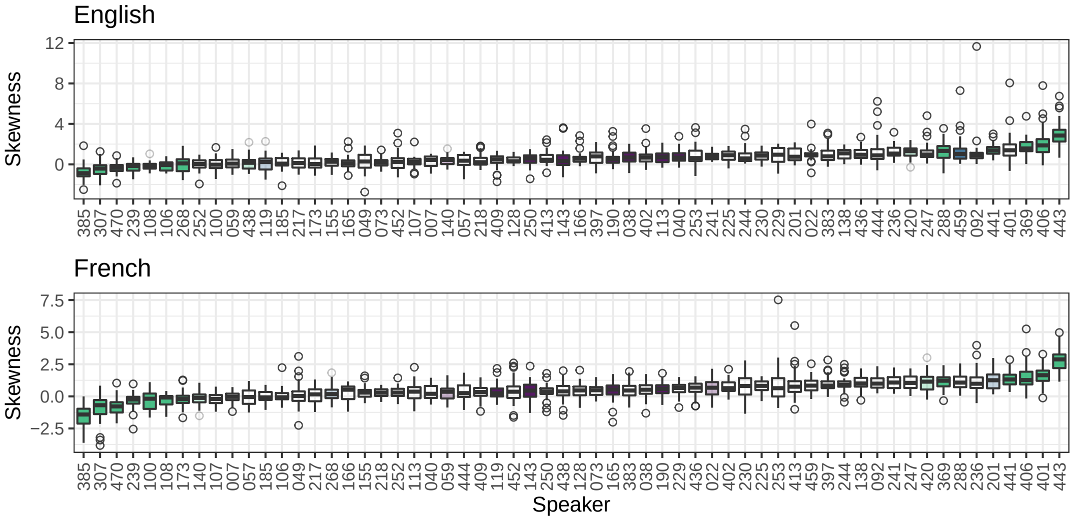

Figure 7.13: Boxplot of skewness for all speakers in English and French, arranged in ascending order of speaker mean by language, with doves in green, worms in purple and phantoms in blue. Light colours denote near-members.

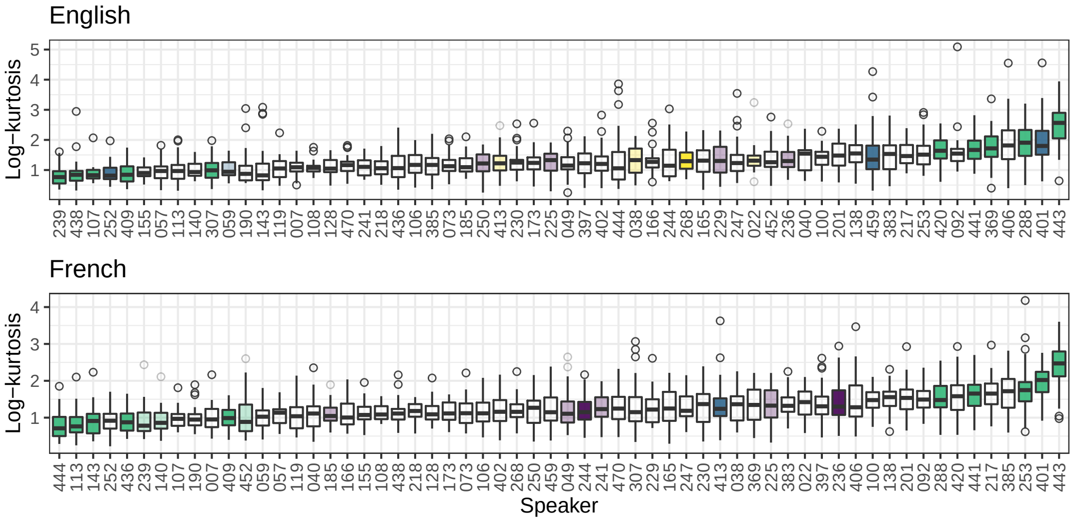

Figure 7.14: Boxplot of log-kurtosis for all speakers in English and French, arranged in ascending order of speaker mean by language, with doves in green, worms in purple, phantoms in blue and chameleons in yellow. Light colours denote near-members.

isons, these speakers are set apart from the others by having extremely high or low CoG. The few phantoms do not form a separate cluster by themselves, but are mixed in with doves at the high end of the distribution. As doves and phantoms are distinguished by their performance in SS comparisons, it might be expected that phantoms would demonstrate greater within-speaker variation while doves would display more lower within-speaker variation. A close inspection of Figure 7.11, however, find no evidence for this. In fact, the group of doves encompass speakers with little intra-speaker variation (e.g., 007 and 140 in French) and those who demonstrate much greater intra-speaker variation (e.g., 307 in French) exceeding that of phantom speakers. A possible explanation for the divergence between doves and phantoms may thus lie in how variation within individuals is manifested in the samples tested. As the current study makes use of a single recording divided into two halves for each speaker, speakers whose variability in CoG remains relatively consistent throughout the whole recording would encounter little issue in SS comparisons, whereas speakers whose /s/ CoG differs considerably between the two parts of the recording would perform poorly in SS comparisons and thus more likely be classified as phantoms. Worms, on the other hand, demonstrate a tendency to have mean CoG close to the centre of the group. Although the worms are not clustered together, they all have similar distributions of CoG to other speakers in the central part of the group, such that given a particular sample there could be little evidence of it originating from one speaker over another similar speaker.

The same patterns extend to the distributions of other spectral moments as well. Figures 7.12 to 7.14 all show speakers classified as doves for SD, skewness and kurtosis to be among those with the highest or lowest means. These findings may further account for the high level of correspondence in dove membership between English and French. As the spectral moments produced by individual speakers were found to be highly consistent across languages, speakers ranked close to the top or bottom of the group's distributions in one language were also likely relatively extreme in the other, as evident in each Figure, resulting in exceptionally strong discriminatory performance overall. Similarly, phantoms are typically found near or among doves, with the exception of kurtosis in French for one speaker. Worms (and near-worms) include speakers whose mean spectral moments are far from the extremes but are instead spread over the central part of the group and never overlap with doves (or near-doves). No clear pattern can be found for chameleons, which only rarely occur within this set of systems. For kurtosis, members of this group show very similar distributions to near-worms (no worms were identified), but for SD the near-chameleons are located somewhere between the worms in the middle and the doves with the highest individual SD. A different set of variables that results in a larger group of chameleons would be necessary in order to analyse the relationship between the underlying data and its membership.

7.1.5 Discussion: Same-language comparisons

Returning to the question of whether /s/ offers similar discriminatory potential in English and French, the results above from global- and individual-level analysis have shown that there is a consistent gap between how each parameter performs in each language. Despite their similar acoustic targets (and minor differences in dynamic trajectories), /s/ in French outperformed /s/ in English in terms of both C_{llr} and EER. It is thus suggested that the lower within-speaker variability of /s/ in French, found in Section 5.1, allows for more effective speaker discrimination using this feature.

Individual-level analysis conducted in Section 7.1.3 further demonstrates that stronger performance in French is not merely driven by individual outliers but a more general finding applicable to most speakers. Zooplot analysis shows that there were indeed a number of outliers with extreme scores in each system, which in some cases constituted a significant presence of doves. Acoustic analysis in Section 7.1.4 indicates that such speakers were highly distinctive in their production of /s/, in clear contrast with worms (and chameleons) whose spectral moments were much lower in distinctiveness and who performed poorly in both SS and DS comparisons. At the same time, even when taking into account speakers with particularly poor performance, speakers in French as a group produced DS-LLR of much higher magnitude. Although performance of SS comparisons in the two languages did not show such drastic differences, there were generally fewer speakers found to have a negative mean SS-LLR (and hence could not be matched with themselves on average) in French.

Regarding the relationship between different spectral moments, the relatively stronger discriminatory power of CoG and skewness over SD and kurtosis was similarly demonstrated on an individual level, particularly in terms of their performance in DS comparisons. In systems using CoG and skewness, the evidence could provide moderate support in favour of the different-speaker hypothesis for the majority of speakers in both languages. The strength of evidence in systems using SD and (log-)kurtosis was much weaker, as evidenced by the overall smaller DS-LLR. Poorer performance of SD and kurtosis in SS comparisons, while not as substantial, could be attributed to a higher proportion of speakers who produced negative mean SS-LLR and were on average difficult to match with themselves. The complementarity of different spectral moments is evident not only from the gradual improvement of C_{llr} and EER as more spectral moments were combined in a system, but also from the highly distinct speaker classifications summarised in Figure 7.10, as each spectral moment captured a largely different group of speakers with outlying performance. There is nevertheless some evidence that the acoustic correlation between CoG and skewness, and between SD and kurtosis, percolates through to discriminatory performance of individual speakers, which is especially clear seen from the overlap of doves between each of these pairs of spectral moments. As Figures 7.11 to 7.14 show, speakers with a distinctively low CoG or SD simultaneously show /s/ with a distinctively high value of skewness or kurtosis, consequently producing outlying performance in both.

7.2 Experiment 2: Cross-language comparisons

7.2.1 Global metrics

7.2.1.1 Language mismatch

Figures 7.15 and 7.16 present C_{llr} and EER obtained from En–Fr systems, where an English QS was compared with a French KS, with reference to those from same-language comparisons. Figure 7.15 shows that, across all tested parameter combinations, systems in both cross-language conditions reported higher C_{llr} than corresponding En–En and Fr–Fr systems. Meanwhile, systems in Condition C1, where the language of the background data matched with the KS, and Condition C2, where the language of the background data matched with the QS, produced very similar C_{llr} values. Across the various modes of data input tested, the relative hierarchy of performance established in Experiment 1 appears to be largely preserved. In the systems using individual spectral moments, midpoint-only systems yielded higher C_{llr} than either set of dynamic systems. The pattern was reversed in the combined system, with generally lower C_{llr} in the midpoint-only system than in the quadratic and three-point systems.

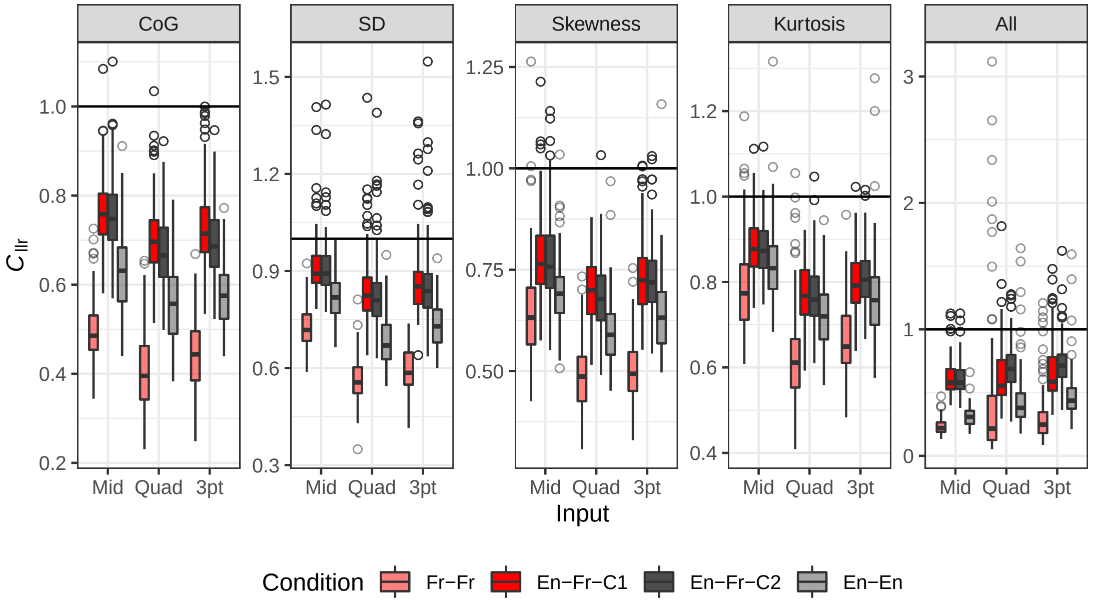

Figure 7.15: C_{llr} from all tested En–Fr /s/ systems in Conditions C1 and C2, with reference to same-language comparisons in English (En–En) and French (Fr–Fr). Horizontal rule for reference at $C_{\text{llr}} = 1$.

The effect of condition on C_{llr} was not uniform across different modes of input, as a significant interaction between condition and input was found in the models fitted to each set of parameters (see Table D.1 in Appendix D for full results of model comparisons). C_{llr} rose more sharply in systems using dynamic input when compared to systems using static input, as demonstrated by the example of SD in Figure 7.17. Post-hoc Tukey-adjusted pairwise comparisons by input confirmed that differences between En–Fr systems in either condition C1 or C2 and systems with no language mismatch were all significant ($p < .0001$). The same comparisons further found no significant differences between C1 and C2 ($p > .05$) in each case, except for the pairs in the dynamics systems for CoG, where C_{llr} in Condition C2 was lower than that in Condition C1 by 0.035 in the case of quadratic input and by 0.037 in the case of three-point input.

Figure 7.16 shows that the measure of EER largely followed the same patterns as C_{llr}, with similar EERs reported in the two cross-language conditions higher than those in the same-language conditions, especially when compared to the Fr–Fr systems. Model

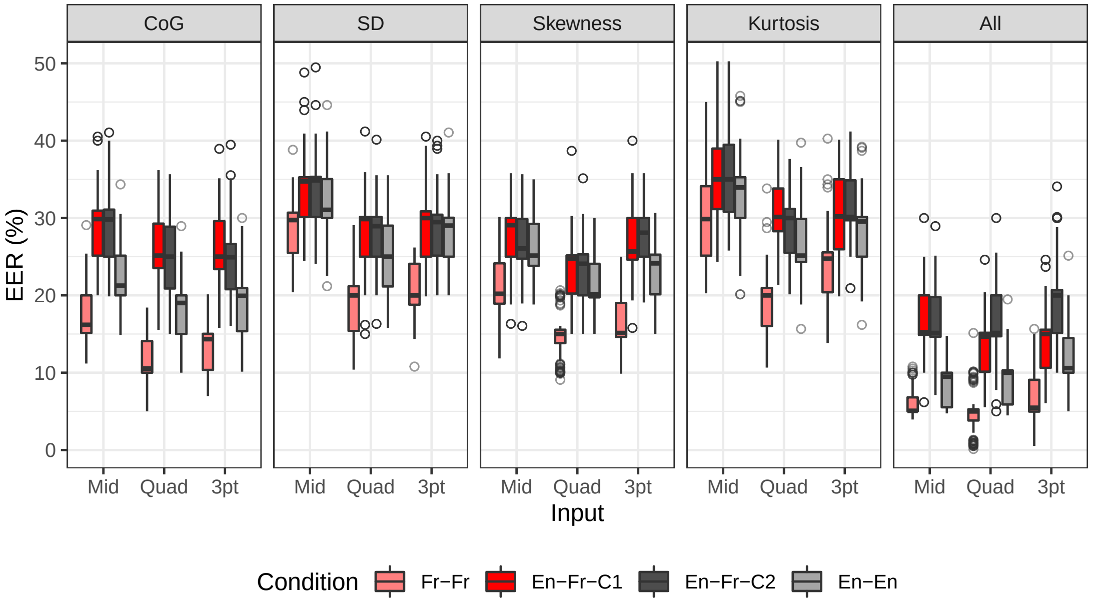

Figure 7.16: EER from all tested En–Fr /s/ systems in Conditions C1 and C2, with reference to same-language comparisons in English (En–En) and French (Fr–Fr).

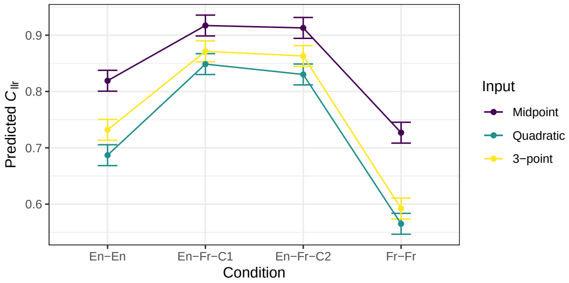

Figure 7.17: Predicted C_{llr} for each condition by mode of input in systems using SD.

comparisons similarly found a significant interaction between condition and mode of input for all sets of parameters. Post-hoc Tukey-adjusted pairwise comparisons by input showed that there were no significant differences between Conditions C1 and C2 in all sets of systems, with the exception of the combined systems using dynamic input, where Condition C1 produced EER lower than Condition C2 by 2.6% and 4.6% in the quadratic and three-point systems respectively. EER from both cross-language comparisons was consistently higher than those from corresponding Fr–Fr systems, and was in most cases, but not always, significantly higer than those from corresponding En–En systems. Specifically, in the midpoint-only and three-point systems for SD, no significant differences were found between En–En and the two En–Fr conditions: In the midpoint-only systems for skewness, EER from the En–En system was only significantly lower than EER from Condition C1 but not Condition C2.

These findings suggest that, overall, in cross-language comparisons with an English QS and a French KS, the performance of the the same parameters for /s/ was weaker than in same-language comparisons. On the other hand, modelling the background in either the KS or the QS language made little to no difference to system performance as measured by C_{llr} and EER.

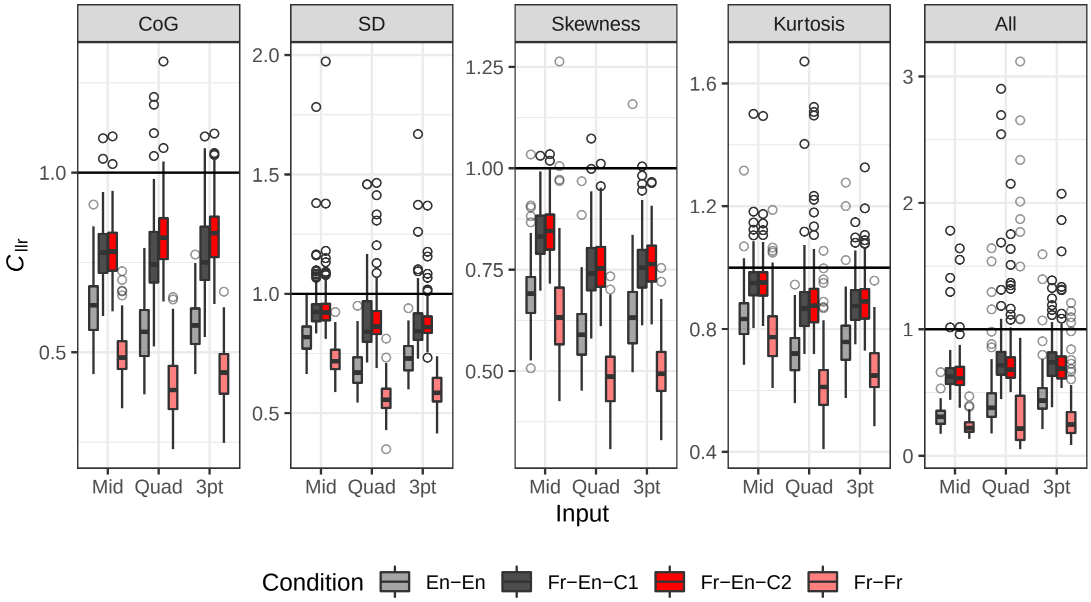

Figure 7.18: C_{llr} from all tested Fr–En /s/ systems in Conditions C1 and C2, in addition to same-language comparisons in English (En–En) and French (Fr–Fr). Horizontal rule for reference at $C_{\text{llr}} = 1$.

Results from the opposite language pairing (i.e., French QS and English KS) are presented in Figures 7.18 and 7.19. The general trends for C_{llr} and EER found in these systems show remarkable similarities to those observed above, with the notable anomaly in CoG that dynamic input did not result in stronger performance than static input. Statistical analysis confirmed the same differential effects of condition on C_{llr} and EER across different modes of input, as suggested by the common finding of a significant interaction of the two factors in the mixed-effects models fitted. The sole exception was the model fitted to C_{llr} from the combined systems, in which each factor was independently significant, but the interaction was not. As in the comparisons above involving En–Fr systems, Conditions C1 and C2 were not found to yield significantly different C_{llr} in the post-hoc Tukey-adjusted pairwise comparisons except for the dynamic systems using CoG: C_{llr} was lower in Condition C1 by 0.059 in the case of quadratic input and 0.070 in the case of three-point input. Post-hoc comparisons similarly found no differences in EER between Conditions C1 and C2 in all cases. Without exception, the differences between the cross-

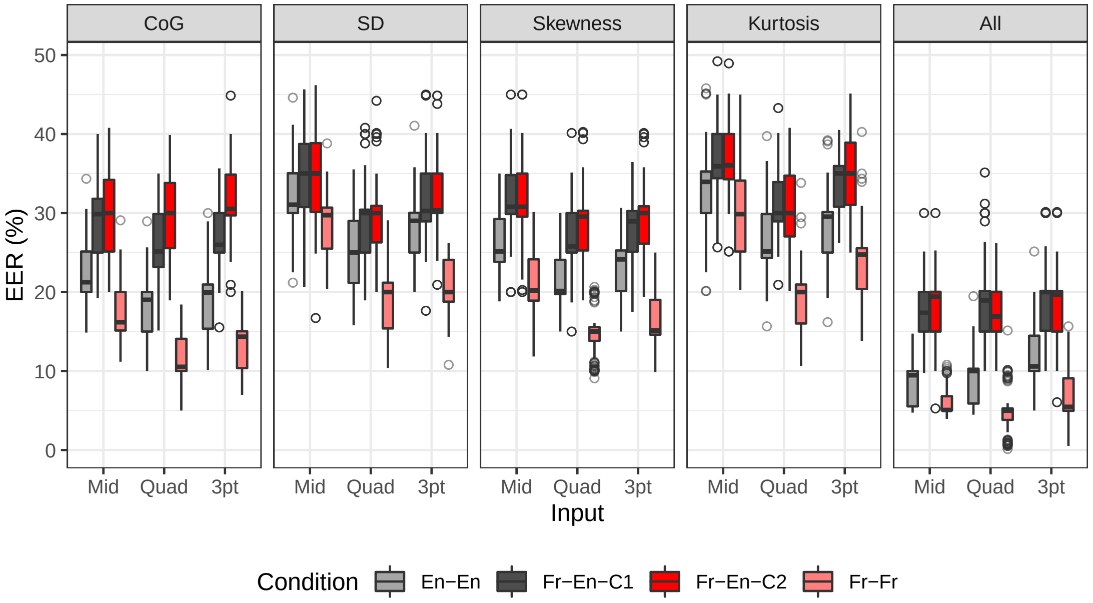

Figure 7.19: EER from all tested Fr–En /s/ systems in Conditions C1 and C2, in addition to same-language comparisons in English (En–En) and French (Fr–Fr).

language and the same-language conditions were shown to be significant. These results thus closely mirror what was found for the other language pairing. Nonetheless, there were minor quantitative differences in the values of validity metrics obtained in En–Fr and Fr–En systems, in favour of the former, as Fr–En systems reported higher C_{llr} than En–Fr systems in each set of parameters (Table 7.2).

When compared to same-language comparisons, the deterioration in system performance was further evidenced by an expansion of the range of C_{llr} and an increased number of replications with $C_{llr} > 1$. Among systems using individual spectral moments, such effects were relatively small for CoG and skewness, but particularly prominent for SD, where C_{llr} exceeded 1 in 8–11 replications in En–Fr systems and 9–20 replications in Fr–En systems, reaching a maximum of 1.55 and 1.97 respectively. Similarly, for kurtosis, the Fr–En systems recorded 10–22 replications with $C_{llr} > 1$. While overall C_{llr} and EER were lower for the combined systems, there were numerous replications yielding $C_{llr} > 1$, particularly in the case of dynamic input (En–Fr: 6–12; Fr–En: 9–11). The maximum C_{llr} out of all replications from the combined systems was also much higher than systems using individual moments in each pairing (Table 7.2; but note the highest C_{llr} outlier of 3.12 in the Fr–Fr systems). Therefore, /s/ performed more poorly in cross-language comparisons than same-language comparisons not only in terms of system validity, but also in terms of system stability.

Table 7.2: Mean (maximum) C_{llr} and EER from En–Fr and Fr–En /s/ systems for each parameter combination, aggregated over all modes of data input and Conditions.

	En–Fr		Fr–En	
	C_{llr}	EER	C_{llr}	EER
CoG	0.72 (1.10)	26.3% (41.1%)	0.80 (1.31)	28.8% (44.9%)
SD	0.87 (1.55)	30.2% (49.5%)	0.91 (1.97)	32.0% (46.2%)
Skewness	0.74 (1.21)	25.8% (40.0%)	0.79 (1.07)	29.5% (45.0%)
Kurtosis	0.82 (1.12)	32.1% (50.2%)	0.92 (1.67)	34.0% (49.2%)
All	0.66 (1.82)	16.1% (34.1%)	0.73 (2.90)	18.3% (35.1%)

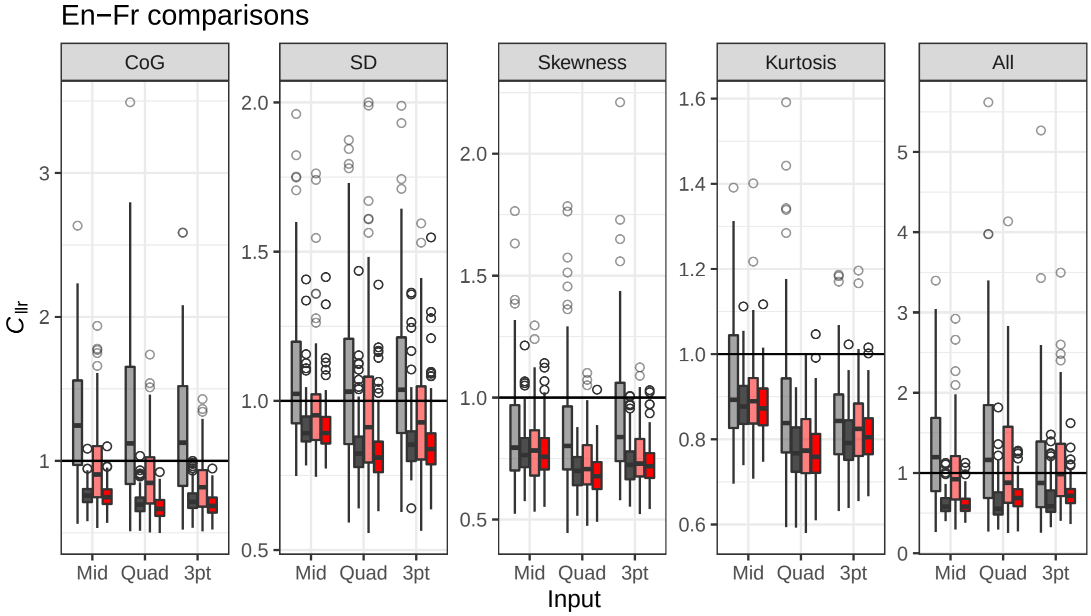

Figure 7.20: C_{llr} from all tested En–Fr /s/ systems in Conditions C1 (grey) and C2 (red), with (light) and without (dark) training mismatch. Horizontal rule for reference at $C_{\mathrm{llr}} = 1$.

7.2.1.2 Training mismatch

This section turns to present findings on the effect of mismatch between training data and test circumstances. Figure 7.20 illustrates C_{llr} from systems of En–Fr comparisons with and without training mismatch and Figure 7.21 illustrates C_{llr} from systems of Fr–En comparisons. As mentioned in Section 6.3, EERs were not analysed because using different sets of training data for calibration in the score-to-LR stage does not change the system's EER.

As both Figures show, systems with training mismatch generally performed worse than those without training mismatch. In most cases, C_{llr} rose to above 1, suggesting that the systems were poorly calibrated and could not provide useful speaker-specific information. Systems using only CoG or all four spectral moments suffered the effects of

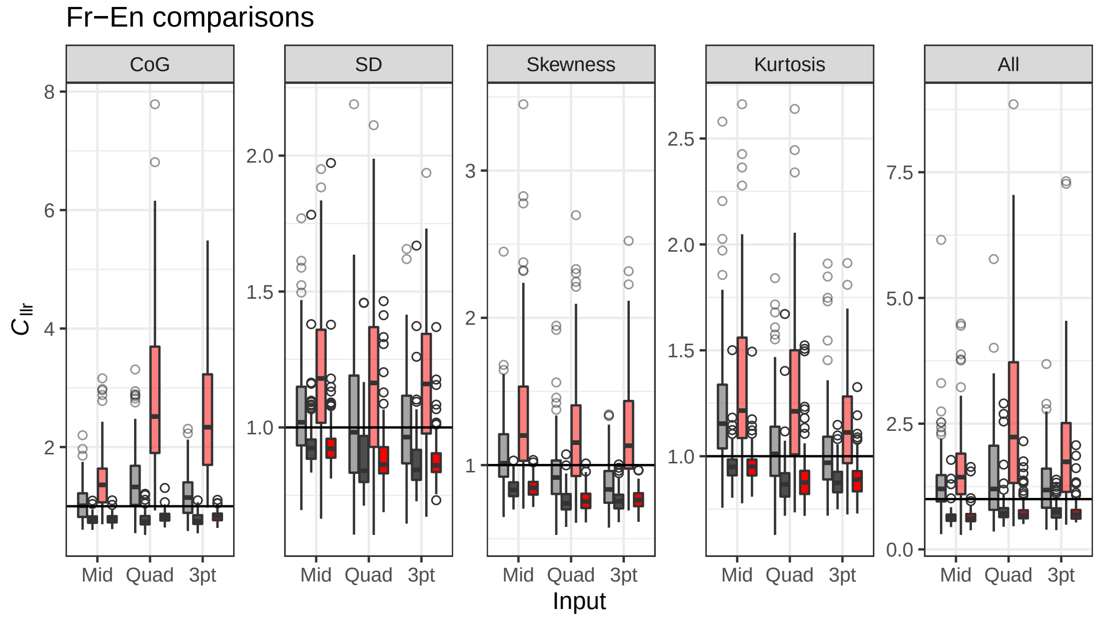

Figure 7.21: C_{llr} from all tested Fr−En /s/ systems in Conditions C1 (grey) and C2 (red), with (light) and without (dark) training mismatch. Horizontal rule for reference at $C_{\text{llr}} = 1$.

training mismatch particularly badly. Its impact not only manifested in overall high C_{llr}, but was also patent in the extremely wide range of C_{llr}, extending up to 3.49 and 7.79 in En−Fr and Fr−En comparisons respectively for CoG, and up to 5.62 and 8.85 for the combined systems. Other spectral moments in En−Fr comparisons, as well as those in Condition C1 in Fr−En comparisons, appear to be relatively unaffected. Their C_{llr} remained around or even below 1 on average, pointing to their limited utility in speaker discrimination even though these systems were not in fact poorly calibrated. Even so, these systems shared the expansion of range of C_{llr} when compared to corresponding systems with appropriate calibration with case-matched training data. Unlike results from earlier parts of both experiments, no consistent trends could be found for mode of data input in systems with training mismatch. In comparison to static systems, slightly stronger performance in dynamic systems could only be observed in a limited subset of systems using individual moments. In other cases, system performance of dynamic systems with training mismatch was either highly similar or worse than that of corresponding static systems.

Statistical analysis by LMEM shows that, in the En−Fr systems, significant two-way interactions of training mismatch with both condition and mode of input were found for CoG and SD, but not the interaction between condition and mode of input (full results for model comparisons in Table D.2). Figure 7.22 shows that, in both cases, C_{llr} experienced a greater upward shift due to training mismatch in Condition C1 than in Condition

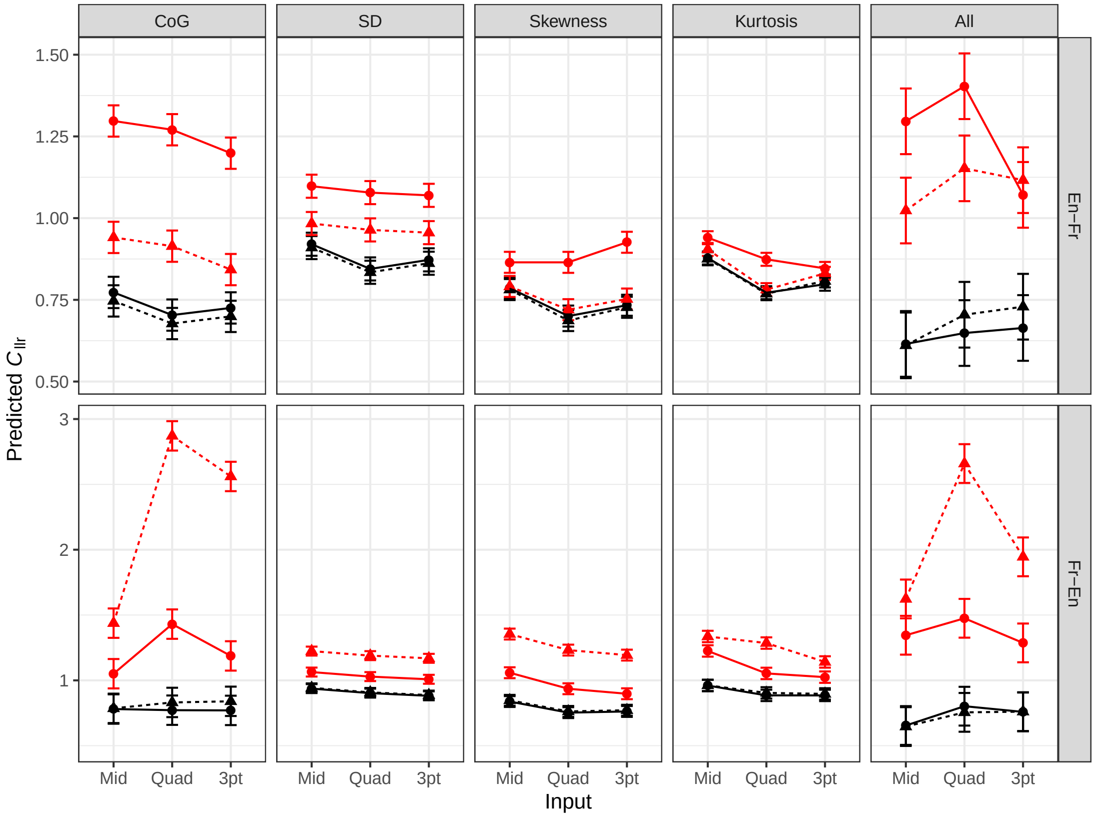

Figure 7.22: Model predictions for all tested parameter combinations in En–Fr and Fr–En comparisons. Systems with training mismatch in red and those with no mismatch in black. Conditions C1 in circles joined by solid lines and C2 in triangles joined by dashed lines.

C2, while the effect of mismatch was largest for quadratic systems and smaller (but to different degrees) for midpoint-only and three-point systems. A more complex three-way interaction between all three fixed factors was found for the remaining sets of parameters, meaning that the differential impact of training mismatch in each condition was further dependent on the mode of input in question. For both skewness and kurtosis, training mismatch had negligible effects in Condition C2, but resulted in higher C_{llr} in Condition C1, especially for one (quadratic, for kurtosis) or both dynamic modes of input. For the combined systems, however, both midpoint-only and quadratic forms of input were affected by training mismatch to a greater extent in Condition C1, whereas three-point systems showed no such distinction.

As for the Fr–En systems, it is clear from the bottom row of Figure 7.22 that training mismatch had a greater impact on Condition C2 than Condition C1, resulting in higher C_{llr}. Indeed, model comparisons for all sets of parameters found either a significant two-way interaction between condition and training mismatch or a significant three-way interaction of the two factors as well as mode of input. A three-way interaction for CoG indicates that the differential effect of training mismatch on the two conditions was particularly pronounced for the two systems using dynamic input, with C_{llr} reaching close to 3, the highest out of all systems. The three-way interaction was similarly shown to be significant for kurtosis, though in this case, as is evident from Figure 7.22, the increase in C_{llr} arising from training mismatch was greater in Condition C2 than in Condition C1 only when quadratic input was used but not the midpoint-only or three-point systems. The only other set of parameters for which there was a significant three-way interaction is the combination of all spectral moments, where C_{llr} followed similar patterns to the case of CoG, namely that the systems using dynamic input, especially quadratic input, experienced a steeper increase than the static systems in Condition C2 over Condition C1. For SD, only a main effect of input was found alongside the two-way interaction between condition and training mismatch, whereas in the case of skewness mode of input also significantly interacted with training mismatch but not with condition, suggesting the rise in C_{llr} due to training mismatch was smaller for the three-point systems using skewness than other modes of input in both conditions.

7.2.1.3 Interim discussion

The results presented in this section demonstrate the clear impact of language mismatch on the speaker-discriminatory potential of acoustic parameters of /s/, even when systems are appropriately calibrated. Compared to same-language comparisons, systems in cross-language comparisons generally yielded C_{llr} that was both higher and of a wider range, alongside higher EER, although the extent to which systems were impacted varied across spectral moments and modes of data input. While the finding that mean C_{llr} generally does not rise above 1 suggests that the utility of these parameters is not fully neutralised in cases of cross-language comparison, increases in both EER and C_{llr} mean that the strength of evidence that can be offered is much lower than in cases of same-language comparison.

The deleterious impact of miscalibration due to mismatch in the conditions (specifically, languages) of the case and of the reference data is evident in the high C_{llr} in all systems in Conditions C1b and C2b, in which the systems were calibrated using scores from same-language comparisons instead of from cross-language comparisons, and is particularly marked for systems using CoG, whether on its own or in combination with other spectral moments. In the majority of cases, the use of acoustic parameters of /s/ in miscalibrated cross-language comparisons simply provided no useful speaker-specific information. Although training mismatch did not severely impact C_{llr} from skewness and kurtosis in En–Fr comparisons, these systems were nevertheless considerably less reliable than their counterparts that were appropriately calibrated. These findings imply that, in the event that a forensic analyst were tasked to compare samples that were in different languages but only had access to reference materials in one of those languages, it would not be possible to set up a valid system using acoustic parameters of /s/ for a reliable comparison to be carried out.

Theoretically, as Morrison (2018) argues, the conditions of the background data should match the circumstances of the KS rather than the QS, so that any effects of mismatch between the QS and the KS can be balanced by the same effects of mismatch between the QS and the background data. In cases of channel mismatch, matching the conditions of the background data with those of the KS rather than the QS resulted in substantially better system performance (Morrison, 2011). In the present case of language mismatch, however, at least for /s/, matching the language of the background data with the KS versus the QS resulted only in minimal differences. C_{llr} and EER obtained from Condition C1, where the background data were indeed matched in language with the KS, was significantly lower than those from Condition C2 only in the dynamic systems for CoG but not in other sets of parameters. These findings indicate that, cross-linguistically, the distributions of /s/ spectral moments in the reference population may be considered to be sufficiently similar, such that comparing a QS with background data in either language resulted in similar LRs. This conclusion follows from findings in Chapter 5 that English–French bilinguals did not produce /s/ in the two languages with distinct midpoint spectral moments, which may be a reason why the current findings apparently diverged from Morrison (2018). The recordings in the current study differed only in the language of the materials but not in other conditions that may have had an impact on the acoustics, whereas the recordings in Morrison (2011) involved not only different channels (mobile vs landline) but also other discrepancies in the recording environment and circumstances.

Differences between the two conditions did emerge when training mismatch was introduced to the systems. Its effect was considerably stronger in Condition C1 for En–Fr comparisons and in Condition C2 for Fr–En comparisons. The effect of training mismatch thus did not simply depend on whether the language of the reference data was matched with that of the QS or the KS, but on the language itself. In both language pairings, converting scores to LRs using training scores from same-language comparisons in English resulted in substantially poorer C_{llr} than using scores from same-language comparisons in French, which may be an indication that the stronger discriminatory performance of /s/ in French (in same-language comparisons) can also be carried over to some extent even when training data in a single language are used to convert scores in cross-language comparisons to LRs.

7.2.2 Individual-level analysis

In this section, the effect of language mismatch on individual performance is explored. Following the approach taking in Section 7.1.3, the analysis in this section focuses on midpoint-only systems. Further, only systems with no training mismatch (i.e., systems calibrated using training scores from cross-language comparisons themselves) were considered, in order to reliably assess the effect of the primary factor of interest (language mismatch). Figure 7.23 provides an illustration of the effect of training mismatch on individual performance. The zooplots show a virtually identical distribution of speakers when the same set of test scores were calibrated using different sets of training scores, but shifted and scaled to different extents, as evident from the scales of the axes in each plot.[1] Scores from cross-language comparisons that were calibrated with training scores from same-language comparisons, be it from English or French, resulted in LLRs that were shifted in the negative direction and at the same time more exaggerated in magnitude, in both SS and DS comparisons.

Zooplots for both Conditions C1 and C2 were constructed, but a comparison of speaker classification between the two conditions shows that they produced highly similar speaker distributions. As evidenced by the summary of zoo memberships in Figure 7.24, classification of speakers into different relational groups is nearly identical across all analysed systems. Therefore, only zooplots from Condition C1, where the language of the background data was matched with the KS language, are shown and analysed in further detail.

7.2.2.1 CoG

Zooplots for both En–Fr and Fr–En comparisons using CoG are presented in Figure 7.25. In contrast to zooplots for same-language comparisons in Section 7.1.3, speakers can be found concentrated towards the lower right corner. Overall, no significant correlation could be found between speakers' SS and DS performance in En–Fr comparisons ($r = -.04,\ p = .7880$), whereas SS and DS performance in Fr–En comparisons was weakly correlated ($r = -.26,\ p = .0461$). All speakers produced negative mean DS-LLR, and while the majority of speakers produced positive SS-LLR, negative mean SS-LLR was

[1] Speaker distributions are not fully identical in these zooplots as they were constructed from 100 separately calibrated replications, resulting from the cumulative effect of training mismatch having slightly different shifting and scaling impact on each replication.

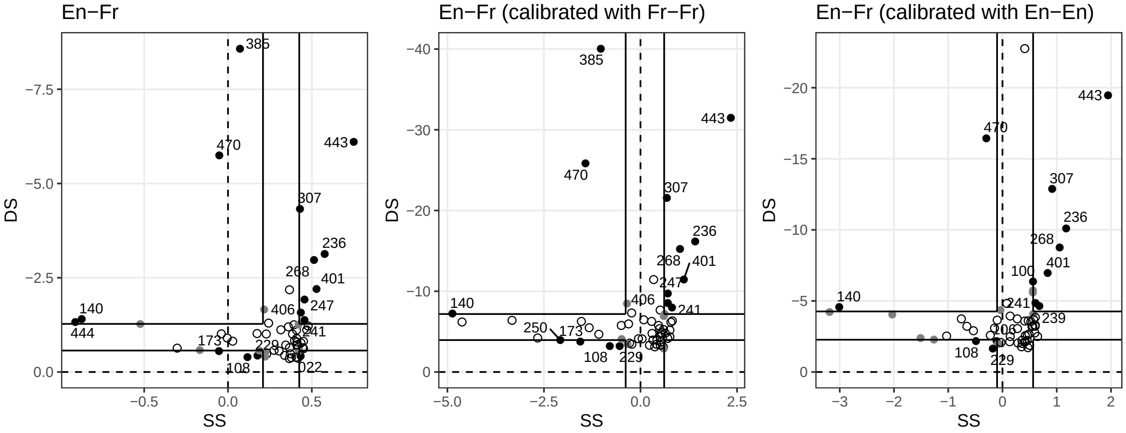

Figure 7.23: Zooplots with speaker labels for systems with midpoint CoG as input, calibrated by training data from En–Fr (left), Fr–Fr (middle) and En–En (right) comparisons.

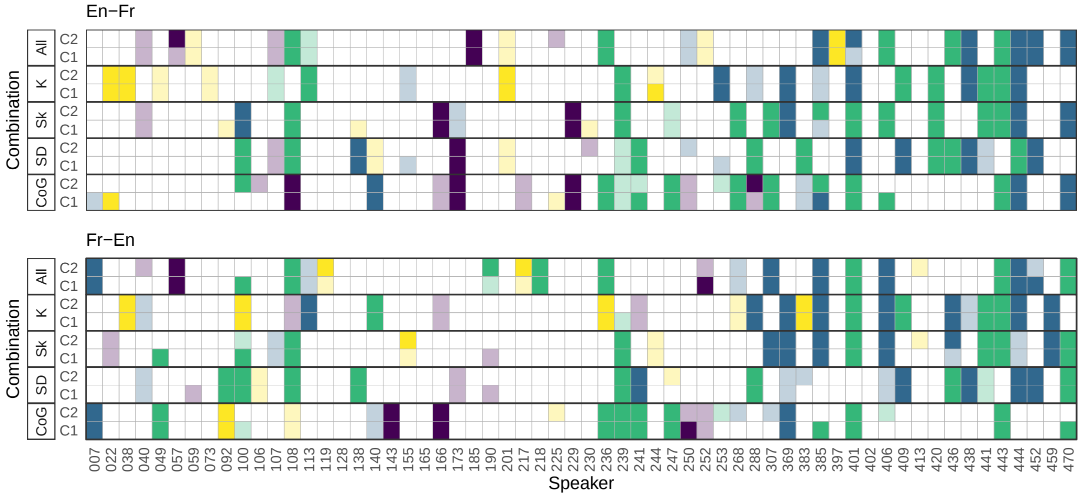

Figure 7.24: Summary of speakers classified as doves (green), worms (purple), phantoms (blue) or chameleons (yellow) in Conditions C1 and C2 for all /s/ systems tested (near-members of each group in lighter shade).

found for a number of speakers (En–Fr: 9; Fr–En: 10). Regardless of polarity, the magnitude of mean SS-LLR was small for all speakers, mostly lying below 0.5 and never exceeding 1 with the exception of one speaker. Mean DS-LLR was similarly small for most speakers, although a small number of clear outliers who fell into the dove or phantom group produced much larger DS-LLR. As summarised in Table 7.3, no relational group was empty and the doves remained the largest group in either En–Fr or Fr–En comparison.

To further analyse the effect of language mismatch on individual performance, a “normalised” version of the zooplot was constructed to show each speaker’s relative po-

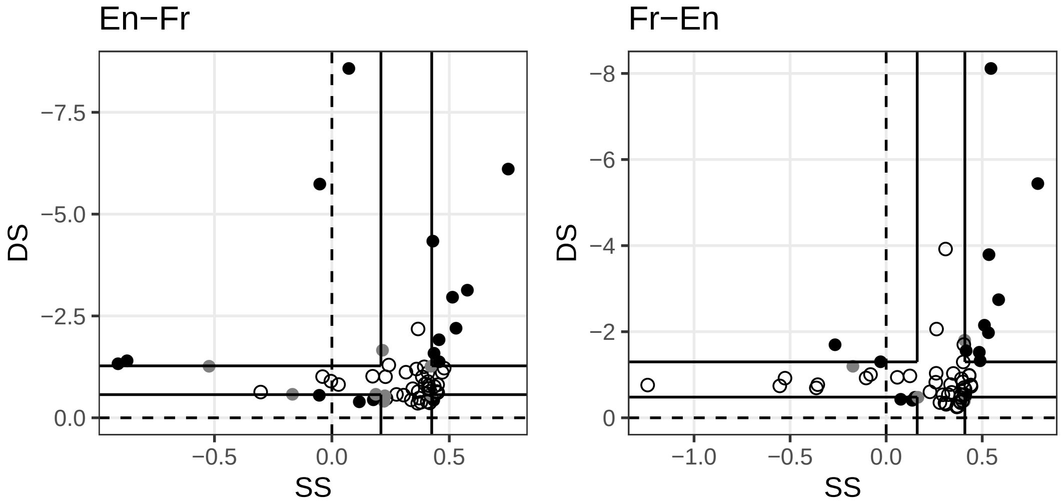

Figure 7.25: Zooplots for systems with midpoint CoG as input in cross-language En–Fr (left) and Fr–En (right) comparisons.

Table 7.3: Number of speakers in each relational group for systems with midpoint CoG as input in cross-language comparisons.

Comparison	Dove	Worm	Phantom	Chameleon
En–Fr	8	3	4	1
Fr–En	9	3	2	1

sition in the population for same- and cross-language comparisons. This was accomplished by plotting the percentile rank of each speaker's mean SS-LLR and DS-LLR, rather than plotting the absolute values themselves, such that the locations of the relational groups in different systems could be aligned. In this case, the main comparison of interest is between same- and cross-language comparisons when the KS language is fixed, so that the impact of comparing samples from the same known speaker (suspect) to QS in different languages could be examined.

The normalised zooplots for CoG are presented in Figure 7.26, with outlying speakers who were classified as members of relational groups in same- and/or cross-language comparisons highlighted for clarity. In both plots, when the KS was in French (left panel) or in English (right panel), an overall similar pattern of speaker distribution could be observed, with movement between same- and cross-language comparisons primarily found horizontally along the SS dimension. Speakers in the top half of the plots, who performed relatively well in DS comparisons, generally shifted towards the left, in the direction of relatively worse performance in SS comparisons. While some speakers classified as doves in same-language comparisons maintained strong performance in cross-

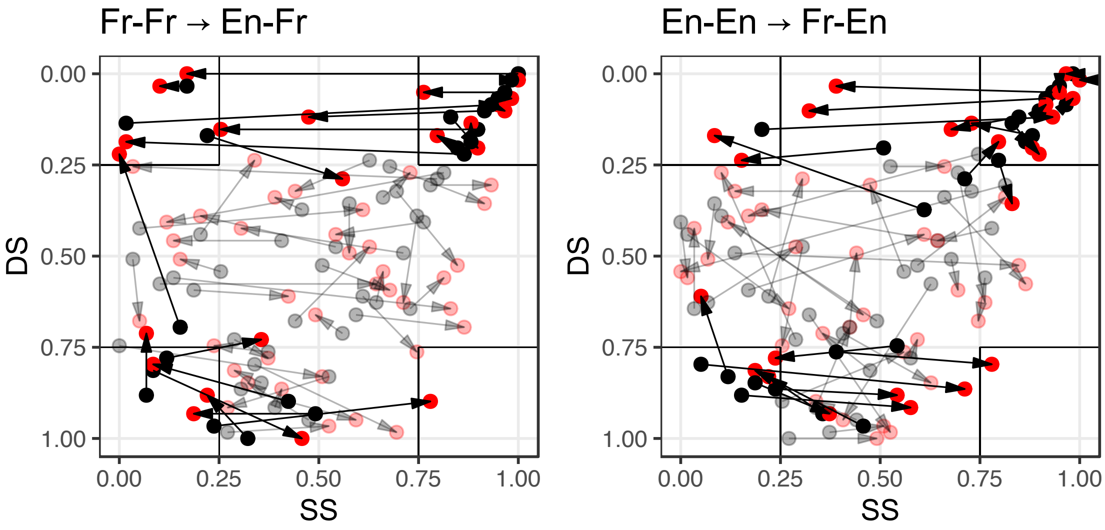

Figure 7.26: Normalised zooplots for systems with midpoint CoG as input in same-language (black) and cross-language (red) comparisons. Values on axes indicate percentile ranks. Arrows connect speakers from same-language to cross-language comparisons.

language comparisons, others were particularly affected by language mismatch and moved to the left half of the normalised zooplots in cross-language comparisons. Speakers in the bottom half of the plots, on the other hand, more commonly shifted in the other direction. Unlike the dove group, almost no speaker who was classified as a worm in same-language comparisons was in the same relational group in cross-language comparisons, although their performance in DS comparisons remained relatively poor. These patterns of movement suggest that, aside from the few speakers with exceptional performance in both same- and cross-language comparisons, speakers with relatively good DS performance in same-language comparisons were more susceptible to the adverse effects of language mismatch, particularly in SS comparisons.

7.2.2.2 SD

The zooplots for cross-language comparisons using SD, illustrated in Figure 7.27, similarly depict a dense cluster of speakers near the lower right corner and a distinct lack of speakers in the lower left corner. This is partially reflected in the low number of speakers classified as worms, although the chameleon group was nonetheless empty in both cases (Table 7.4), suggesting that speakers with the highest mean SS-LLR did not at the same time produce the weakest mean DS-LLR. SS and DS performance was nevertheless found to be weakly negatively correlated in Fr–En comparisons ($r = .29,\ p = .0257$), although no significant correlation was found in En–Fr comparisons ($r = .25,\ p = .0534$).

The magnitude of mean SS-LLR and DS-LLR was small for all speakers with no extreme outliers, with most speakers clustering near DS-LLR = 0, pointing to the weak performance of this parameter for most speakers. In addition to 12 speakers producing negative mean SS-LLR in either En–Fr or Fr–En comparison, there was also one speaker producing positive mean DS-LLR, adding further evidence to the detrimental impact of language mismatch to both types of comparisons.

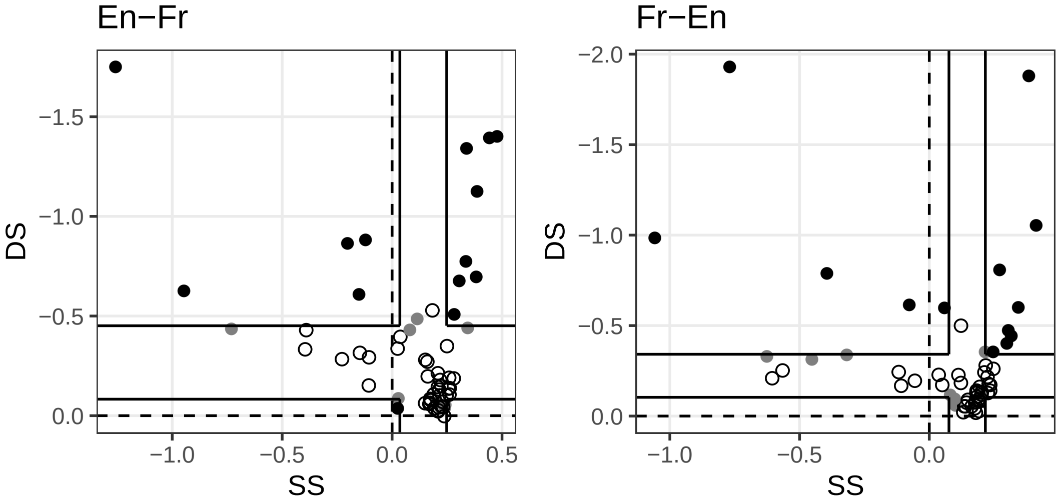

Figure 7.27: Zooplots for systems with midpoint SD as input in cross-language En–Fr (left) and Fr–En (right) comparisons.

Table 7.4: Number of speakers in each relational group for systems with midpoint SD as input in cross-language comparisons.

Comparison	Dove	Worm	Phantom	Chameleon
En–Fr	8	1	5	0
Fr–En	8	0	5	0

Figure 7.28, which presents the normalised zooplots for SD in same- and cross-language comparisons, shows considerable movement displayed by most speakers within each plot, particularly among speakers with relatively good DS performance. Both plots show an overall speaker distribution across both same- and cross-language comparisons in the configuration of an inverted triangle, with speakers towards the bottom of the plot remaining relatively stable across. Stability was particularly low for speakers with the best DS performance in the dove or chameleon groups, as speakers identified as chameleons in cross-language comparisons were mostly dove members in same-language comparisons or otherwise speakers with similarly good performance in SS comparisons, whereas

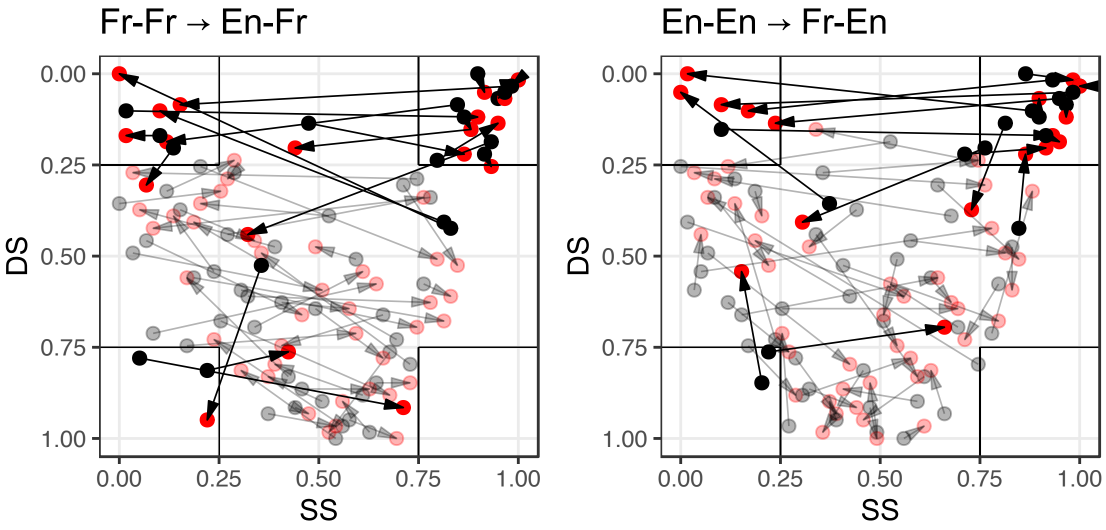

Figure 7.28: Normalised zooplots for systems with midpoint SD as input in same-language (black) and cross-language (red) comparisons.

doves in cross-language comparisons comprised a mixture of speakers who were doves or chameleons in same-language comparisons and other speakers with generally poorer SS performance in same-language comparisons.

7.2.2.3 Skewness

The distribution of speakers for skewness, shown in the zooplots in Figure 7.29, is broadly in line with that for CoG or SD, with a high concentration of speakers close to SS-LLR and DS-LLR $= 0$ located towards the lower right corner of the zooplots. Only a few outlying speakers produced mean SS-LLR or DS-LLR of higher magnitudes, most of whom classified as either doves or phantoms. Like SD, these systems were characterised by the complete absence of chameleons and a low number of worms, with doves and phantoms forming the largest groups (Table 7.5). SS and DS performance was weakly positively correlated in En–Fr comparisons ($r = -.39,\ p = .0020$), but was not significantly correlated in Fr–En comparisons ($r = -.20,\ p = .1321$). Although no speakers produced positive mean DS-LLR, the number of speakers with a negative mean SS-LLR in Fr–En comparisons (14) was the highest among all tested systems, indicating consistently poor individual performance in SS comparisons.

When individual performance is compared across same- and cross-language comparisons, Figure 7.30 demonstrates that speakers with the best DS performance in same-language comparisons also showed a tendency to perform well in DS comparisons when there was language mismatch. Vertical movement of speakers near the top of both nor-

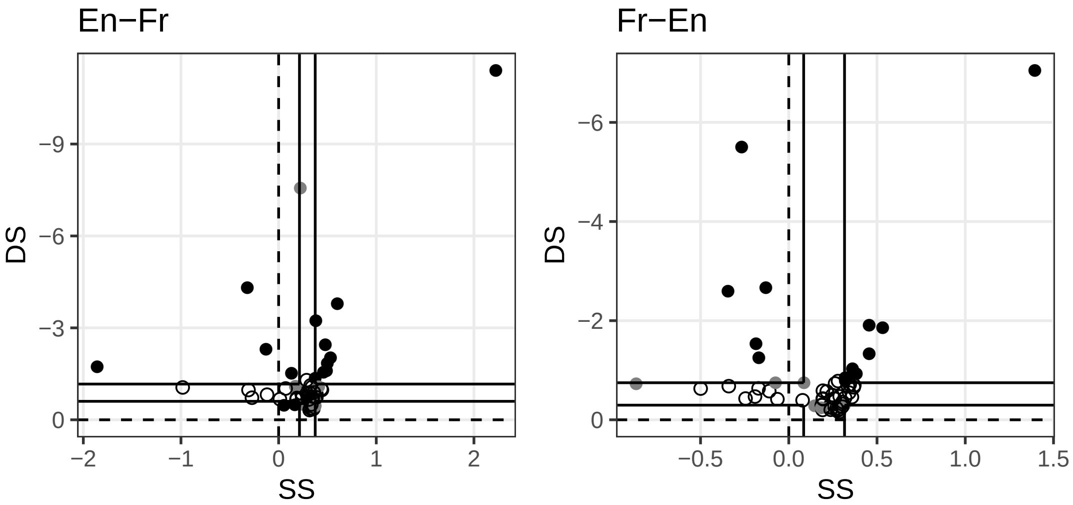

Figure 7.29: Zooplots for systems with midpoint skewness as input in cross-language En−Fr (left) and Fr−En (right) comparisons.

Table 7.5: Number of speakers in each relational group for systems with midpoint skewness as input in cross-language comparisons.

Comparison	Dove	Worm	Phantom	Chameleon
En−Fr	9	2	4	0
Fr−En	8	0	5	0

malised zooplots was limited, while speakers largely fell into two main groups in terms of their horizontal movement. One group of speakers, most of whom were doves, underwent negligible overall movement, so that their individual performance was relatively similar in same- and cross-language comparisons. The other group of speakers, all of whom were again doves in same-language comparisons with one exception, displayed considerable movement along the SS dimension, to the extent that they were located in or near the phantom region in cross-language comparisons. These trends largely echo the patterns of movement found for other spectral movements above. For speakers who were not among those who performed best in DS comparisons, there was substantial movement along both SS and DS dimensions, but there was otherwise no strong indications of a systematic pattern in how their individual performance varied.

7.2.2.4 Kurtosis

Zooplots for the fourth spectral moment, kurtosis, are presented in Figure 7.31. While the vast majority of speakers similarly formed a dense cluster near the lower right corner, when compared to other spectral moments, systems using kurtosis were marked

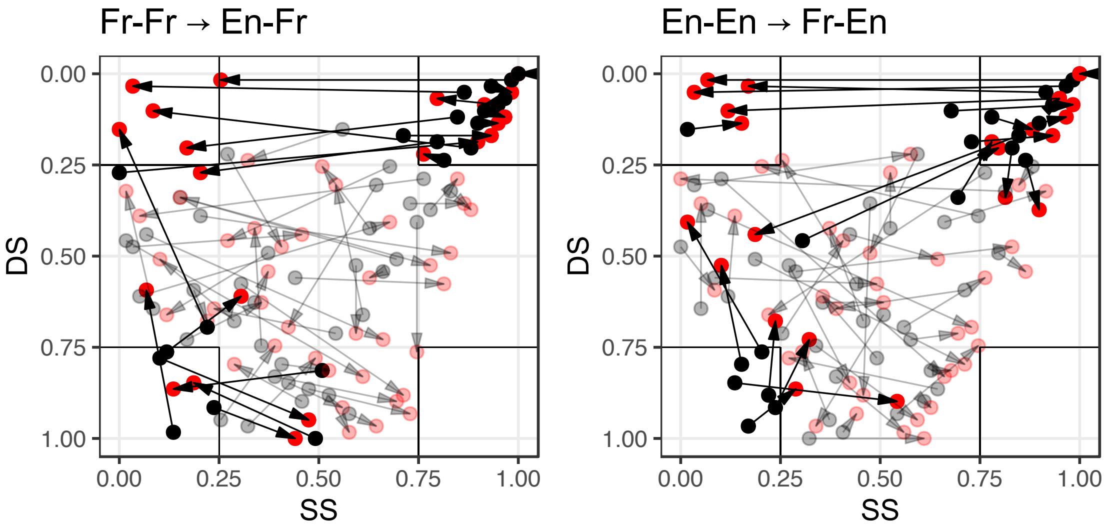

Figure 7.30: Normalised zooplots for systems with midpoint skewness as input in same-language (black) and cross-language (red) comparisons.

by a particularly sizeable group of phantoms. In fact, in Fr–En comparisons, phantoms overtook doves as the largest relational group (Table 7.6). With the exception of one outlier in each case, mean SS-LLR was constrained to a narrow range; the positions of the 25th and 75th percentiles delineating the boundaries of relational groups indicate that half of the speakers produced a mean SS-LLR between 0.05 and 0.24 in En–Fr comparisons and between 0.06 and 0.16 in Fr–En comparisons. The range of mean DS-LLR was similarly narrow, with its value exceeding −0.57 in En–Fr comparisons and −0.23 in Fr–En comparisons for only 25% of speakers. Correlation between SS and DS performance was significant but weak in En–Fr comparisons ($r = -.19,\ p < .0001$), and not significant in Fr–En comparisons ($r = -.19,\ p = .1372$). Particularly poor performance was evidenced for 11 individuals producing negative mean SS-LLR in both cases, as well as two individuals producing positive mean DS-LLR in Fr–En comparisons.

Table 7.6: Number of speakers in each relational group for systems with midpoint log-kurtosis as input in cross-language comparisons.

Comparison	Dove	Worm	Phantom	Chameleon
En–Fr	6	0	5	4
Fr–En	5	0	8	4

In the normalised zooplots for kurtosis, presented in Figure 7.32, most speakers underwent considerable movement in all directions between same- and cross-language comparisons. Outlying speakers who were members of relational groups in either same-

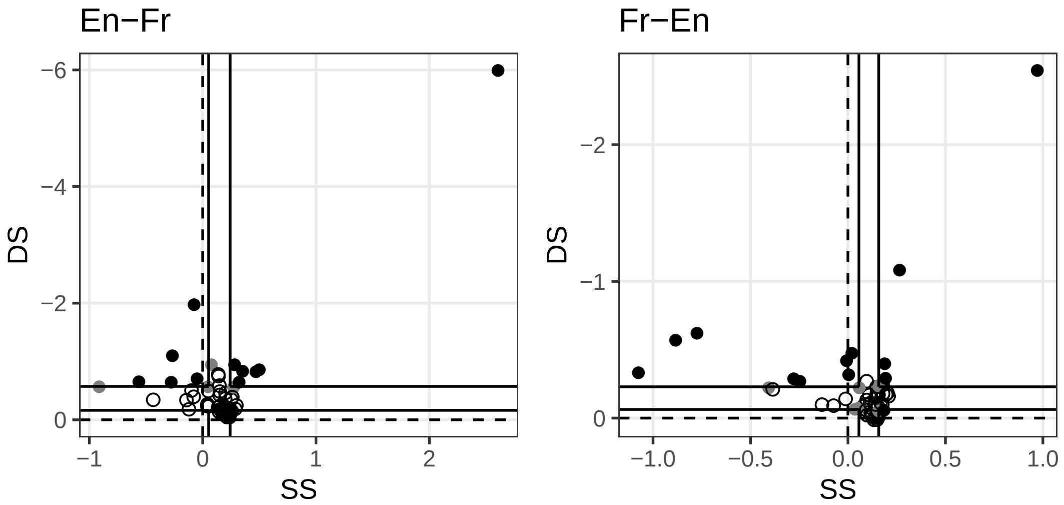

Figure 7.31: Zooplots for systems with midpoint log-kurtosis as input in cross-language En–Fr (left) and Fr–En (right) comparisons.

or cross-language comparisons appear to be most affected, where only very few speakers retained the same classification across both sets of comparisons. Doves in same-language comparisons demonstrated a tendency to be most adversely affected in SS comparisons but largely maintained good DS performance, thus shifting leftwards to the phantom region of the zooplots, though there were also individual exceptions whose DS performance also became relatively worse within the population. While there were few speakers near the top left corner in same-language comparisons, they had relatively unstable performance as a group and shifted to various other regions in the zooplots, meaning that the groups of phantoms identified in same- and cross-language comparisons had very little overlap. At the same time, speakers with the weakest DS performance in same-language comparisons generally also maintained weak performance in cross-language comparisons, with little upward movement seen in the zooplots. SS performance of these speakers relative to the population generally improved in cross-language comparisons, resulting in the emergence of the chameleons and the small group of worms in same-language comparisons becoming empty.

7.2.2.5 All moments

When all four spectral moments were combined in cross-language comparisons, zooplot analysis in Figure 7.33 shows that, despite the presence of extreme outliers, mean SS-LLR and DS-LLR remained generally small, with SS-LLR and DS-LLR exceeding 1 and −3 only for very few speakers. Meanwhile, eight and 10 speakers produced mean SS-

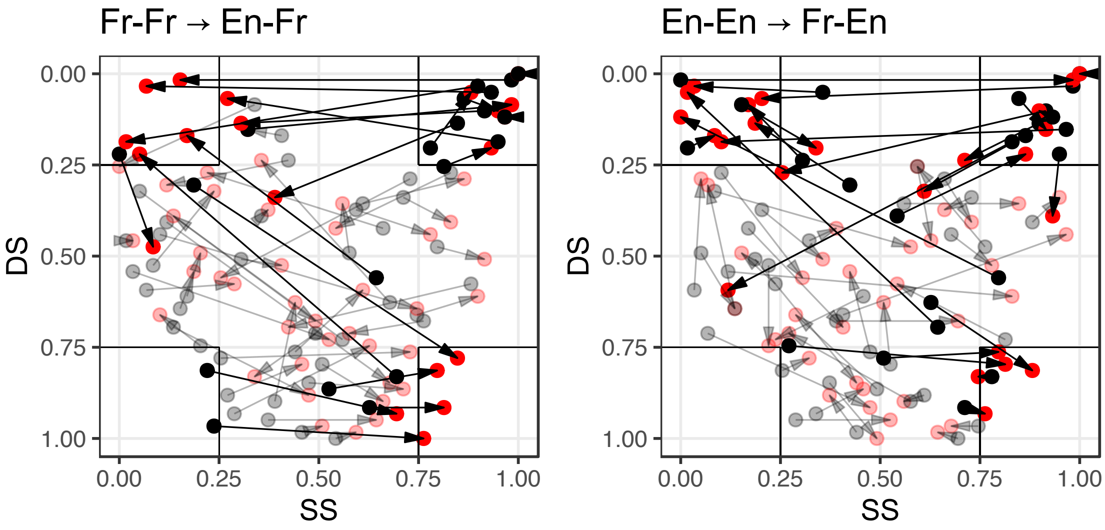

Figure 7.32: Normalised zooplots for systems with midpoint log-kurtosis as input in same-language (black) and cross-language (red) comparisons.

LLR below 0 in En–Fr and Fr–En comparisons respectively. While most speakers still produced similar levels of performance, their distribution in the zooplots here was more evenly spread out than in the case of individual moments. With the exception of chameleons in Fr–En comparisons, none of the relational groups were empty, and even though doves and phantoms constituted the largest groups, no significant presence of a dove or phantom population could be established. The distributional differences with the other systems are also evident in the lack of correlation between SS and DS performance in both En–Fr ($r = -.07,\ p = .5799$) and Fr–En comparisons ($r = .15,\ p = .2585$).

Table 7.7: Number of speakers in each relational group for systems with midpoint of all four spectral moments as input in cross-language comparisons.

Comparison	Dove	Worm	Phantom	Chameleon
En–Fr	5	1	5	1
Fr–En	7	2	6	0

Figure 7.34 further presents the normalised zooplots for the combined systems. In broad agreement with the normalised zooplots for individual spectral moments, particularly with that for CoG (Figure 7.26), speakers primarily varied along the SS dimension between same- and cross-language comparisons, and speakers in the upper half of the plot typically moved in the opposite direction to those in the lower half. Speakers situated in the upper half, who performed relatively well in DS comparisons, mostly moved towards the left. Most dove speakers in same-language comparisons in fact exhibited

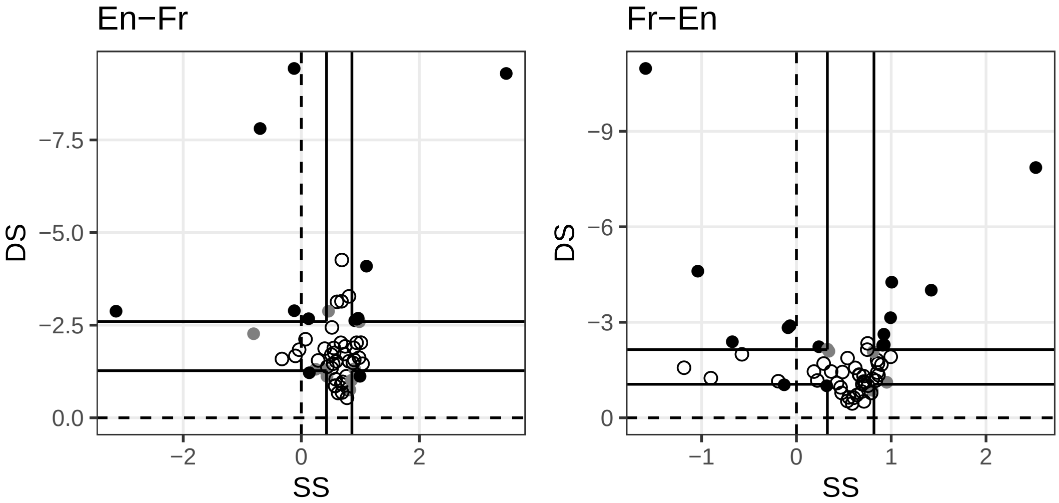

Figure 7.33: Zooplots for systems with midpoint of all four spectral moments as input in cross-language En–Fr (left) and Fr–En (right) comparisons.

little movement, performing exceptionally well also in cross-language comparisons, although SS performance of a number of speakers were particularly affected by language mismatch, such that they were classified as phantoms in cross-language comparisons. Speakers located in the lower half of the normalised zooplots, with relatively poor performance in DS comparisons, chiefly shifted towards the right. The degree of movement was relatively small when the KS was in English (right panel), meaning that their relative SS performance did not markedly improve. With French KS (left panel), however, the worst-performing worms in same-language comparisons were among speakers who shifted to the right most substantially.

7.2.3 Acoustic analysis

The section above revealed that, compared to same-language comparisons using acoustic parameters of /s/, the consequences of language mismatch on individual performance consistently included the proliferation of phantom speakers and the emergence of chameleons, coinciding with a reduction in the number of doves and worms. This section goes on to examine the acoustic characteristics of speakers identified as members of relational groups in En–Fr and Fr–En cross-language comparisons to explore the relationship between individual performance and underlying acoustic data.

Figures 7.35 to 7.38 present a cross-linguistic illustration of the four spectral moments to analyse systems using midpoints of individual spectral moments as input. Following the previous section, the analysis here drew on Condition C1 only, in which the

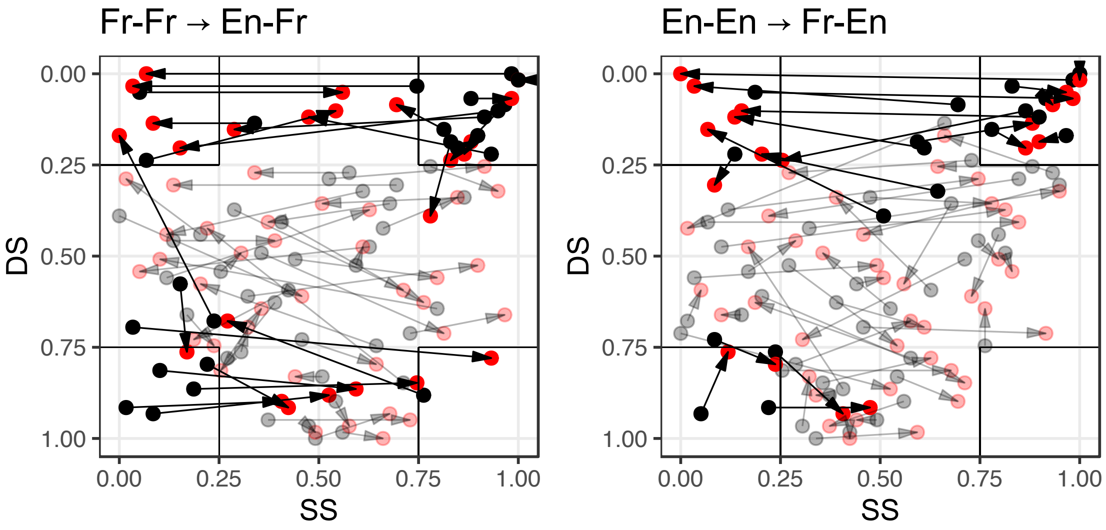

Figure 7.34: Normalised zooplots for systems with midpoint of all four spectral moments as input in same-language (black) and cross-language (red) comparisons.

language of the background data was matched with the KS language. The top panel of each Figure identifies speakers classified as members of relational groups in En–Fr comparisons, where English QS was compared with French KS, and plots the difference between English and French by speaker. Each speaker is represented by two circles, one for each language, the size of which represents the within-speaker standard deviation of the spectral moment in question. Speakers are arranged in ascending order of their means in the KS language (i.e., French). To visualise the cross-linguistic differences more clearly, speaker means are centred on the French means (see Figures 7.11–7.14 above for unadjusted values). The bottom panels illustrate Fr–En comparisons in an analogous manner. For example, in the top panel in Figure 7.35, speaker 385 was identified as a phantom in En–Fr comparisons and, being located on the far right, had the highest mean CoG in French among all 60 speakers. The respective sizes and locations of the circles show that his mean CoG in English was over 500 Hz lower than that in French, and that he had very low within-speaker variability of CoG in French but somewhat higher within-speaker variability in English.

Figure 7.35 shows that, as in the case of same-language comparisons, speakers identified as doves generally could be found among speakers with the highest or lowest values. Furthermore, compared with phantom speakers, who were also typically found on the margins of the population, especially for En–Fr comparisons, doves showed generally small differences between English and French. There were nonetheless speakers who exhibited acoustic characteristics similar to those of doves but did not perform well enough

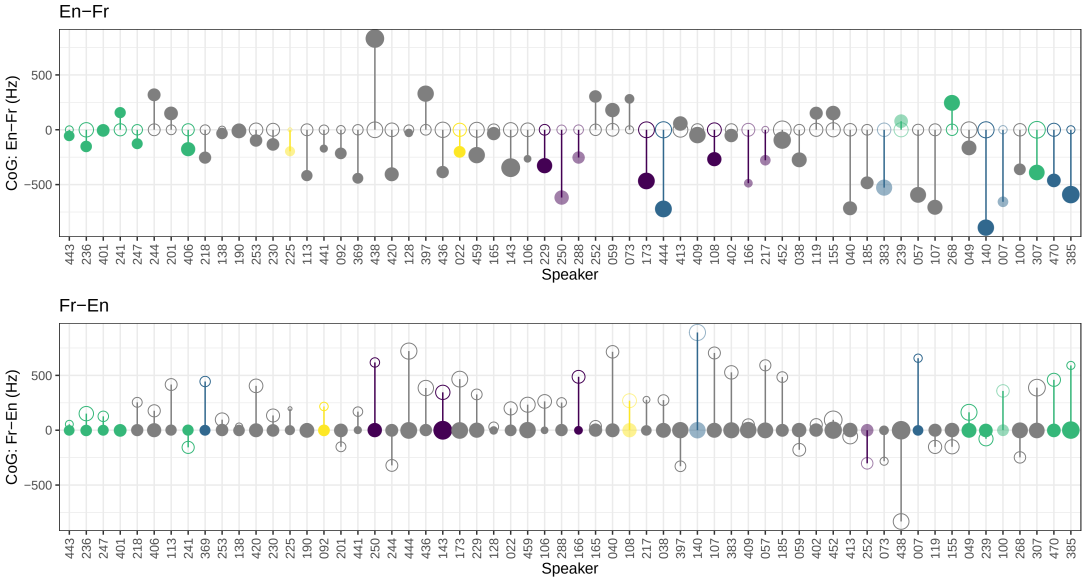

Figure 7.35: Dumbbell plot showing cross-linguistic difference in CoG by speaker (solid: English; hollow: French). Top panel highlights members of relational groups in En–Fr comparisons of systems using midpoint CoG as input (green: doves; purple: worms; blue: phantoms; yellow: chameleons; dark: full members; light: near members), with CoG centred on French means and speakers arranged in ascending order of French means. Bottom panel analogously illustrates Fr–En comparisons, with CoG centred on English means and speakers arranged in ascending order of English means. Size of each circle is proportional to individual within-language standard deviation (scaled separately in each panel).

to be classified as doves (e.g., 049 in En–Fr comparisons, 268 in Fr–En comparisons), which may be attributed to the direction of the difference between English and French. As these speakers produced CoG that was higher than average in the language of the KS and the background data, and their CoG in the QS language was relatively lower, their QS would more closely resemble the norm of the reference population and as such lead to relatively poorer performance. This can be contrasted with the doves nearby, who mostly had either CoG in their QS even further away from the rest of the speakers or especially outlying CoG values (cf. Figure 7.11). By contrast, even though phantoms were often found near other doves, they typically showed much larger distance across the two languages, such that the two samples did not show a high degree of similarity despite their distinctiveness.

In both En–Fr and Fr–En comparisons, worms were commonly found within the central portion of the population. They also commonly displayed a marked difference between English and French, such that both similarity between the QS and the KS and distinctiveness of their CoG values were low. For instance, speaker 250, who was classified as a near-worm in En–Fr comparisons and a worm in Fr–En comparisons, had CoG distributions that were over 500 Hz apart. The few (near-)chameleons were similarly found near the central portion of the population. None of the chameleons found produced CoG that was highly distant across the two languages, but there were no clear characteristics shared by these speakers that distinguished them from other speakers who did not receive the same classification. Indeed, the zooplots in Figure 7.25 show a large cluster of speakers with very similar levels of performance near the bottom right corner, suggesting that similarities in the speakers' acoustic behaviour and the general acoustic proximity between speakers are also reflected in their individual discriminatory performance.

As Figure 7.36 shows, speakers identified as members of relational groups in cross-language comparisons using midpoint SD broadly shared the acoustic characteristics demonstrated by those in the case of CoG. There are, however, a number of speakers who appeared to not follow the same set of patterns, which warrants further discussion. Speaker 441, for example, was a near-phantom in En–Fr comparisons, but did not exhibit great difference in SD between English and French. The reason for his relatively poor SS performance in En–Fr comparisons may alternatively be attributed to the relatively large within-speaker variability in both languages, coupled with the fact that the SD in his English QS was higher than that in French. As this speaker's SD was on the low end within the population, a higher SD in English would then be closer to the population norm, thus lowering his ability to be matched with himself. This can be contrasted with Speaker 239, who was classified as a near-dove in En–Fr comparisons and similarly had a slightly higher SD in English than in French. As this speaker had one of the highest values of SD in French, having an even higher SD in English would not have the effect of making the QS more similar to the majority of other speakers and so would not result in particularly poor performance in SS comparisons. In fact, speaker 441's situation in En–Fr comparisons can be contrasted with his own classification as a near-dove in Fr–En comparisons. Given that his SD in English was among the lowest in the population, his SD in French that was further lower would be even more dissimilar to the population norm. Also noteworthy are the particular cases of speakers 190 and 059, who were both classified as near-worms in Fr–En comparisons but had very similar SD means in English and French. These speakers, however, did exhibit relatively very high within-speaker variability in both languages, which likely contributed to their relatively poor performance in SS comparisons when considered in conjunction with their highly typical SD values. At the same time, it is noted from the zooplot in Figure 7.27 that the performance of near-worms was in fact not particularly poor in absolute terms. Unlike the phantoms in this set of comparisons, their SS-LLR was not negative but was slightly above 0 and fairly similar to the cluster of other speakers in the bottom half of the zooplot. Their classification as near-worms may thus be accounted for in more relative terms, as their high intra-speaker variability has led them to be unfavourably compared to other speakers who may have exhibited a greater difference across languages.

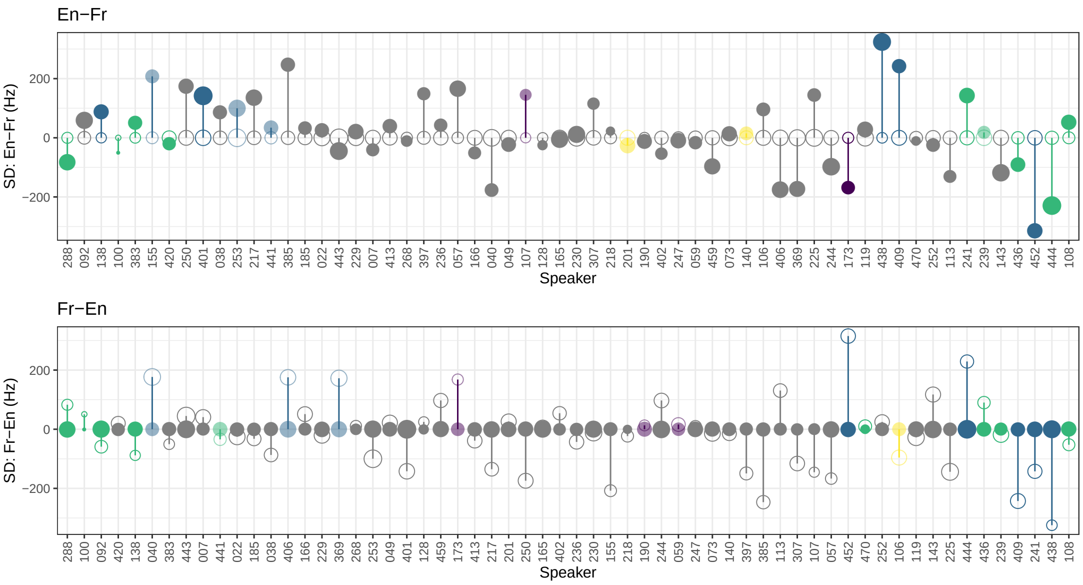

Figure 7.36: Dumbbell plot showing cross-linguistic difference in SD by speaker (see Figure 7.35 for explanation).

In the case of skewness, doves and phantoms were mostly concentrated among speakers with the most extreme values. There are, however, notable exceptions of speakers far from the margins of the population who were nonetheless identified as doves (e.g., 268 in En–Fr; 049 in Fr–En) or phantoms (e.g., 444 in En–Fr). A close inspection of the zooplots in Figure 7.29 suggests that these exceptions may be difficult to account for following the adopted cut-offs for each relational group. While a number of doves and phantoms showed more extreme individual performance that set them apart, many of the remaining speakers, including some doves and phantoms close to the 25% cut-off points, produced mean SS-LLR and DS-LLR within a very narrow range and so performed at almost indistinguishable levels. The range of acoustic variation in skewness between speakers is also relatively narrow (see Figure 7.13), in comparison with CoG or SD (Figures 7.11–7.12), apart from the very few speakers with a clearly lower or higher skewness distribution. The acoustic distance between speakers identified as doves or phantoms who were placed near the centre of the group and those nearer either end was therefore relatively small.

Moreover, the distinction between doves and phantoms is arguably weaker than that observed for CoG and SD, with some doves exhibiting similarly large or even larger differences between English and French than phantoms. Examples of such cases include speakers 307 and 470, who were respectively identified as a dove and a phantom in En–Fr comparisons but belonged to the opposite groups in Fr–En comparisons. The overall acoustic similarity between the two speakers (see Figure 7.13) suggests that the relative performance of an individual may be sensitive to subtle changes in the skewness distributions, not only within these individuals, but also among the sample of background speakers randomly selected to model the reference population.

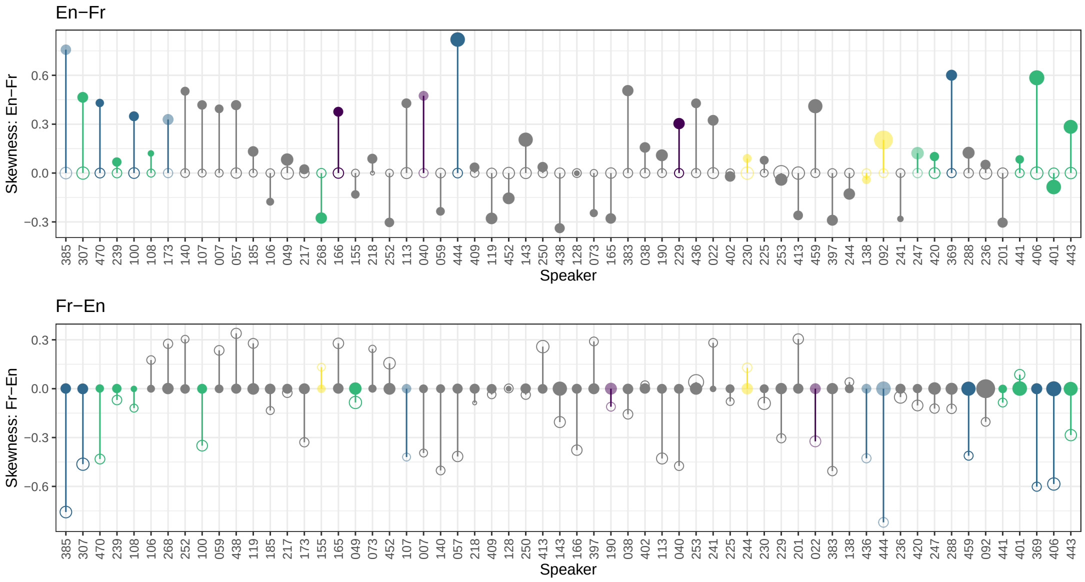

Figure 7.37: Dumbbell plot showing cross-linguistic difference in skewness by speaker (see Figure 7.35 for explanation).

The final spectral moment, (log-)kurtosis, is shown in Figure 7.38. In both En–Fr and Fr–En comparisons, doves can be found on both low and high ends of the population and all exhibited relatively small differences between English and French. Notably, however, the same characteristics were also shared by some other speakers who did not perform as well as the doves (e.g., 252). Phantoms, on the other hand, were not limited to speakers in certain parts of the population, but included both speakers on the margins and those in the centre (e.g., 444 in Fr–En) who sat alongside chameleons and worms. Phantoms in both En–Fr and Fr–En comparisons all produced /s/ with relatively distinct kurtosis in English and French, with the exception of Speaker 401 in En–Fr comparisons, whose distribution of log-kurtosis showed relatively high within-speaker variability in each language but only little cross-linguistic difference (in Fr–En comparisons, he was classified as a dove). In the systems analysed, chameleons, like doves, showed very small differences between English and French, which may account for their relatively strong performance in SS comparisons, but did not produce /s/ with kurtosis as extreme as doves did, which may in turn account for their relatively weak performance in DS comparisons. As in the case of skewness, chameleons were not in fact well distinguished from most other speakers who were not similarly categorised but performed at generally the same level. The acoustic characteristics that chameleons displayed were thus unsurprisingly shared by many such speakers.

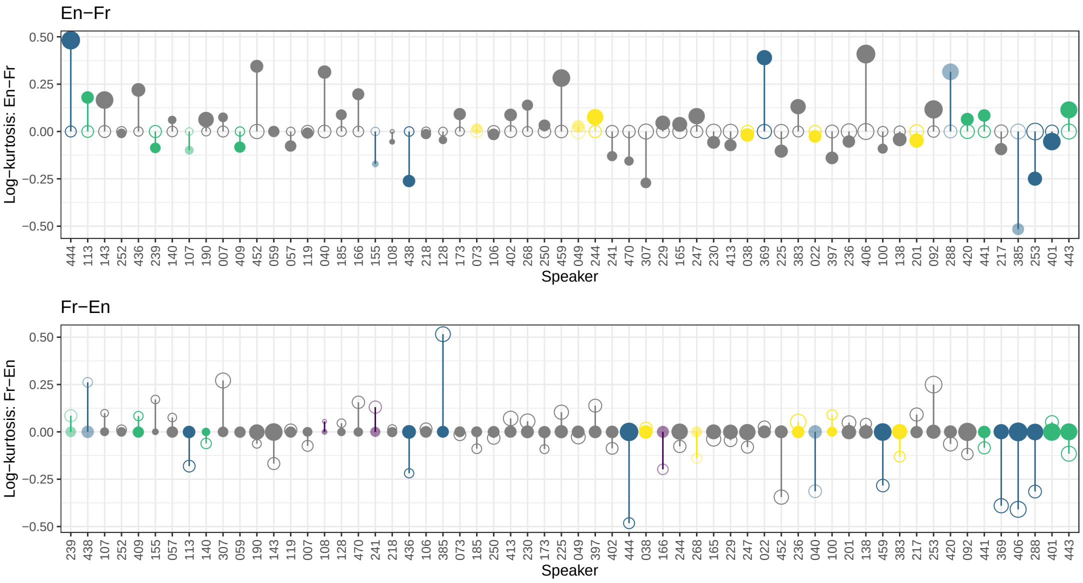

Figure 7.38: Dumbbell plot showing cross-linguistic difference in log-kurtosis by speaker (see Figure 7.35 for explanation).

7.2.4 Discussion: Cross-language comparisons

Overall, the results from Experiment 2 show that, in cases of cross-language comparison, /s/ spectral moments are only of limited value, with their discriminatory power much reduced when compared to same-language comparisons. Global-level analysis in Section 7.2.1 found that C_{llr} and EER both increased to near 1 for individual spectral moments, while combining all four spectral moments only resulted in limited improvement over using individual moments. While dynamic systems still maintained a small advantage over static systems for individual moments, that advantage was completely offset when spectral moments were combined, suggesting that the added value of dynamic representation is restricted when there is language mismatch. Notably, there was virtually no difference between systems which matched the language of the background data with the KS language and those matched with the QS language.

Individual analysis focused on midpoint-only systems in Section 7.2.2 provides more information as to the sources of degradation of performance. In SS comparisons, speakers showed a general shift towards more negative LLRs, with a greater proportion of speakers producing negative mean SS-LLR, thus contributing to high error rates. As for DS comparisons, although speakers with mean DS-LLR > 0 remained a rare occurrence, the clustering of speakers close to DS-LLR $= 0$ provides evidence of a general weakening in this aspect. Even though the strength of evidence in DS comparisons was still reported to be strong for a small number of outliers, it was nonetheless much reduced from the extreme values obtained from same-language comparisons. Further, zooplots showed a concentration of speakers near the lower right corner of the zooplot, as opposed to the lower left corner of the zooplot, indicating that there was in fact little difference, in terms of SS comparisons, between speakers who reported the strongest performance in that type of comparison and the majority of the population. When compared with the systems using individual moments, individual performance in the combined systems improved only in small increments, especially in SS comparisons, where the strength of evidence remained weak for the vast majority of speakers.

Findings from analysis of normalised zooplots, in conjunction with the acoustic analysis in Section 7.2.3, illustrate the uneven impact of language mismatch on relative individual performance, depending on the acoustic characteristics displayed by the speaker. Many speakers classified as doves in same-language comparisons were able to maintain their strong performance, albeit much less extreme, such that doves remained the largest relational group. These speakers generally produced highly distinctive spectral moments in the KS language and were relatively stable across languages, showing little cross-linguistic differences.

Not all doves were able to maintain their strong performance, and those who did not were typically severely affected in SS comparisons, but not so much in DS comparisons, resulting in a sizeable phantom group. Such movement was reflective of the overall trend of language mismatch having a more variable impact on SS comparisons, where speakers with stronger DS performance were relatively more affected, while very few speakers exhibited much vertical movement along the DS dimension (except in the case of kurtosis). Phantoms typically displayed large within-speaker cross-linguistic variation, such that their performance in SS comparisons was particularly weak. While their distributions in the KS language were generally distinctive (typically doves in same-language comparisons), this was not necessarily the case, particularly for skewness and kurtosis. Indeed, when the distributions for these two spectral moments (in Figures 7.13–7.14) were examined closely, the range of variation was low except for a few speakers on the margins. As such, where phantoms, with large cross-linguistic differences, had a distribution in the QS that was highly dissimilar from the norm of the reference population, the resultant LRs from these speakers would be more biased towards support for the different-speaker hypothesis.

Speakers towards the bottom of the normalised zooplots also exhibited much horizontal movement, such that speakers who performed relatively poorly and were classified as worms in same-language comparisons performed relatively better in cross-language comparisons. In same-language comparisons, worms were typically speakers who had highly uncharacteristic distributions of spectral moments. Here, in cross-language comparisons, both worms and chameleons were generally indistinctive in the KS language. The difference in their SS performance, then, appeared to be largely determined by their within-speaker variability, either within the same language or across languages. Speakers who showed a strong difference between the KS and the QS, or high within-speaker variability within each language, ended up performing more poorly and being classified as worms. Chameleons, on the other hand, consisted of speakers with typical values of spectral moments who displayed relatively low within-speaker variability. As noted above, however, the chameleons in these sets of cross-language comparisons were often not a well-defined group. The clustering of speakers near the lower right corner meant that the performance of chameleons was in fact similar to many other speakers in the population, which was also reflected in the similar individual distributions of spectral moments across English and French.

The findings here demonstrate that, even where speakers as a group do not produce cross-linguistically distinct /s/, the discriminatory power of the feature can be sensitive to subtle acoustic shifts on an individual level. As speakers show considerable variability in their response to language shift, their performance in both SS and DS comparisons deteriorates in cross-language comparisons. How they perform individually, relative to the population, is in turn tightly tied up with how their cross-linguistic patterns are positioned in relation to the group. The strong adverse effect of training mismatch, arising from the conversion of scores in cases of language mismatch using scores trained without language mismatch, further underscores the sensitivity of LR output to individual differences among bilingual speakers and to caution against cross-language FVC analysis when only training/reference data in a single language are available.

Chapter 8

Results: LTFDs

This chapter presents results from Experiments 1 (Section 8.1) and 2 (Section 8.2) for the semi-automatic linguistic-phonetic variable of LTFDs, following the same structure as Chapter 7. Each section first outlines global-level results on C_{llr} and EER, followed by an interim discussion. An individual-level analysis is then presented along with an acoustic analysis of outlying speakers in each system, before the findings from each experiment are discussed.

8.1 Experiment 1: Same-language comparisons

8.1.1 Global metrics

C_{llr} and EER for all tested systems using LTFDs, grouped by the number of LTFs included as input in the system, are presented in Figures 8.1 and 8.2. Metrics for each individual LTF are additionally presented in Figures 8.3 and 8.4.

Overall, all systems using parameters from individual LTFs produced mean C_{llr} below 1. In fact, C_{llr} did not exceed 1 in any replication for any of the tested systems, suggesting that these systems were relatively well calibrated and could offer some useful speaker-specific information. As Figures 8.3 shows, it is clear that the inclusion of formant bandwidths consistently led to lower C_{llr} and EER than F-only systems. Across both languages, F-only systems of individual LTFs produced mean C_{llr} between 0.61 and 0.74 and EER between 18.8% and 27.2%, while F+BW systems yielded mean C_{llr} between 0.40 and 0.51 and EER between 10.8% and 14.6%. The quantitative advantage of F+BW systems over F-only systems, however, varied across languages and LTFs, as shown by a significant three-way interaction of language, the LTF used and the mode of input for both C_{llr} ($\chi^2(3) = 42.85,\ p < .0001$) and EER ($\chi^2(3) = 63.35,\ p < .0001$).

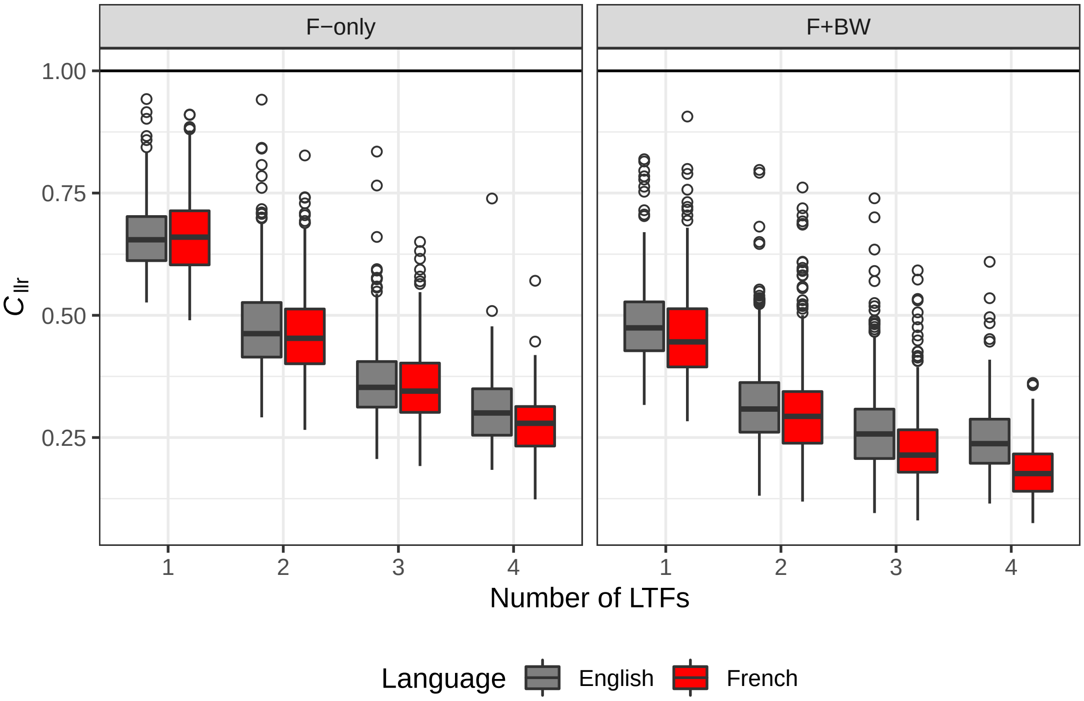

Figure 8.1: C_{llr} from system testing of all F-only and F+BW systems in English (grey) and French (red), grouped by number of LTFs included. Horizontal rule reference at $C_{\text{llr}} = 1$.

Within F-only systems, it can be seen from Figures 8.3 and 8.4 that LTF1 generally performed the best, while LTF2 consistently produced the highest C_{llr} and EER overall. Although the relative ranking of each LTF appeared to be stable across languages, the difference between LTFs showed a tendency to be greater in French. Post-hoc Tukey-adjusted pairwise comparisons by mode of input, performed using the *emmeans* package (Lenth, 2020), confirmed these observations. Within the French systems, EER and C_{llr} from LTF1 and LTF4 did not significantly differ from each other, but they were significantly lower than the metric values from LTF3, which were in turn lower than those from LTF2. In English, LTF1 produced significantly lower C_{llr} than the other LTFs, but C_{llr} from LTF2–4 was not significantly different from one another. LTF4 produced the lowest EER that was significantly lower than the others, and while EER from LTF3 was not significantly different than EER from either LTF1 or LTF2, EER from LTF1 was significantly lower than that from LTF2.

As for F+BW systems, no consistent trends could be easily discerned between each LTF, though notably LTF2, which performed worst in F-only systems on the whole, produced the lowest C_{llr} and EER here. Post-hoc comparisons showed that, in French, LTFs could be grouped in pairs in terms of their performance, with LTF2 and LTF4 performing better in both metrics than LTF1 and LTF3. LTF2 further produced significantly lower

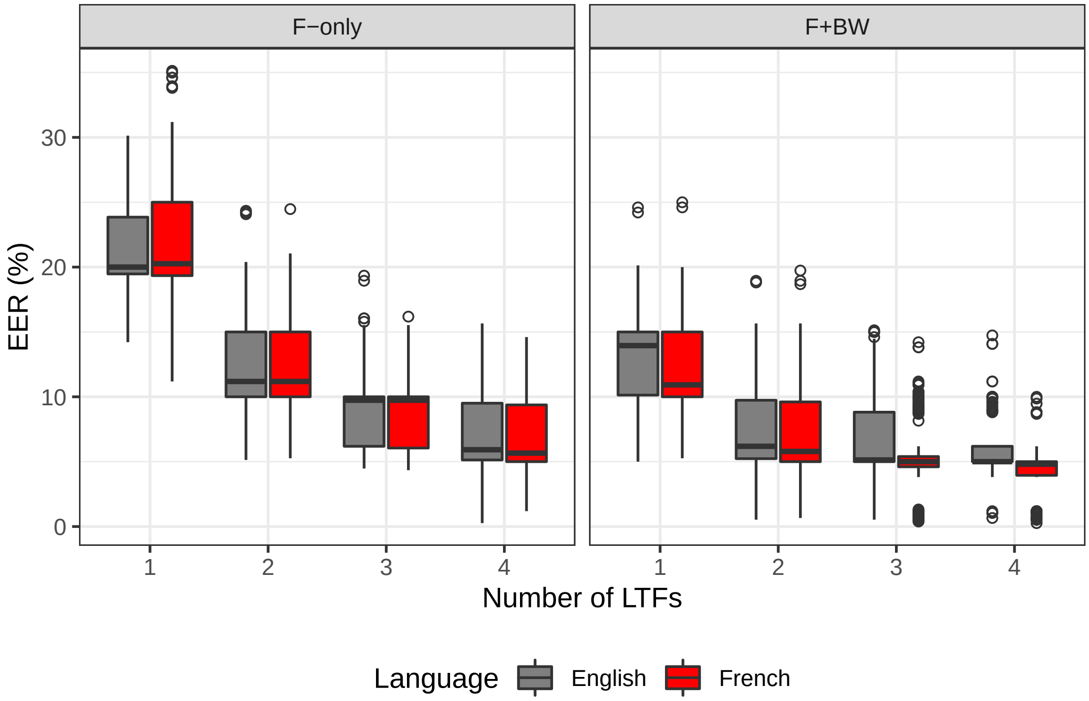

Figure 8.2: EER from system testing of all F-only and F+BW systems in English (grey) and French (red), grouped by number of LTFs included.

C_{llr} than LTF4, but performance within each pair otherwise showed no significant differences. In English, LTF2 stood out as the best-performing LTF with significantly lower EER and C_{llr} than the other LTFs. While LTF3 and LTF4 did not significantly differ from each other in either C_{llr} or EER, they performed significantly better than LTF1, but only in terms of EER.

Figures 8.3 and 8.4 further show that systems using individual LTFs performed generally similarly in both languages, with no clear advantage of one language over the other to be discerned. Post-hoc comparisons carried out above showed that, within F-only systems, only LTF2 demonstrated cross-linguistic differences in both C_{llr} and EER, with stronger performance produced in English. LTF1 produced lower metric values in French than in English, but only the differences in EER reached significance. LTF3 and LTF4, on the other hand, did not produce significantly different EER or C_{llr} in the two languages. In F+BW systems, C_{llr} and EER demonstrated patterns in opposite directions. C_{llr} was significantly lower in French for LTF2 and LTF4, but not for LTF1 and LTF3, whereas EER was significantly higher in French for LTF1 and LTF4, but not for LTF2 and LTF3.

The overall similarity between the two languages is partially substantiated when looking at systems incorporating multiple LTFs, the results of which are summarised in Table 8.1. Along with Figures 8.1 and 8.2, Table 8.1 shows that, as the number of LTFs

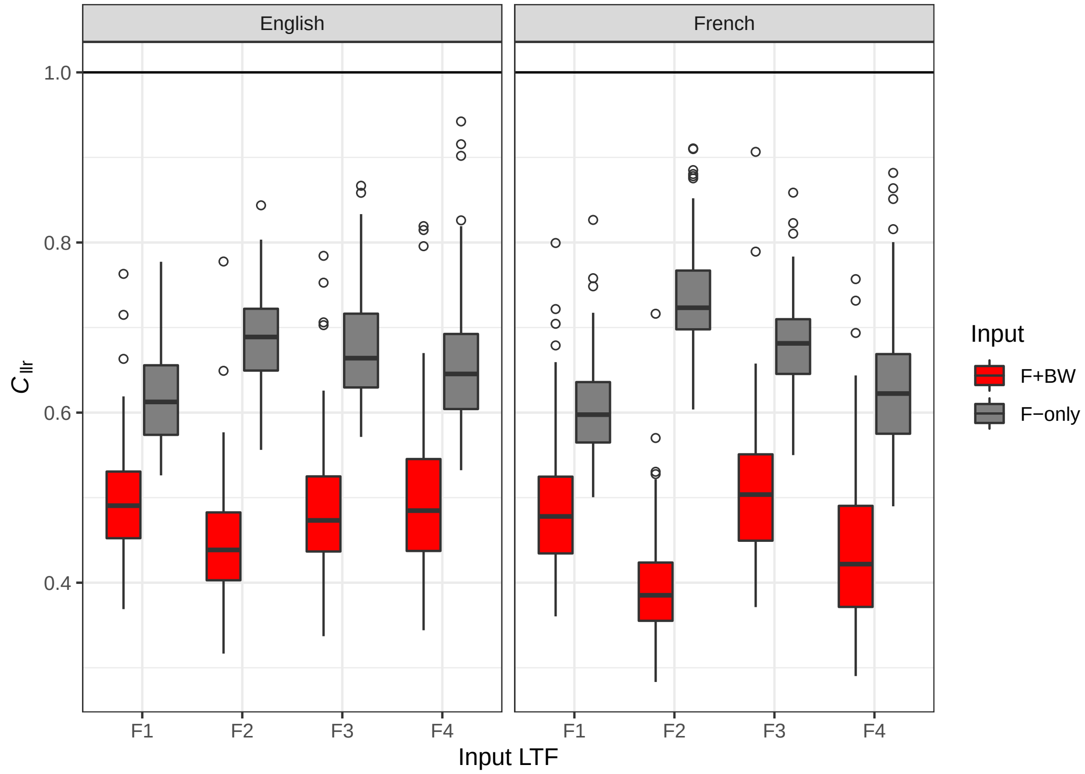

Figure 8.3: C_{llr} from system testing of all F-only (grey) and F+BW systems (red) using individual LTFs in English and French. Horizontal rule reference at $C_{llr} = 1$.

in the system increased, C_{llr} and EER in English and French both consistently decreased. Nevertheless, LMEM analysis did find a significant three-way interaction of number of LTFs included, language and the mode of input for both C_{llr} ($\chi^2(3) = 11.01,\ p = .0117$) and EER ($\chi^2(3) = 15.70,\ p = .0013$). Overall, C_{llr} and EER decreased at higher rates for F-only systems than F+BW systems, such that the quantitative advantage of including formant bandwidths became less substantial as more LTFs were included. Meanwhile, C_{llr} and EER decreased at slightly higher rates in French than in English, but only in F+BW systems, such that there was a growing discrepancy in performance between the two languages as the number of LTFs increased.

Post-hoc Tukey-adjusted comparisons within each mode of input showed that the incremental improvements of C_{llr} and EER when an additional LTF was included the system were significant, with the exception of the changes between three and four LTFs in English. Cross-linguistically, in F-only systems, the difference between English and French was only significant for EER when one formant was used, but not when multiple formants were included, and was not significant in any case for C_{llr}. In F+BW systems, differences in C_{llr} and EER between English and French were all found to be significant,

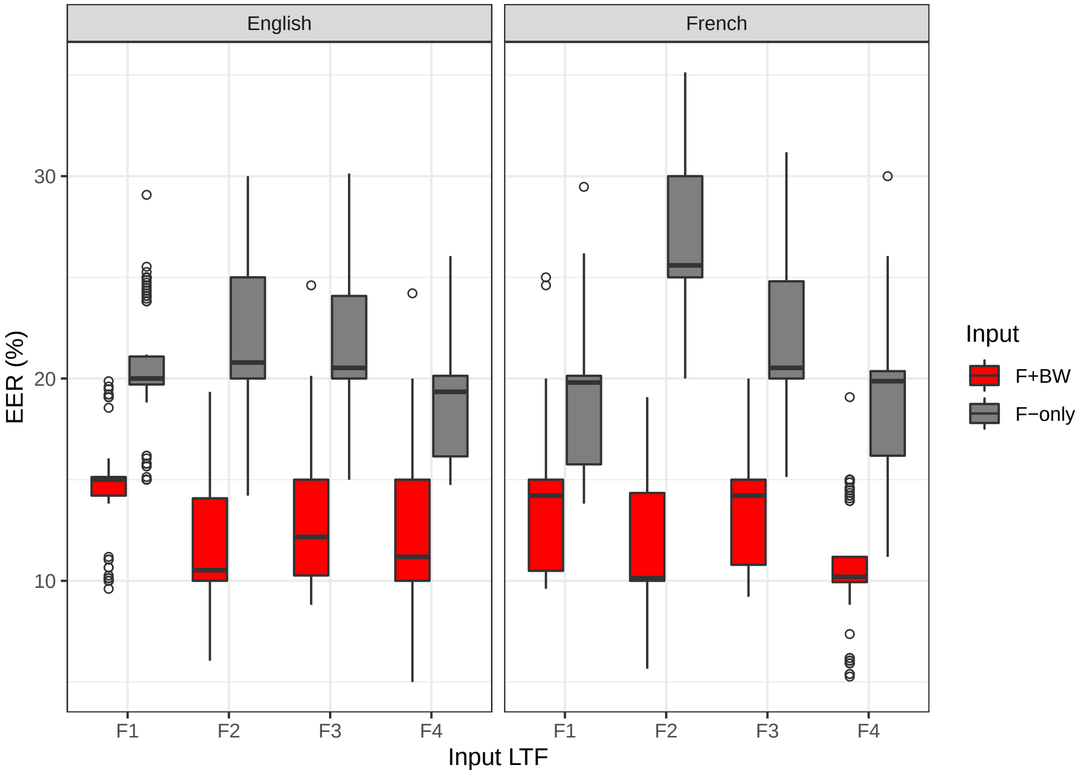

Figure 8.4: EER from system testing of all F-only (grey) and F+BW systems (red) using individual LTFs in English and French.

except for EER when two LTFs were included. In other words, the current findings show that English and French exhibited little difference in F-only systems, regardless of the number of LTFs included, but in F+BW systems French outperformed English, slightly but consistently, especially when the performance of English levelled off when three or more LTFs were included.

Table 8.1: Mean C_{llr} and EER from F-only and F+BW systems in English and French by number of LTFs included.

		F-only		F+BW	
	No. of LTFs	C_{llr}	EER	C_{llr}	EER
English	1	0.66	20.7%	0.48	12.9%
	2	0.48	12.7%	0.32	7.6%
	3	0.37	9.3%	0.27	6.3%
	4	0.31	7.3%	0.25	6.2%
French	1	0.66	21.9%	0.46	12.3%
	2	0.46	12.8%	0.31	7.2%
	3	0.35	9.0%	0.23	5.3%
	4	0.28	7.0%	0.19	4.3%

8.1.2 Interim discussion

The key issue in the present experiment is the cross-linguistic stability of LTFDs as speaker discriminants in FVC. The findings above did not show straightforward, clear-cut patterns, although there was an overall trend of LTFDs in French performing better, or at least on the same level as the same parameters in English. The differences in performance between English and French remained generally small, but reached significance in a number of systems, depending on whether formant bandwidths were included and on the specific LTFs in question. Overall, LTFDs in French and in English were comparable in terms of performance in F-only systems, but French outperformed English slightly when formant bandwidths were included. These findings indicate that, despite the cross-linguistic differences described in Chapter 5, the factor of language only had a minor effect on the discriminatory power of LTFDs. The differences between French and English in F+BW systems nevertheless highlighted the language-specific discriminatory value that formant bandwidths could potentially add to FVC, which may be in part due to the presence of nasal vowels in the French vowel inventory. As nasal vowels are associated with higher formant bandwidths as a consequence of the coupling of nasal and oral vocal tracts, the degree of nasality in these vowels may act as a systematic source of speaker individuality in French that lends itself to making higher contributions to the discriminatory potential of formant bandwidths, particularly in a heterogeneous bilingual population that includes speakers who acquired French—and the contrastive use of nasality—as an L2.

Out of all the factors considered, the strongest influences on C_{llr} and EER clearly came from the use of multiple LTFs and the inclusion of formant bandwidths. The result that combining multiple LTFs led to better performance replicates findings from many previous studies on the discriminatory potentials of LTFDs and, more generally, vowel formants (Becker et al., 2008; Gold et al., 2013b; Hughes et al., 2016; McDougall, 2004), providing further support for the idea that each formant carries complementary speaker-specific information that can be effectively combined to discriminate between speakers. The positive impact of the latter, however, runs contrary to the prediction made in Section 5.3 that adding formant bandwidths would not be able to substantially improve system performance, potentially suggesting that even though bandwidths are relatively limited in between-speaker variability, they could nonetheless capture some speaker-specific information complementary to what is present in the corresponding formants. As such, their inclusion remains to have a conducive, rather than adverse, effect on the discriminatory power of LTFDs. Alternatively, it can be suggested that the effect of including formant bandwidths is confounded by the effect of including more parameters to be modelled in the system. F+BW systems including one LTF use two acoustic parameters, those including two LTFs use four, and so on. This explanation, which would imply that formant bandwidths are no less valuable than formant centre frequencies as speaker discriminants, is supported by the fact that, in terms of both C_{llr} and EER, F+BW systems including one LTF performed at virtually the same level as F-only systems including two LTFs, whereas F+BW systems including two LTFs and F-only systems including all four LTFs also yielded highly similar results. The extent to which improvements from including formant bandwidths can be attributed to their own discriminatory value or simply the inclusion of more data cannot be easily adjudicated on the basis of the current evidence, as the discriminatory potential of formant bandwidths was not tested on its own for comparison, although the finding that F+BW systems displayed language-specificity in their performance discounts an explanation that relies solely on the inclusion of more data. Future research designed specifically to investigate the use of formant bandwidths in FVC would be necessary to address the nature of their contribution.

Other predictions in Section 5.3 on the relationship between different LTFs were similarly not fully borne out. LTF2 did perform the worst out of all individual LTFs, as predicted, but this finding was confined to systems that excluded formant bandwidths. When formant bandwidths were included, LTF2 turned out to gain the most improvement in performance and outperformed all other formants. Contrary to Gold et al. (2013b), higher formants did not offer stronger discriminatory value than lower formants, outperformed by either LTF2, in F+BW systems, or LTF1, in F-only systems. Indeed, earlier studies examining the discriminatory potentials of vowel formants have not yielded consistent findings as to whether higher formants convey a greater amount of speaker-discriminatory information than lower formants (see McDougall, 2004), which McDougall (2004) argues may be dependent on the speech materials and specific conditions. While the current study examines vowel formants in a long-term context rather than within individual segments, the findings here suggest that the discriminatory power of lower formants should not be underestimated and, subject to bandwidth limitations due to telephone transmission (Byrne & Foulkes, 2004), LTF1 may emerge as one of the most speaker-specific LTFs.

8.1.3 Individual-level analysis

This section presents results from zooplot analysis for systems using each individual LTF and the combination of all LTFs. To facilitate acoustic analysis of LTFDs from individual speakers, the current section focuses on the analysis of F-only systems. Zooplots and overall speaker classification derived from F+BW systems are not considered further but are included in Appendix C.

8.1.3.1 LTF1

The zooplots shown in Figure 8.5 illustrate the individual performance of all 60 speakers in the systems based on LTF1. For English, as demonstrated in the zooplot on the left, all speakers produced a negative mean DS-LLR. While mean SS-LLR was positive for most speakers, 12 speakers (20%) produced a negative mean SS-LLR, suggesting that they were not well matched with themselves on average. Performance in SS and DS comparisons was strongly correlated ($r = -.74,\ p < .0001$): Speakers with stronger performance in SS comparisons similarly produced stronger performance in DS comparisons. This is further supported by the absence of any phantoms or chameleons, as only members of the other two relational groups (10 doves and five worms) were identified. Overall, the system produced a narrow range of mean LLRs in English, with SS-LLR between -0.40 and 2.55, and DS-LLR between -0.48 and -1.47. Most speakers could be found in a dense cluster near the lower left corner of the zooplot.

The zooplot in French shows a similar distribution of speakers. As in English, all speakers produced a negative mean DS-LLR, while nine speakers (15%) produced a negative mean SS-LLR, and a strong correlation was found between speakers' performance in SS and DS comparisons ($r = -.82$, $p < .0001$). The range of mean LLRs was likewise narrow, with the exception of a few doves, which produced mean SS-LLR and DS-LLR of up to 4.96 and -2.36 respectively. The size of each relational group closely mirrored their counterpart in English, with a significant absence of any phantom or chameleon speakers. A significant presence of the dove group was also found, consisting of 11 speakers, whereas the group of worms comprised seven speakers.

8.1.3.2 LTF2

As shown in the zooplots in Figure 8.6, all speakers also produced a negative mean DS-LLR in the LTF2 systems, and a majority of speakers produced a positive mean SS-LLR. In English, although there were the same number of speakers with negative mean SS-LLR (12) as in the case of LTF1, the magnitude of these negative SS-LLRs was slightly higher (up to -0.80). In French, the number of speakers with negative mean SS-LLR rose to 14 (23%), with the lowest mean SS-LLR being -0.83. These findings provide an indication of poorer performance in SS comparisons in both languages. Relatively poorer performance in DS comparisons is also observed in French, as the zooplot shows a lack of outlying well-performing speakers alongside a cluster of speakers closer to DS-LLR $= 0$. Performance in SS and DS comparisons was only moderated correlated in English

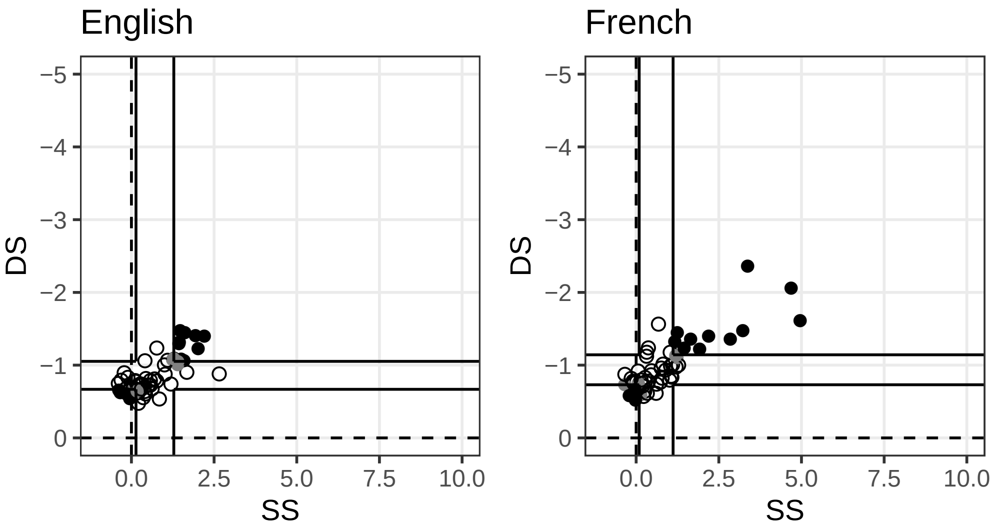

Figure 8.5: Zooplots for systems with LTF1 as input in English (left) and French (right). Abscissa and ordinate respectively show mean SS-LLR and DS-LLR. Solid line segments represent 25th and 75th percentiles; dotted lines indicate mean LLR = 0. Members and near-members of relational groups respectively in black and grey.

($r = -.51,\ p < .0001$) and not significantly correlated at all in French ($r = -.08,\ p = .5198$), with a notable presence of speakers in the upper left and lower right regions of the zooplot. These include three speakers identified as phantoms in English and four speakers similarly identified in French, as well as two chameleons in French. By contrast, only nine doves and three worms were identified in English, and only seven doves and no worms were identified in French, fewer than any other individual LTFs.

8.1.3.3 LTF3

The distribution of speakers in the zooplots for LTF3, as displayed in Figure 8.7, shows broad similarities with those for LTF1. No speakers were classified as phantoms or chameleons in both English and French, and strong positive correlations were similarly found between SS-LLR and DS-LLR (English: $r = -.92,\ p < .0001$; French: $r = -.90,\ p < .0001$). As in the case of LTF1, a dense cluster of speakers could be located near the lower left corner of the zooplot, including eight worm speakers in both English and French. Despite the relatively high number of worms, in English only eight speakers produced a negative mean SS-LLR, none of which exceeded -0.15, indicating smaller contrary-to-fact LLRs on average. In French, the number of speakers with negative mean SS-LLR stands higher at 12, with a mean SS-LLR of up to -0.35. Stronger individual performance is evident among the large group of best-performing doves (12 members and three near-

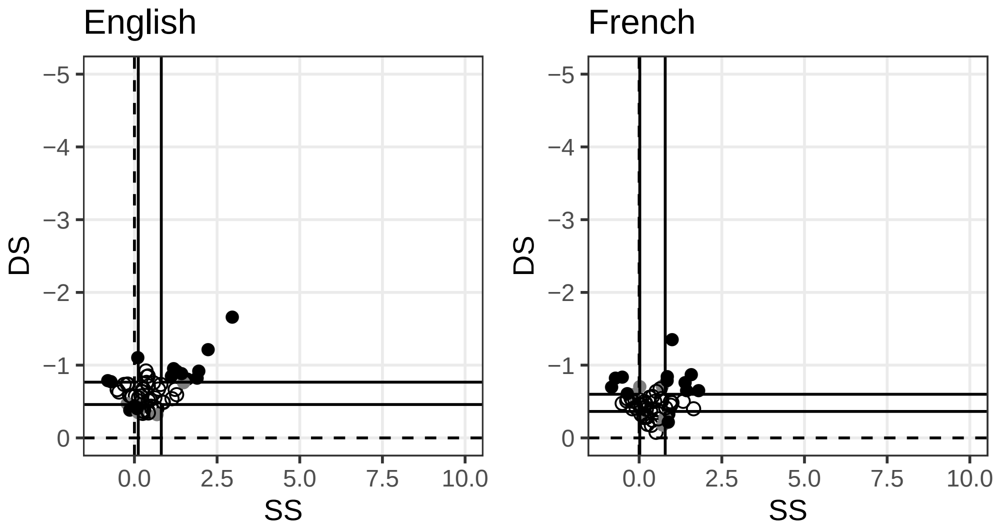

Figure 8.6: Zooplot for systems with LTF2 as input in English (left) and French (right).

members in English; 13 members in French), who were distant from the dense cluster of speakers and spread further upward and rightward in the zooplot, particularly in English, as a result of more positive mean SS-LLR and more negative mean DS-LLR.

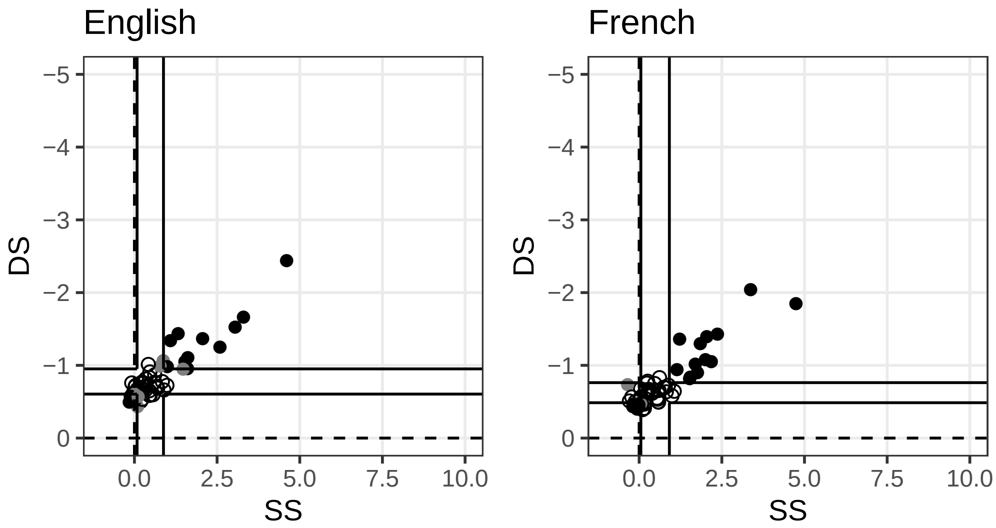

Figure 8.7: Zooplot for systems with LTF3 as input in English (left) and French (right).

8.1.3.4 LTF4

The zooplots for the LTF4 systems, as shown in Figure 8.8, illustrate a distribution of speakers resembling that for LTF3, although speakers were less clustered towards the lower left corner in both languages, indicating an overall greater degree of between-speaker variation in LR performance. Performance between SS and DS comparisons was also strongly correlated in both English ($r = -.87,\ p < .0001$) and French ($r = -.90,\ p < .0001$), with particularly high mean SS-LLR reported for a number of speakers. Mean SS-LLR reached a maximum of 4.78 in English and an even higher 7.10 in French. At the same time, this system reported the highest number of speakers with outlying performance in English, where a total of 11 speakers were classified as doves and 13 were classified as worms. The number of doves in French was similarly high in French at 11, although there was no significant presence of a worm group, which contained only eight speakers. The high proportion of outlying speakers suggests that, in English, individual performance is more extremely distributed in the case of LTF4. A similar case can be made for French, where there was noticeable separation between the cluster of speakers in the lower left corner with relatively poor performance and those in the upper right region.

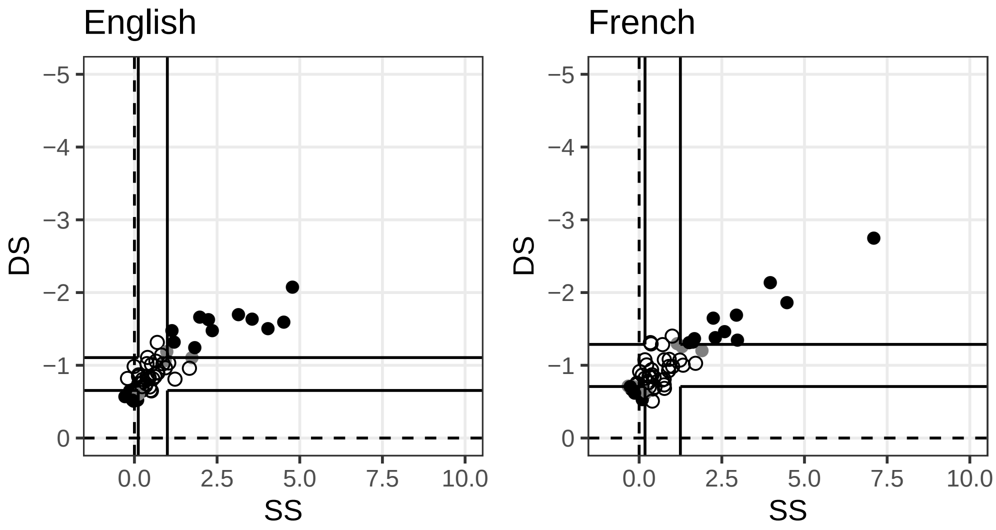

Figure 8.8: Zooplot for systems with LTF4 as input in English (left) and French (right).

8.1.3.5 All LTFs combined

Compared to the zooplots for individual LTFs, there is evidently a wholesale shift of speakers towards the top of the zooplots in Figure 8.9, brought on by more strongly negative DS-LLR. No speaker produced a mean DS-LLR higher than -1.70 in either language, and the most extreme mean DS-LLR was beyond -4.00 in both languages. Speakers also tended to have much higher SS-LLR, when compared with zooplots for the other systems, as evidenced by the rightward spread of speakers in Figure 8.9. The majority of speakers (42 in both cases) produced a mean SS-LLR greater than 1, while only a small number of speakers (three in English; four in French) produced marginally negative mean SS-LLR. In this set of systems, mean SS-LLR and DS-LLR were moderately correlated in English ($r = -.59,\ p < .0001$) and more strongly correlated in French ($r = -.72,\ p < .0001$). No phantoms or chameleons were found, although one speaker was classified as near-phantom in English. Eight doves and six worms were identified in English, while in French the distribution is more extreme, with 11 speakers classified as doves and nine as worms. Overall, the distribution of speakers in Figure 8.9 is clearly much less clustered than in any of the individual LTF systems.

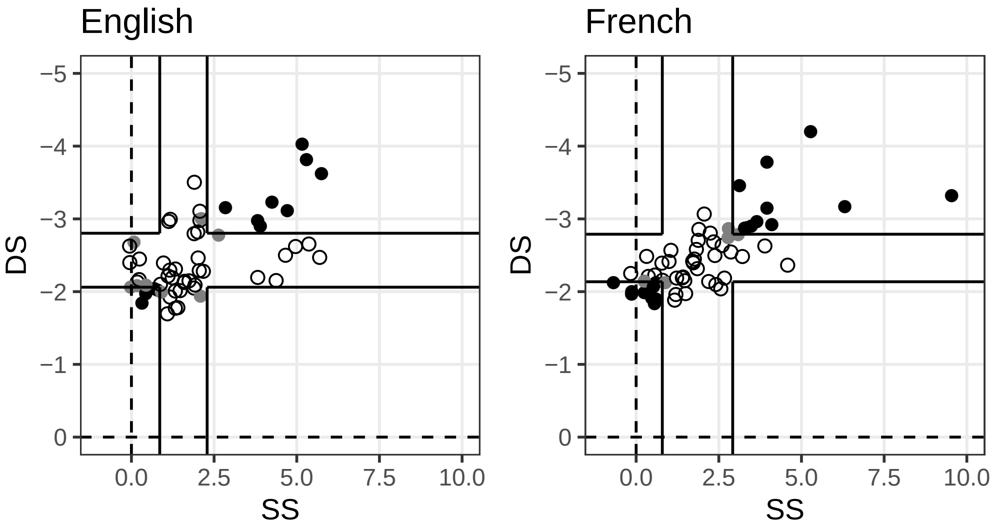

Figure 8.9: Zooplot for systems with all four LTFs as input in English (left) and French (right).

8.1.3.6 Speaker classification

Following separate analyses within each system above, this section considers the performance of individuals across all LTFD systems and across languages. Figure 8.10 summarises the animal group classification of all speakers for each system tested.

To first consider the English systems, while there are some overlaps between different LTFs, it is clear that each LTF generally captured a different group of outlying speakers. Out of 60 speakers, 21 (35%) were in or near a relational group for more than one LTF, but only three speakers (5%) were in or near a relational group for all four LTFs: 441 was consistently classified as a dove; 470 was similarly always in or near the dove group; 119 was classified as a worm for all LTFs except for LTF2, where he was classified as a

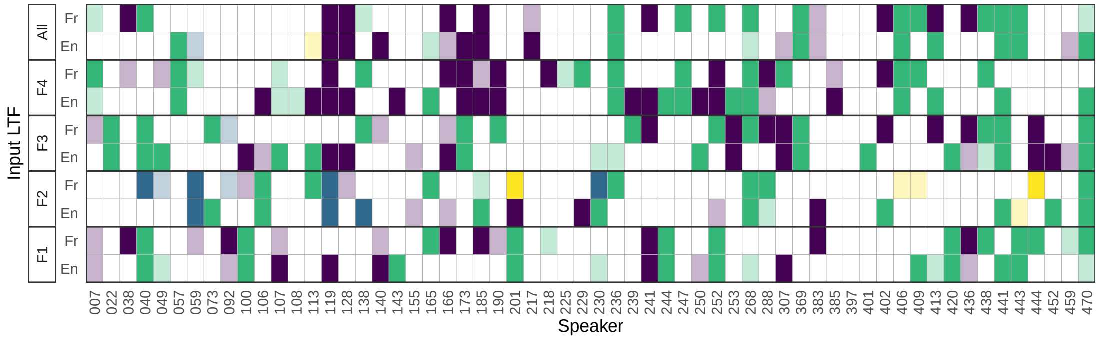

Figure 8.10: Summary of speakers classified as doves (green), worms (purple), phantoms (blue) or chameleons (yellow) in systems tested (near-members of each group in lighter shade).

phantom. The difference between systems is further illustrated by the finding that 12 (20%) speakers were classified as the best-performing doves for one LTF but as the worst-performing worms for another. Across all four LTFs, only five speakers (8%) were not in or near any group, meaning that the vast majority of speakers could be considered an outlying speaker for at least one LTF.

In the combined system for English, all but two speakers who were in or near an animal group were similarly classified for at least one of the individual LTFs. Out of those 22 speakers, 12 of them shared their classification with the LTF4 system. While systems using other LTFs shared fewer (near-)members of animal groups with the combined system, with LTF2 having only five speakers in common, each of the individual LTFs could uniquely account for at least one classified animal in the combined system. The only exceptions were 217, who was a worm in the combined system but not a member of the outlying groups for any individual LTF, and 113, who was a near-chameleon in the combined system. The latter could be explained by the mixed contribution of his relatively poor DS performance in LTF4, for which he was classified as a worm, and relatively excellent SS performance in LTF3, which resulted in a dove classification.

A similar picture emerged for the systems in French. Half of all speakers were in or near a relational group for more than one LTF, 13 (22%) were in or near a group for three LTFs, but none was in or near a relational group for all four LTFs. Most speakers could be considered an outlying speakers for at least one LTF, with only eight speakers (13%) not in or near any group across all four LTFs, but classification consistency across LTFs is weak. Out of the 30 speakers who found (near-)membership in any relational group more than once, only 12 were always in or near the same group. With one exception, the other speakers were all well-performing (near-)doves for at least one LTF, but received a different classification in other cases.

In the combined system for French, all but one speaker who were in or near an animal group shared the same classification with at least one individual LTF. Speaker 217, who was discussed above as a worm in the English combined system unaccounted for by any individual LTF, emerged again as a near-worm in the French combined system here but was otherwise not a member of any outlying groups. Of the other 24 speakers identified as outlying in the combined system, systems using LTF3 and LTF4 each shared 13 of them. As in the case of English, LTF2 had the fewest classifications (only four) in common with the combined system, but the contribution of each LTF is reaffirmed by the fact that each individual LTF could account for at least one classification in the combined system that was not shared by the other LTFs.

Overall, the analysis thus far provides evidence on an individual level that complementary speaker-specific information is available from LTFDs of different formants. As the findings above illustrate, speaker-specific information from each LTFD contributes, albeit in varying degrees, to the classification in the combined systems.

When speaker classification in English is compared with that in French, some degree of correspondence can be found. Table 8.2 summarises the number of speakers who were in or near the same relational group in both languages, in comparison with the total size of group membership in each language. In all cases, speakers with the same classification in both languages amounted to roughly half of the total size of the groups. Moreover, with the exception of two phantoms, all common pairs of members were either doves or worms, which is not unexpected given the low number of chameleons and phantoms across all systems, though the relatively high proportion of phantoms that were shared is notable.

For individual LTFs, only two speakers received different classifications in English and French. Both cases came from LTF2, where 201 was classified as a worm in English but a chameleon in French, and 230 was classified as a dove in English but a phantom in French. The differences in animal group membership thus lie primarily within their performance in SS comparisons. Crucially, between systems using the same LTF as input, no speakers were found to be (near-)members of the best-performing doves in one language, but classified as the worst-performing worms in the other. The same pattern is extended to the combined systems, although in this case one exceptional speaker (413) was identified as a dove in English and a worm in French. This particular speaker was only ever classified as a dove and not as another animal in English, and only as a worm and never as a member of other groups in French. The anomaly thus appears to follow on from his classification in individual LTFs.

Table 8.2: Total number of speakers in or near any relational group in each system and the number of speakers who shared a classification in both languages.

Input LTF	Total size (En)	Total size (Fr)	N shared
F1	24	25	14
F2	20	20	7
F3	27	25	12
F4	28	26	14
All	22	25	13

8.1.4 Acoustic analysis

Figures 8.11 to 8.14 reproduce the individual distributions of LTFDs, previously displayed in Figure 5.17, with speakers who were classified as outlying speakers additionally marked in each plot. Results of acoustic analysis from the higher formants are presented first, before the lower formants are discussed.

As discussed in Section 5.2, the distributions of LTF3 and LTF4, displayed in Figures 8.11 and 8.12, showed considerable between-speaker variation in the location of the peak frequency. Peaks in individual LTF3 distributions ranged from just above 2000 Hz to around 2900 Hz, while peaks in LTF4 distributions ranged from below 3000 Hz to around 3600 Hz in the case of English and exceed 4000 Hz in the case of French. A high degree of variability is similarly found in the shape of the distribution, in terms of both the height and the sharpness of the peaks. Bimodal distributions were also evidenced, albeit rarely, for individual speakers.

A close examination of the speakers classified as doves in LTF3 and LTF4 suggests that, in both languages, these are all speakers with relatively extreme peak frequencies, on both the low and high ends of the collection of LTFDs. The doves in Figures 8.11 and 8.12 showed a clear separation from the worms (and near-worms), who were generally close to the group norm, represented by the pooled distribution. The distributions of the worm speakers showed peak frequencies that were much further away from the extremes. Their shapes were also relatively unremarkable, with no particularly sharp peaks.

Moving to the lower formants, Figure 8.13 shows that LTF1 peaks were limited to a narrow range in English with relatively little between-speaker variation, whereas in French there was greater variability in peak location. Nevertheless, the shape of the LTF1 distributions demonstrated substantial variability in both languages, especially within the region of frequencies above 500 Hz, where secondary peaks could be found for numerous speakers. Figure 8.14 further demonstrates the variability in the shapes of the distributions of LTF2. While the peaks for most speakers resided within a narrow range,

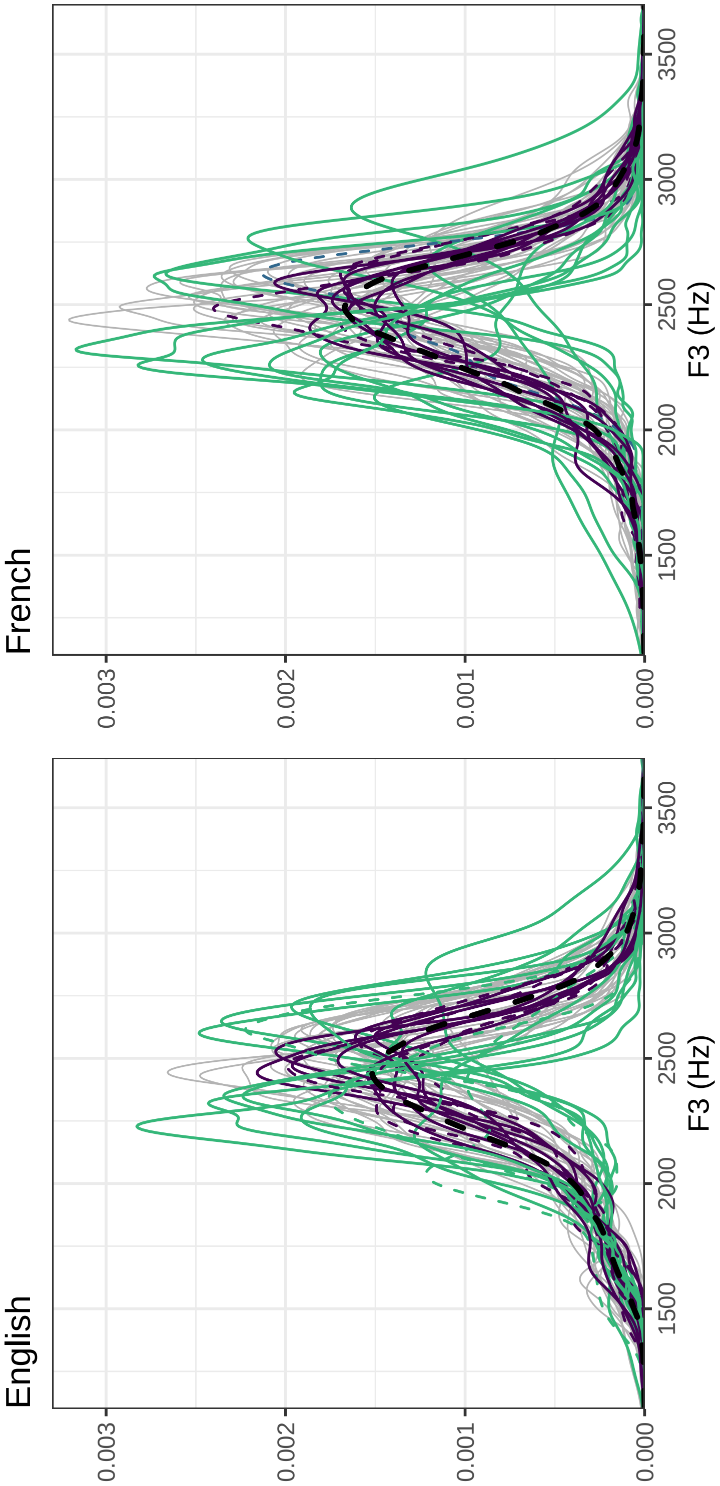

Figure 8.11: Distributions of LTF3 for all speakers in English (left) and French (right), with doves in green, worms in purple and other speakers in grey. Dashed lines indicate near-members of animal groups. Black dashed line indicates group distribution pooled from all 60 speakers.

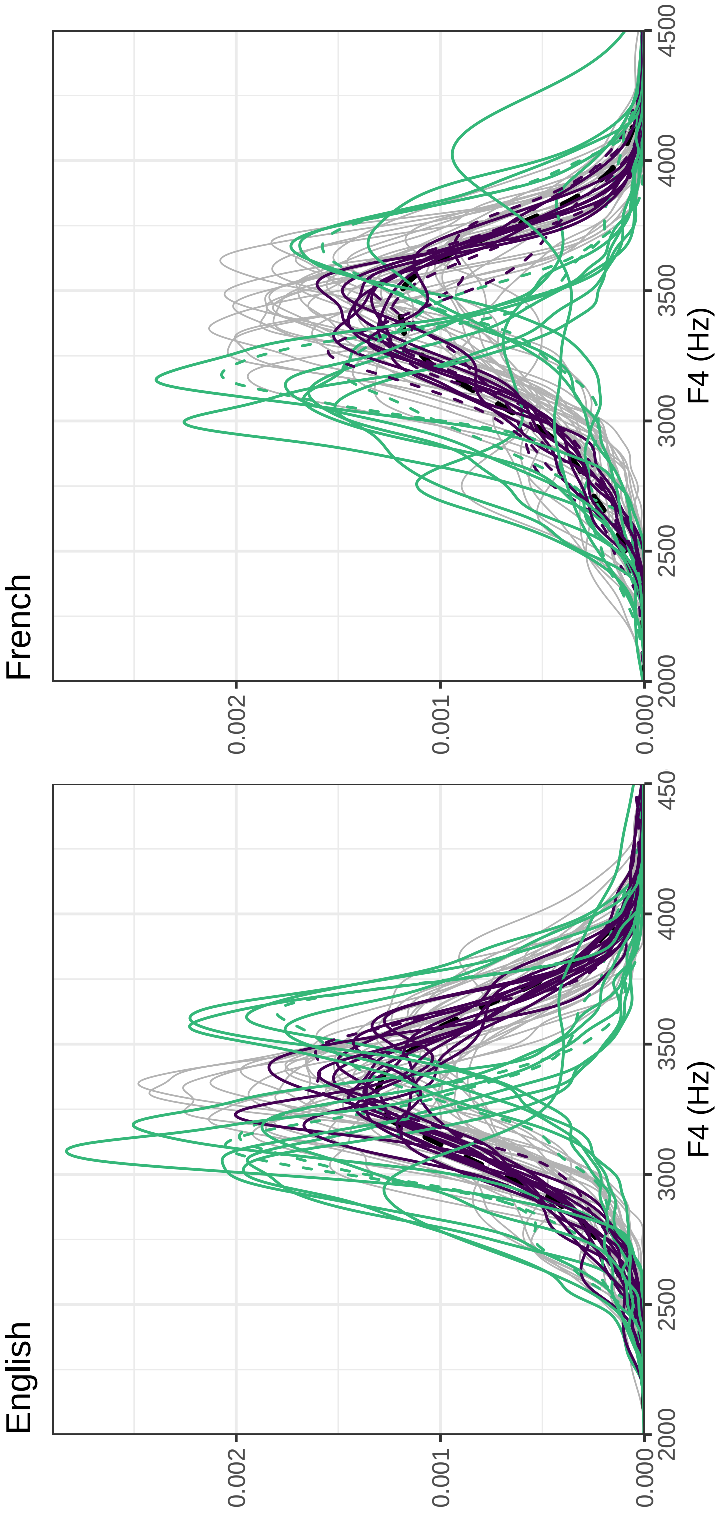

Figure 8.12: Distributions of LTF4 for all speakers in English (left) and French (right), with doves in green, worms in purple and other speakers in grey. Dashed lines indicate near-members of animal groups. Black dashed line indicates group distribution pooled from all 60 speakers.

the presence of secondary peaks and bimodality was not uncommon among this group of speakers in both English and French.

As in the case of LTF3 and LTF4, speakers classified as worms (and to a lesser extent, near-worms) showed distributions that strongly resembled the group norms, both in LTF1 and LTF2. Similarly, many dove speakers could be identified as those whose distributions show peaks at especially low or high frequencies within this population. However, it is also clear that the distributions of some speakers in either language had peaks at frequencies that were by no means extreme, but indistinguishable from the group average. The distributions of these speakers nonetheless exhibited distinctiveness in other ways, through either particularly sharp peaks or bimodality. While phantoms and chameleons were also present in the LTF2 systems, the acoustic correlates of these groups are not analysed in detail due to their low count, although it can be noted that strong bimodality (or even trimodality) appears to be a common characteristic of the distribution of most of these speakers.

In summary, acoustic analysis presented in this section demonstrates a clear contrast between doves and worms in their LTFDs that holds across languages. Dove speakers are mostly accounted for by peaks of relatively extreme frequencies. This is most clearly demonstrated in LTF3 and LTF4, but can also be found in the lower formants. While the remaining doves display unremarkable peak frequencies, they show other distinctive characteristics in their distributions. By contrast, the distributions of speakers classified as worms all tend to be very similar to the overall distributions of the group, in terms of both peak frequency and shape.

8.1.5 Discussion: Same language comparisons

Returning to the main issue of the cross-linguistic stability of the discriminatory potential of LTFDs, findings from the current experiment show that, notwithstanding consistent acoustic-phonetic differences in both formant centre frequencies and bandwidths, LTFDs provided highly similar discriminatory power in English and French. As more LTFs were included in the system, performance in French and in English improved largely in parallel, though French emerged with a small but significant advantage when formant bandwidths were also included in systems using multiple LTFs.

Whereas the analysis in Section 8.1.1 established this on a global, system level, the zooplot analysis in Section 8.1.3, focusing on F-only systems, supplemented this cross-linguistic analysis, adding further evidence that cross-linguistic stability of performance holds on an individual level. Speakers formed highly similar LLR distributions in both languages for each individual LTF and, to a lesser extent, the systems combining all four

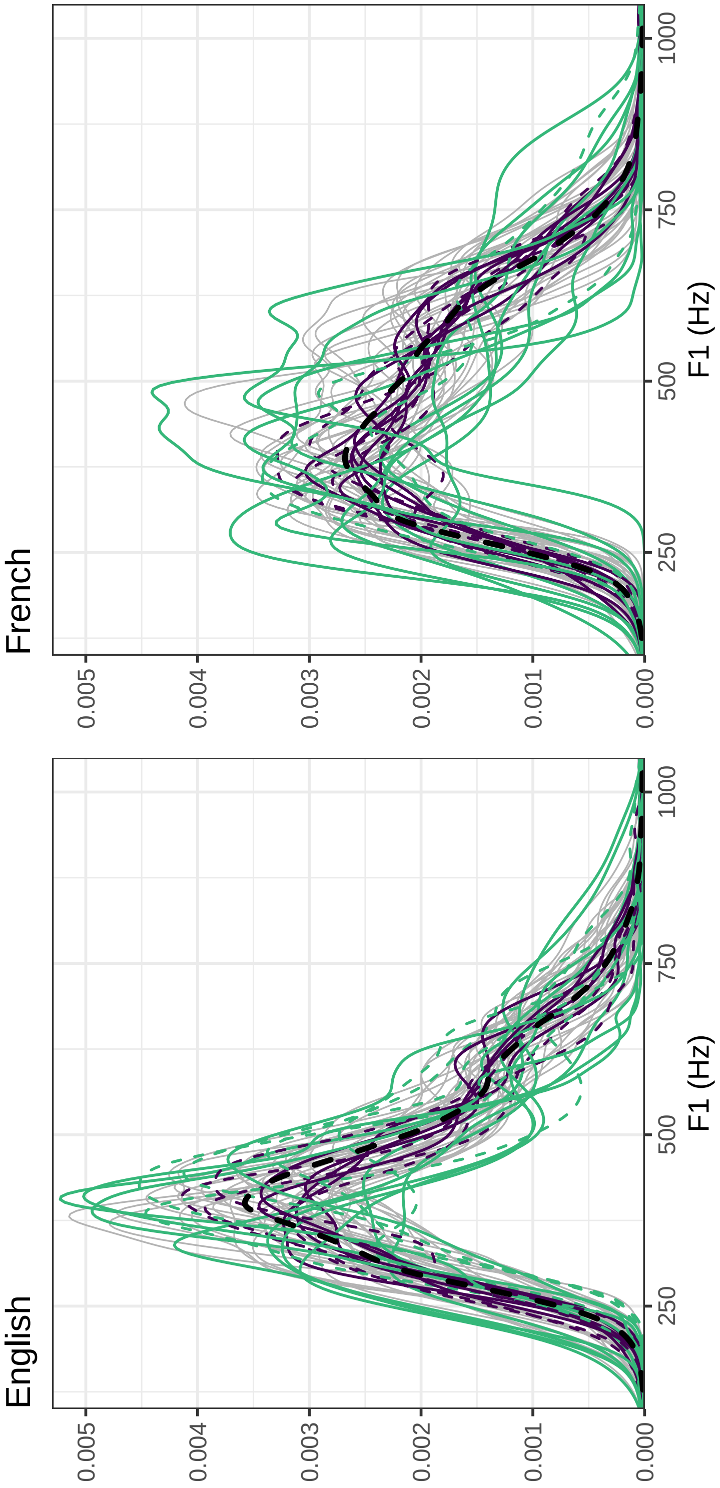

Figure 8.13: Distributions of LTF1 for all speakers in English (left) and French (right), with doves in green, worms in purple and other speakers in grey. Dashed lines indicate near-members of animal groups. Black dashed line indicates group distribution pooled from all 60 speakers.

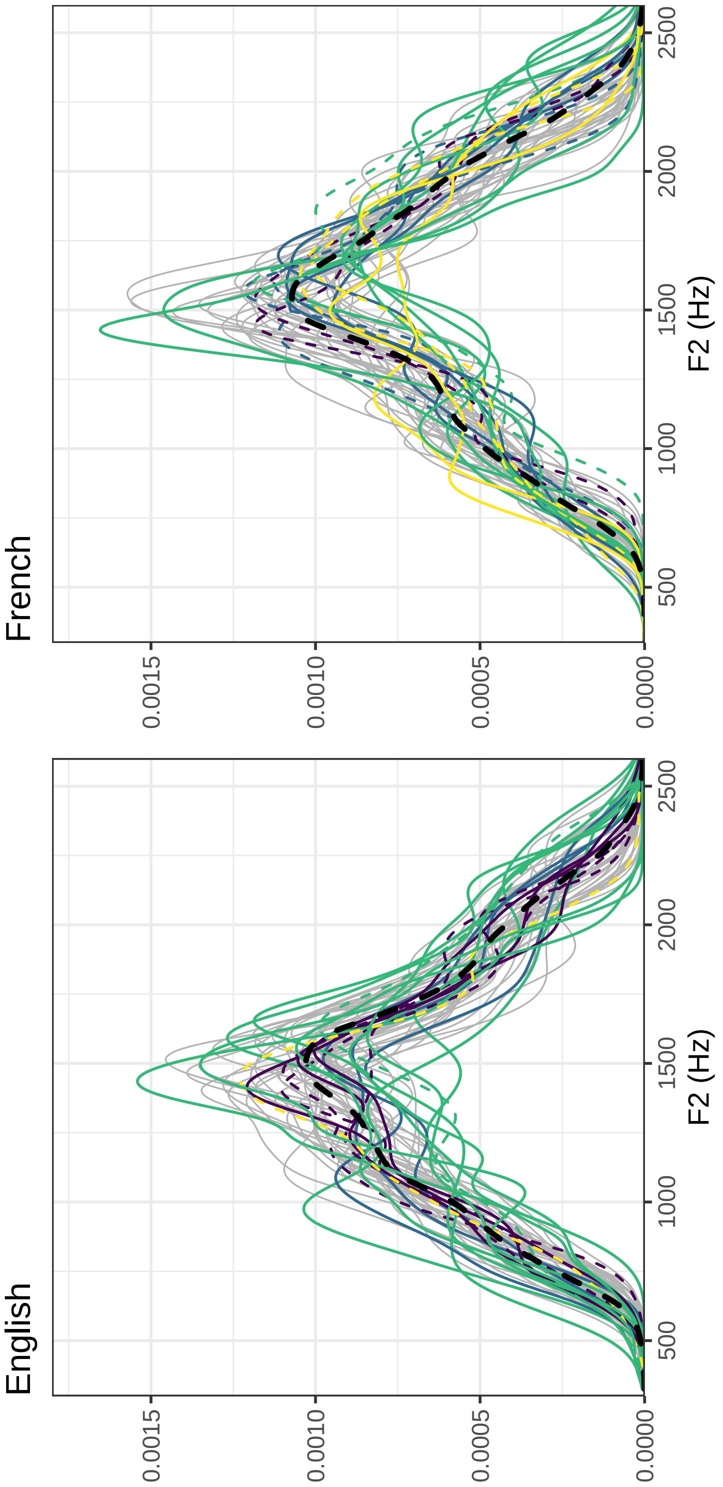

Figure 8.14: Distributions of LTF2 for all speakers in English (left) and French (right), with doves in green, worms in purple, phantoms in blue, chameleons in yellow and other speakers in grey. Dashed lines indicate near-members of animal groups. Black dashed line indicates group distribution pooled from all 60 speakers.

LTFs. Furthermore, speaker classification based on relational groups identified clear correspondences in the sets of best-performing (dove) and worst-performing (worm) speakers, indicating that cross-linguistic stability on the global level is not simply coincidental but accrues from stability of discriminatory performance on the level of individual speakers.

Further evidence to substantiate this conclusion can be found in the acoustic analysis from Section 8.1.4. In systems using each LTF, speakers with the most distinctive distributions, either in the location of the peak frequencies or in the shape of the distribution itself, consistently proved to be the ones who produced the strongest individual performance, while those with uncharacteristic distributions were frequently among the worst performers. Although speakers produced language-specific LTFDs, results from the acoustic analysis in Section 5.2 highlight the strong cross-linguistic consistencies within individual speakers. As the systems tested are relatively effective in harnessing speaker-specific information from the LTFDs themselves, even on the level of individual formants, bilingual speakers with highly distinctive LTFDs in one language were thus capable of maintaining their distinctiveness in the other, whereas speakers whose LTFDs were low on the scale of distinctiveness ended up performing poorly regardless of language. These findings may in part explain why LTF1–3 means could not predict individual performance in Hughes et al. (2018), which explored this relationship using linear regression. Within individual LTFDs, speakers with very low or very high means can both be considered atypical, whereas it is speakers with means near the middle, not near the lower end, of the group who are considered highly typical and tend to provide weak evidence. The current study differs from Hughes et al. (2018) in that, in the present analysis, the underlying LTFDs were only compared with systems using individual formants, and not the combined system as was the case in Hughes et al. (2018), which also included formant bandwidths and delta coefficients in their systems. Therefore, other factors likely also contributed to the lack of association between the acoustic data and speaker performance. Nevertheless, it remains that the relationship between LTFD means and LR performance cannot be captured linearly.

Individual-level analysis was also useful for diagnosing the relative performance and contribution of each LTF. In the present study, although higher formants did not outperform lower formants, when speaker classification is compared across systems, the influence of higher formants appears to dominate in the combined systems, as evidenced by the high proportion of classifications shared with LTF3 and LTF4, whereas the contribution of the lower formants is comparatively limited. The zooplots above show that, in the cases of LTF3 and LTF4, although most speakers had an SS-LLR of low magnitude and were clustered near 0, there were a number of (dove) speakers with exceptional mean LLRs, such that they were distant from the other speakers. LTF1, on the other hand, had relatively few such speakers in French and no such speakers in English. Despite producing a narrow range of mean SS-LLR and DS-LLR overall, speakers not classified as outlying in LTF1 were not as clustered near 0 as they were in LTF3–4. Thus, it may be the case that the performance of higher formants here was driven by a small number of individuals who performed exceptionally well and thus carried over to the combined systems, whereas the performance of LTF1 was driven by the population as a whole.

8.2 Experiment 2: Cross-language comparisons

8.2.1 Global metrics

8.2.1.1 Language mismatch

C_{llr} and EER from En–Fr systems, in which the QS was in English and the KS was in French, are presented in Figures 8.15 and 8.16. As Figure 8.15 shows, systems in Condition C1, where the language of the background data was matched with the KS, and C2, where the language of the background data was matched with the QS, all yielded higher C_{llr} than corresponding En–En and Fr–Fr systems in each set of parameters. In all cases, mean C_{llr} remained below 1, although in the F-only LTF2 systems, C_{llr} was only marginally lower at 0.92–0.94. The effect of including formant bandwidths found above was mirrored in the current set of findings, as F+BW systems resulted in lower C_{llr} than F-only systems in all cases, although the differences between the two sets of input were relatively small in the cases of LTF4 and all formants combined. EER exhibited similar patterns (Figure 8.16), with both cross-language conditions producing much higher EER than same-language comparisons. In systems using individual LTFs, EER was generally higher for F-only systems than for F+BW systems, with the worst performance of 39.6–41.1% coming from F-only LTF2 systems. EER from systems combining all four LTFs, on the other hand, were all within 10–15% and did not show any clear differences between F-only and F+BW systems.

Conditions C1 and C2 generally had a differential impact on C_{llr}, but the patterning varied across each LTF combination. For LTF1 and LTF2, Condition C1 produced lower C_{llr} than Condition C2, while for LTF4 and the combination of all four formants, the trend was reversed. For LTF3, systems in the two conditions produced very similar C_{llr} values. The effect of condition on C_{llr} was also not uniform between F-only and F+BW systems, as demonstrated by a significant interaction between condition and input in the

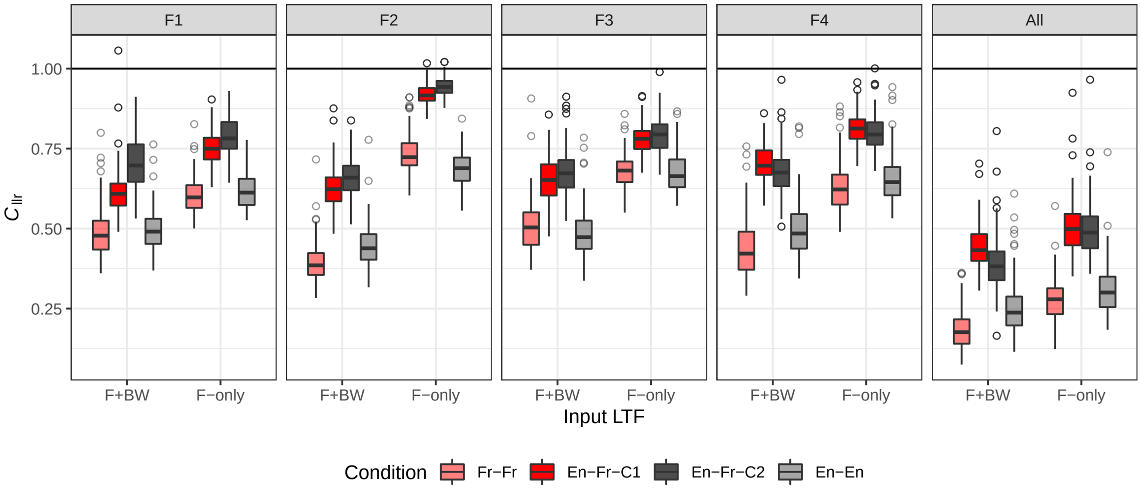

Figure 8.15: C_{llr} from all tested En–Fr LTFD systems in Conditions C1 and C2, with reference to same-language comparisons in English (En–En) and French (Fr–Fr). Horizontal rule reference at $C_{\text{llr}} = 1$.

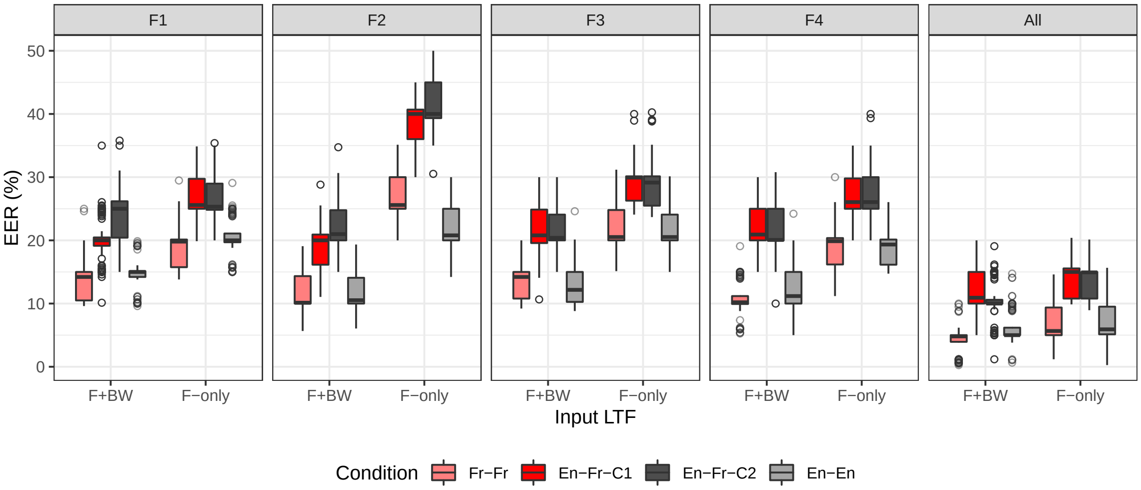

Figure 8.16: EER from all tested En–Fr LTFD systems in Conditions C1 and C2, with reference to same-language comparisons in English (En–En) and French (Fr–Fr).

models fitted to each set of parameters (see Table D.3 in Appendix D), with a tendency for F+BW systems to have a greater distance between same- and cross-language comparison conditions than F-only systems. Post-hoc Tukey-adjusted comparisons by input found that the differences between Conditions C1 and C2 were significant in each case ($p < .01$), with the exception of F-only systems using LTF3 ($p = .2923$), LTF4 ($p = .4423$) and all formants combined ($p = .9827$).

EER, however, showed generally little difference between Conditions C1 and C2, with the exception of LTF2 systems. Model comparisons found a significant interaction between condition and input in all cases, except LTF3, where there were only main effects of condition and mode of input. Post-hoc comparisons further showed that, in F-only systems, the difference between Conditions C1 and C2 was only significant for LTF2 ($p = .0024$) but not for other systems ($p > .05$), while in F+BW only systems, Condition C1, relative to Condition C2, produced significantly lower EER for LTF1 and LTF2 ($p < .0001$), significantly higher EER for all LTFs combined ($p < .0001$), but EER that was not significantly different for LTF3 and LTF4 ($p > .05$).

Figures 8.17 and 8.18, which illustrate C_{llr} and EER obtained from the opposite language pairing, clearly demonstrate the effect of language mismatch when French QS was compared with English KS. Conditions C1 and C2 both resulted in higher C_{llr} and EER than same-language comparisons across all tested systems. Similar to the En–Fr systems above, F-only systems in cross-language conditions generally produced higher C_{llr} and EER than F+BW systems, but the differences for systems using LTF1, as well as for the combination of all four formants, were negligible. Within each set of parameters, Conditions C1 and C2 produced largely similar metrics values, although C_{llr} and EER in Condition C2 showed a tendency to be higher than those in Condition C1, with particularly strong differences in the combined systems. Significant interaction between condition and input found in model comparisons for each set of input parameters suggests that the effect of condition was conditioned by the mode of input, with cross-language conditions generally causing a steeper rise of C_{llr} and EER in F+BW systems than in F-only systems, such that the distance between F+BW and F-only systems narrowed in cross-language comparisons. Post-hoc comparisons by input in each case indicated that most of the differences in C_{llr} between Conditions C1 and C2 did not reach significance ($p > .05$). Condition C2 produced significantly higher C_{llr} only in the combined systems ($p < .0001$) and in the F+BW systems for LTF3 ($p = .0002$). EER, on the other hand, was significantly lower in Condition C1 for most systems ($p < .05$), with the exception of LTF2 and F-only systems using LTF4.

Overall, system validity in cross-language comparisons involving LTFDs was gener-

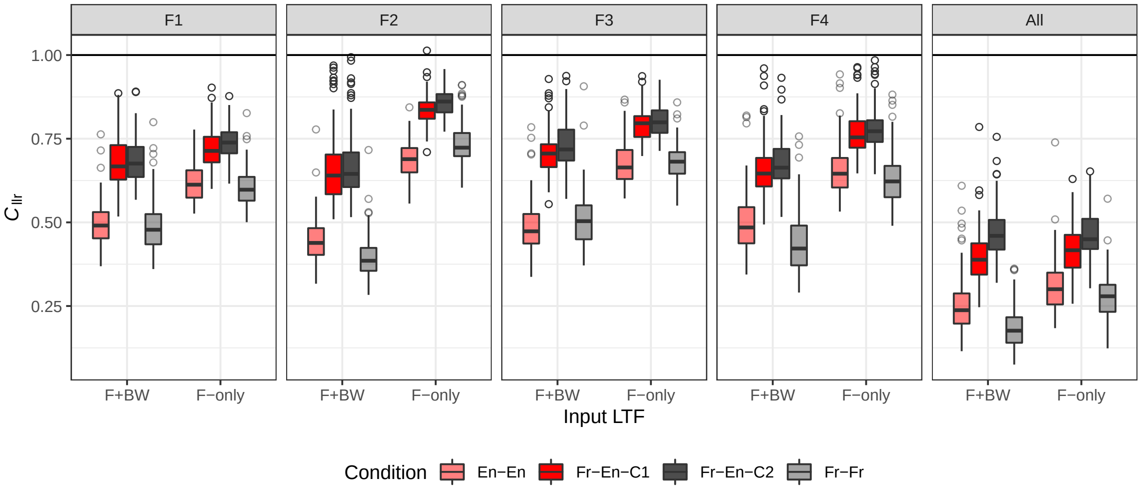

Figure 8.17: C_{llr} from all tested Fr–En LTFD systems in Conditions C1 and C2, with reference to same-language comparisons in English (En–En) and French (Fr–Fr). Horizontal rule reference at $C_{\text{llr}} = 1$.

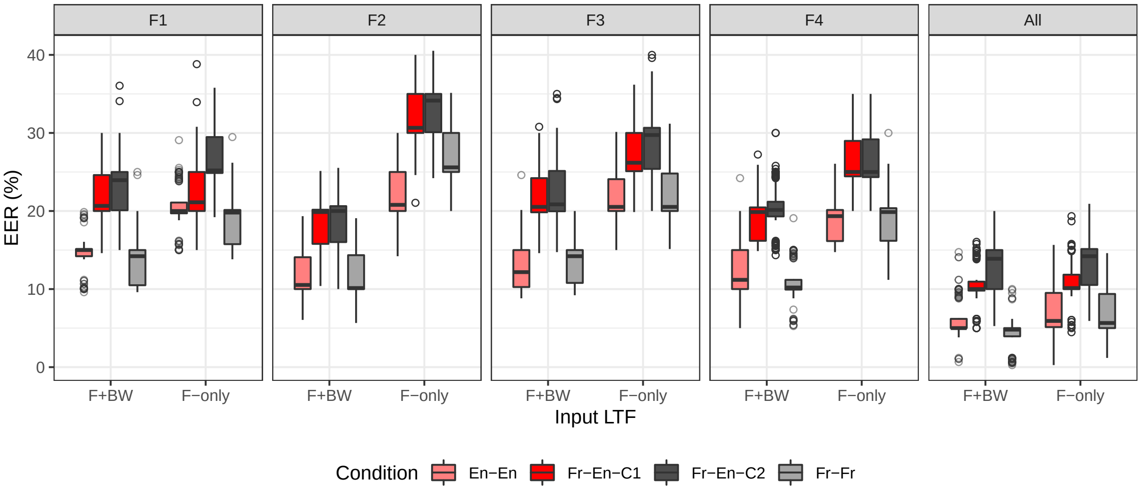

Figure 8.18: EER from all tested Fr–En LTFD systems in Conditions C1 and C2, with reference to same-language comparisons in English (En–En) and French (Fr–Fr).

ally worse than that in same-language comparisons. While C_{llr} remained below 1 on average and indeed in the vast majority of replications for systems using individual LTFs, the magnitude of C_{llr} across these systems also indicates that the discriminatory power they could provide was very limited. This is further supported by the substantial increase in EER, particularly in En–Fr comparisons. In comparison, systems combining multiple LTFs can be said to be more robust, as mean C_{llr} did not rise above 0.5. Between En–Fr and Fr–En comparisons, Table 8.3 shows that system performance was generally similar for the two language pairings. Notably, none of the systems resulted in a substantial expansion of the range of C_{llr}, as the largest C_{llr} out of 100 replications barely exceeded 1. The deterioration of system performance in cross-language comparisons was, in this case, not accompanied by a corresponding fall in system reliability.

Table 8.3: Mean (maximum) C_{llr} and EER from En–Fr and Fr–En LTFD systems for each parameter combination, aggregated over all modes of data input and Conditions.

	En–Fr		Fr–En	
	C_{llr}	EER	C_{llr}	EER
F1	0.71 (1.06)	24.3% (35.8%)	0.71 (0.90)	23.4% (38.8%)
F2	0.79 (1.02)	30.6% (50.0%)	0.76 (1.01)	25.9% (40.5%)
F3	0.73 (0.99)	25.7% (40.1%)	0.76 (0.94)	25.1% (40.0%)
F4	0.75 (1.00)	24.7% (40.0%)	0.72 (0.99)	23.1% (35.0%)
All	0.46 (0.97)	12.8% (20.4%)	0.44 (0.79)	12.0% (20.9%)

8.2.1.2 Training mismatch

Figure 8.19 presents C_{llr} from systems of En–Fr comparisons with training mismatch, namely when the training data originated from same-language comparisons rather than language comparisons, and systems without training mismatch, and Figure 8.20 presents results from systems of Fr–En comparisons. Both Figures demonstrate that systems with training mismatch produced higher C_{llr}, as well as a greater range of C_{llr} across all replications, than systems with no training mismatch. At the same time, there was considerable variation of the effect size of training mismatch in different systems. C_{llr} increased more and reached higher levels in En–Fr comparisons, especially in systems using LTF2. By contrast, there was only a modest increase of C_{llr} in Fr–En comparisons, and as such the primary effect of training mismatch in this case appears to be on the range of C_{llr} obtained. Mean C_{llr} in other systems with training mismatch did not increase to above 1, except for Condition C2 in LTF2 systems of En–Fr comparisons (F+BW: 1.02; F-only: 1.65). Nevertheless, in most systems with training mismatch, C_{llr} of individual replications frequently exceeded 1, with a maximum C_{llr} close to or even beyond 1.5 in many

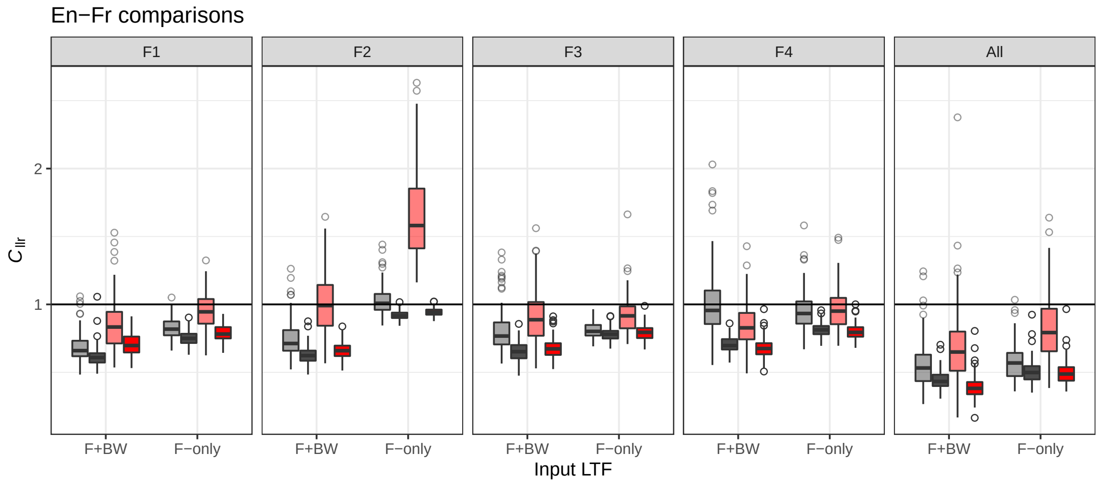

Figure 8.19: C_{llr} from all tested En–Fr LTFD systems in Conditions C1 (grey) and C2 (red), with (light fill) and without (dark fill) training mismatch.

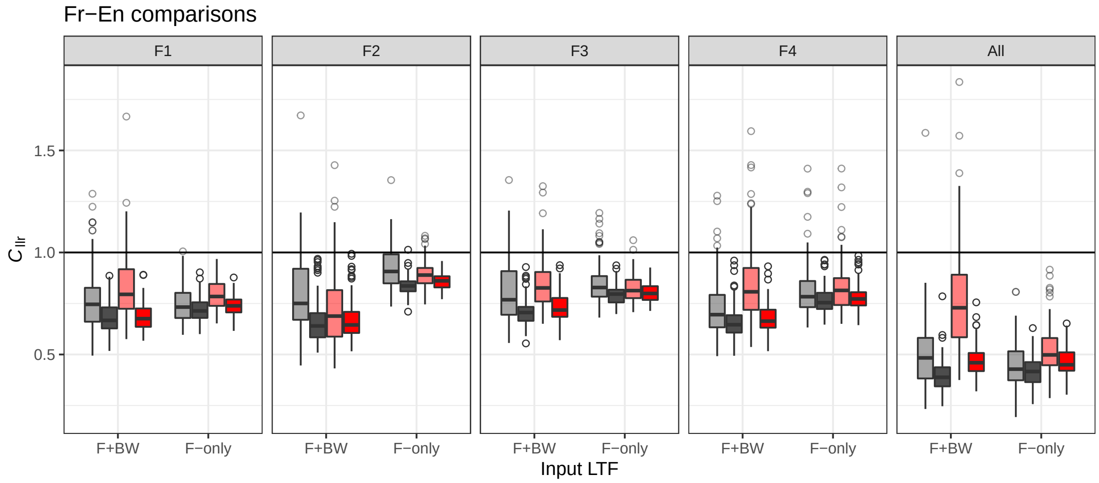

Figure 8.20: C_{llr} from all tested Fr–En LTFD systems in Conditions C1 (grey) and C2 (red), with (light fill) and without (dark fill) training mismatch.

cases.

Model comparisons (presented in full in Table D.2) showed that the presence of training mismatch variably interacted with the factors of condition and mode of input, such that the effect of training mismatch depended on the other two factors.

In En–Fr comparisons, a significant interaction between training mismatch and condition was found for LTF1 and LTF3, such that training mismatch had a greater effect on C_{llr} in Condition C2 than in Condition C1. training mismatch also independently interacted with the mode of input for LTF3, with a more substantial impact on F+BW systems than on F-only systems. Other sets of systems showed a significant three-way interaction of all three factors (training mismatch, condition and input). For systems using LTF2 and the combination of all formants, C_{llr} experienced a greater upward shift due to training mismatch in Condition C2, especially in F-only systems. Systems using LTF4 showed a different trend to the other systems, with C_{llr} rising more in Condition C1 than in Condition C2, but this was limited to F+BW systems only.

In Fr–En comparisons, training mismatch consistently formed a significant interaction with the mode of input, such that it had a greater effect on F+BW systems than F-only systems in all cases. In the case of LTF4, such a differential effect was further conditioned by the factor of condition, as evidenced by a significant three-way interaction, and manifested only in Condition C2. Significant three-way interactions were also found for LTF3 and the combination of all formants, but different trends emerged for C_{llr} in these two sets of systems. For LTF3, training mismatch had a stronger effect on Condition C1, but this was limited to F-only systems. For the combined systems, C_{llr} in Condition C2 experienced a greater upward shift than that in Condition C1, a trend that was amplified in F+BW systems. For LTF1 and LTF2, training mismatch also interacted with condition, but there was no significant three-way interaction. The effect of condition on training mismatch did not run in the same direction for these two LTFs: In the case of LTF1, training mismatch led to a steeper increase of C_{llr} in Condition C2 than in Condition C1, but for LTF2 Condition C1 was instead more affected.

In summary, the effects of training mismatch on systems using LTFDs in cases of cross-language comparisons are twofold. While it generally resulted in higher mean C_{llr}, the extent to which each system was affected differed considerably. Across both En–Fr and Fr–En comparisons, there was an overall tendency for F+BW systems to be more affected by training mismatch than F-only systems, as well as for systems in which the background data matched in language with the QS (Condition C2) to experience a more substantial impact than systems in which the language of the background matched with the KS language (Condition C1). Training mismatch also consistently led to a deterioration of system stability. This was clearly evidenced in Figures 8.19 and 8.20, which show a clear expansion of the C_{llr} range over all 100 replications, even in systems where training mismatch did not result in a sharp increase of C_{llr}.

8.2.1.3 Interim discussion

The results in this section demonstrate that the speaker-discriminatory potential of LTFDs suffered as a result of language mismatch between the QS and the KS. Both C_{llr} and EER increased when compared to systems with no language mismatch in either English or French, although C_{llr} did not rise above 1 even when only individual LTFs were used without the inclusion of formant bandwidths. These findings suggest that LTFDs might be capable of retaining some speaker-specificity when compared across languages, but the amount of speaker-specific information that can be captured cross-linguistically is limited.

System validity of LTFDs in cases of language mismatch was further shown to depend on the representation of the background population. In particular, matching the language of the background data with the KS language in Condition C1 instead of the QS language in Condition C2 resulted in stronger C_{llr} and EER, although the differences were typically small and the direction of the effect was not wholly consistent. The present findings provide further empirical evidence for the consequence of not matching the conditions of the reference population with those in the KS (Morrison et al., 2021). As LTFDs in this study have been consistently shown to be language-specific, the similar levels of performance of the two conditions may be a reflection of the workings of the GMM–UBM technique, rather than similarities in the distributions of the reference data *per se*. In this approach, the model of the known speaker is intended to be tightly coupled with that of the UBM representing the background population and so is adapted from the UBM, rather than independently built from scratch using their own data (Reynolds et al., 2000). As the language of the reference data varied between Conditions C1 and C2, the speaker model for the KS would be adapted differently. The dependence of the speaker model on the UBM may thus have served to reduce the distance between the GMM of the known speaker and the UBM when language mismatch was introduced, and in turn mitigated any imbalance in the mismatch between the two samples and the mismatch between the QS and the reference data.

Further, training mismatch predictably also impacted system validity, as systems with training mismatch produced poorer C_{llr} and were thus less well-calibrated. There was, however, considerable variation in the extent to which each system was impacted. In general, performance deteriorated more substantially for Condition C2 than for Condition C1. In many cases, C_{llr} for Condition C2b reached close to or above 1. The effect of matching the language of the reference data with the QS language thus appears to be exacerbated when training scores are only derived from a single language. In this experiment, Conditions C1b and C2b were set up to simulate the scenario where the analyst only had access to reference materials in one of the two languages. The set-up of these systems in these conditions minimally differed from those of same-language comparisons, only in the language of the QS for Condition C1b or the KS for Condition C2b. The current findings would mean that, in En–Fr comparisons, attempts to carry on with cross-language comparisons using LTFDs would result in poor system validity if only reference materials in French (the KS language) were available, and far worse system validity if only reference materials in English (the QS language) were available. While similar trends were also found in the other language pairing, Fr–En comparisons were generally much more robust to the impact of training mismatch and showed less drastic rises in C_{llr} than En–Fr comparisons. The differences in C_{llr} between Condition C1b and C2b in Fr–En comparisons were also not as large as those in En–Fr comparisons, pointing to some kind of language effect also at work, potentially carried over from same-language comparisons and amplified in cross-language comparisons, such that systems relying on French reference and training data were less susceptible to the effects of training mismatch than those relying on English data.

In Section 5.3, it was predicted that LTF2–4 would be more adversely impacted by language mismatch than LTF1 due to systematic shifts in their distributions. The results above did not find strong support for this prediction, as C_{llr} for all LTFs increased by similar amounts in both En–Fr and Fr–En comparisons. The discriminatory power of LTFDs in cross-linguistic comparisons, as modelled through the GMM–UBM approach, thus appeared to be sensitive not only to overall shifts of the distribution. Indeed, as Figure 8.13 demonstrates, while LTF1 distributions in English and French did not differ in the location of the peaks, there were subtle cross-linguistic shifts in the shape of the distributions, where LTF1 in French typically had less sharp peaks than LTF1 in English. The results in this section suggest that these distinctions were sufficient to have a similar level of impact on speaker-specificity as other LTFs which exhibited arguably greater cross-linguistic differences in both mean/peak frequency and shape.

On the other hand, the prediction that including formant bandwidths would be detrimental rather than advantageous to the discriminatory potential of LTFDs in cross-language comparisons was generally borne out. Both C_{llr} and EER increased more sharply when formant bandwidths were included, suggesting that cross-linguistic differences LTFDs were further compounded by corresponding systematic differences in formant bandwidths found in Chapter 5. The advantages of including formant bandwidths were not fully neutralised by language mismatch, as F+BW systems still outperformed F-only systems in both metrics. As in the case of same-language comparisons, however, the additional value brought by formant bandwidths was very minor when multiple LTFs were combined.

8.2.2 Individual-level analysis

The effect of language mismatch on individual performance is explored in this section. Following Sections 7.2.2 and 8.1.3, as well as to facilitate follow-up acoustic exploration, the current section focuses on F-only systems with no training mismatch. Zooplots for both Conditions C1 and C2 were constructed, but a comparison of speaker classification between the two conditions, presented in Figure 8.21 shows overall good correspondence for LTF1, LTF3, LTF4 and the combined systems. Correspondence of speaker classification was relatively weak for LTF2, which was characterised by an abundance of phantom and chameleon speakers in both En–Fr and Fr–En comparisons. Nevertheless, following Section 7.2.2, only zooplots and speaker classification from Condition C1, where the background data were matched with the KS in the language spoken, are shown and analysed in further detail.

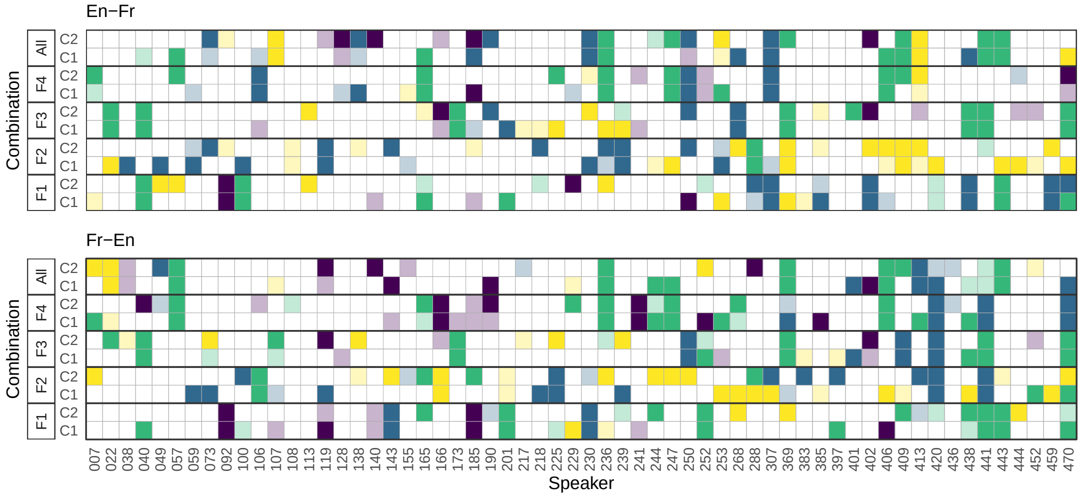

Figure 8.21: Summary of speakers classified as doves (green), worms (purple), phantoms (blue) or chameleons (yellow) in Conditions C1 and C2 for all LTFD systems tested (near-members of each group in lighter shade).

8.2.2.1 LTF1

Figure 8.22 presents the zooplots for both En–Fr and Fr–En comparisons in LTF1 systems. Both plots show that no speakers produced a positive mean DS-LLR in either set of comparisons, but mean DS-LLR was between 0 and -1 for the vast majority of speakers, suggesting that on average speakers were able to be distinguished from other speakers, but the level of performance was weak overall. On the other hand, in both cases, nearly 25% of speakers produced a negative mean SS-LLR, corroborated by the proximity of the 25th percentile to SS-LLR $= 0$, while many other speakers who produced a positive mean SS-LLR nevertheless clustered close to the lower left corner with SS-LLR near 0. There were still some differences which could be found between En–Fr and Fr–En comparisons along the SS dimension, particularly in the behaviour of outlying individuals. The range of mean SS-LLR, from -0.60 to 2.41, was narrower in En–Fr comparisons than in

Fr–En comparisons, in which the lowest SS-LLR was only −0.41 but the highest SS-LLR extended to 3.83. Differences in speaker distribution between En–Fr and Fr–En comparisons were also evidenced through animal classification, summarised in Table 8.4. In particular, En–Fr comparisons contained a higher number of phantoms, clearly visible in the top left corner in Figure 8.22 who performed relatively poorly in SS comparisons but relatively well in DS comparisons. Overall, SS and DS performance showed a weak positive correlation in Fr–En comparisons ($r = -.30,\ p = .0182$), but was not significantly correlated in En–Fr comparisons ($r = -.16,\ p = .2106$).

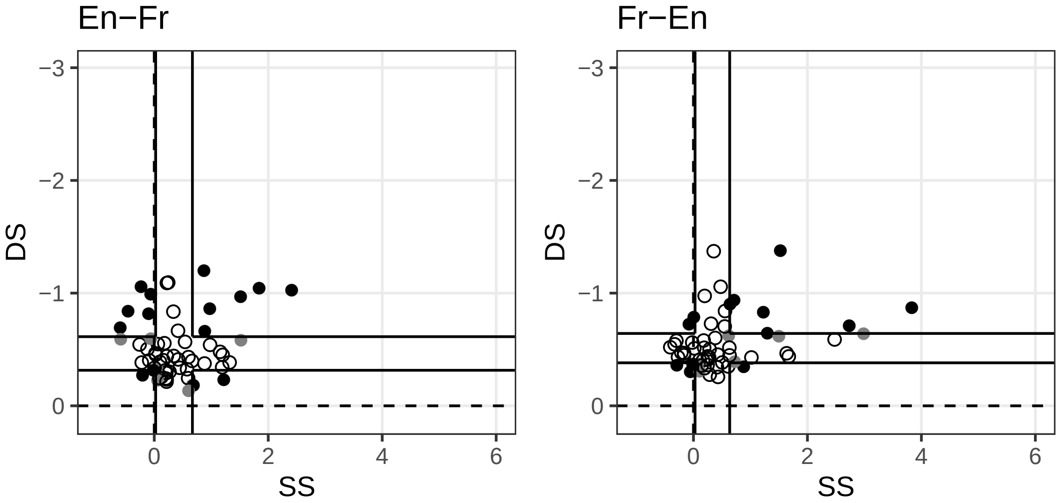

Figure 8.22: Zooplots for systems with LTF1 as input in cross-language En–Fr (left) and Fr–En (right) comparisons.

Table 8.4: Number of speakers in each relational group for systems with LTF1 as input in cross-language comparisons.

Comparison	Dove	Worm	Phantom	Chameleon
En–Fr	6	2	5	2
Fr–En	7	4	2	1

To explore how the relative performance of individual speakers varied between same- and cross-language comparisons, normalised zooplots were constructed following the procedure outlined in Section 7.2.2. The normalised zooplots for LTF1, which juxtapose En–Fr comparisons with Fr–Fr comparisons and Fr–En comparisons with En–En comparisons, are shown in Figure 8.23.

For French KS, shown in the left panel in Figure 8.23, the primary direction of movement between same- and cross-language comparisons was found horizontally in SS comparisons, while the relative vertical position of speakers in terms of DS comparisons remained relatively stable, suggesting that there was greater variation within the population in how speakers' SS performance was affected. In particular, there was a clear distinction between speakers in the top half of the zooplot, who performed relatively well in DS comparisons, and those in the bottom half. Language mismatch had a more adverse effect on SS performance for speakers with relatively strong DS performance, as speakers in the top half commonly shifted towards the left, while speakers in the bottom half tended to move to the right, resulting in the phantom and chameleon groups observed above. There were nonetheless outlying dove and worm speakers whose performance remained relatively stable and exhibited little movement.

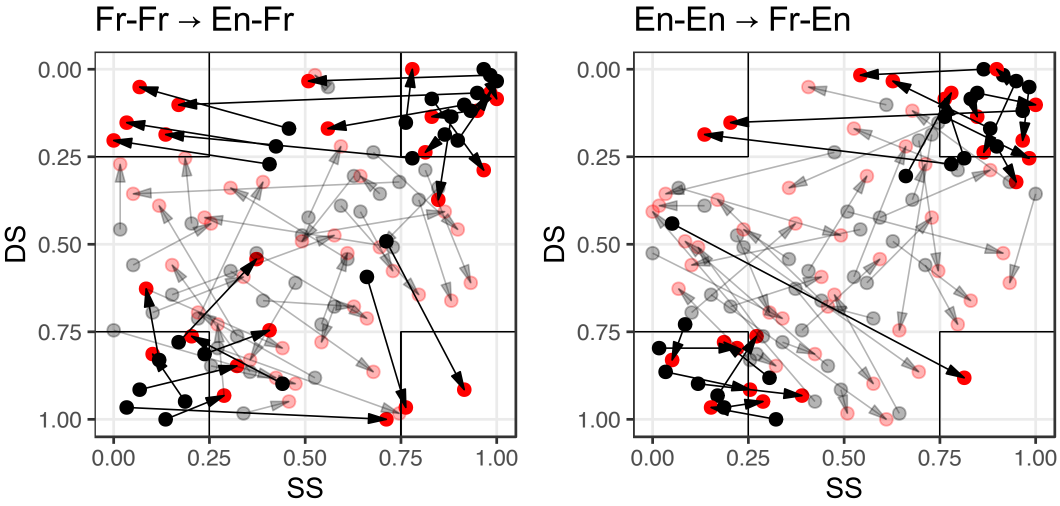

Figure 8.23: Normalised zooplots for systems with LTF1 as input in same-language (black) and cross-language (red) comparisons. Values on axes indicate percentile ranks. Arrows connect speakers from same-language to cross-language comparisons.

For English KS, the pattern of movement was similar to those in French but more limited in degree. With the exception of a small number of speakers, the position of speakers classified as doves or worms in same-language comparisons remained largely stable in cross-language comparisons. More generally, speakers with relatively poor performance in both types of comparisons in the lower left triangle of the zooplot remained within the same area. The same could be said for speakers in the upper right triangle as well, meaning that, in this case, language mismatch did not cause speakers who performed exceptionally well (in both SS and DS comparisons) in same-language comparisons to become some of the worst performers in cross-language comparisons, and vice versa.

8.2.2.2 LTF2

The distribution of speakers in the zooplots for LTF2 systems, illustrated in Figure 8.24, shows a stark contrast to that in LTF1. In both En–Fr and Fr–En comparisons, not only were speakers highly concentrated within a narrow range of SS-LLR and DS-LLR (particularly in En–Fr comparisons), they were also arranged in a top-left-to-bottom-right configuration. Indeed, SS and DS performance was negatively correlated in both cases (En–Fr: $r = .48,\ p = .0001$; Fr–En: $r = .36,\ p = .0046$), such that speakers who performed better in SS comparisons also performed worse in DS comparisons. As Table 8.5 shows, both doves, in the top right corner, and worms, in the bottom left corner, were rare or completely absent, while phantoms and chameleons comprised eight speakers each. Poor performance of LTF2 was strongly evidenced on an individual level, as a large number of speakers produced negative mean SS-LLR (En–Fr: 26, or 43%; Fr–En: 19, or 32%). Poor individual performance was not limited to SS comparisons, as two speakers produced a marginally positive mean DS-LLR in Fr–En comparisons, and seven speakers similarly produced positive mean DS-LLR in En–Fr comparisons.

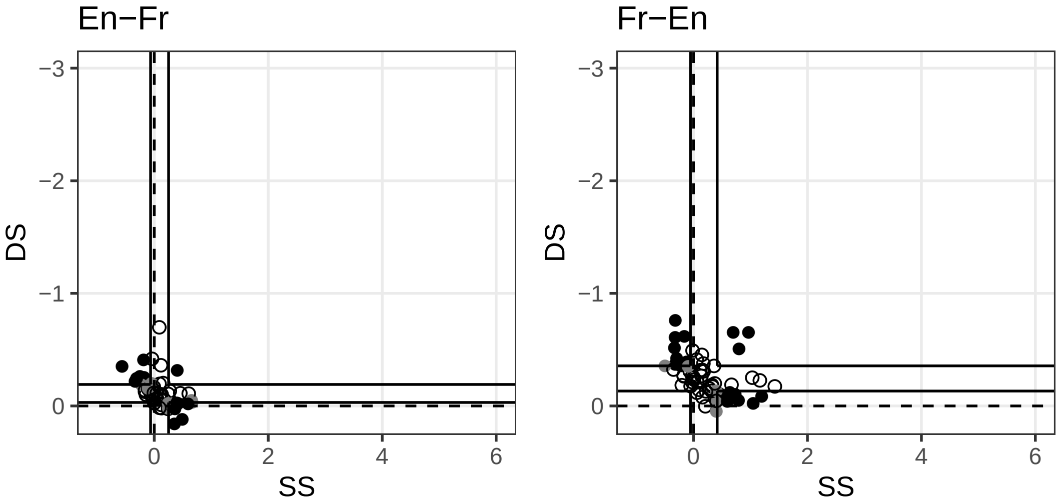

Figure 8.24: Zooplots for systems with LTF2 as input in cross-language En–Fr (left) and Fr–En (right) comparisons.

Table 8.5: Number of speakers in each relational group for systems with LTF2 as input in cross-language comparisons.

Comparison	Dove	Worm	Phantom	Chameleon
En–Fr	1	0	8	8
Fr–En	3	0	8	8

The normalised zooplots for LTF2 are presented in Figure 8.25. Considerable movement between same- and cross-language comparisons can be found along both SS and DS dimensions, especially with English KS (right panel). Regardless of their relative position in SS comparisons, movement from same- to cross-language comparisons for most speakers was towards either the upper left corner or the lower right corner, resulting in the diagonal configuration displayed in Figure 8.24. Consequently, only very few speakers classified as doves in same-language comparisons maintained their classification in cross-language comparisons by having a similar level of performance in both SS and DS comparisons. Instead, many of these speakers ended up as either phantoms or chameleons in either set of cross-language comparisons, while speakers identified as chameleons or phantoms in same-language comparisons in the first place were mostly classified the same way in cross-language comparisons.

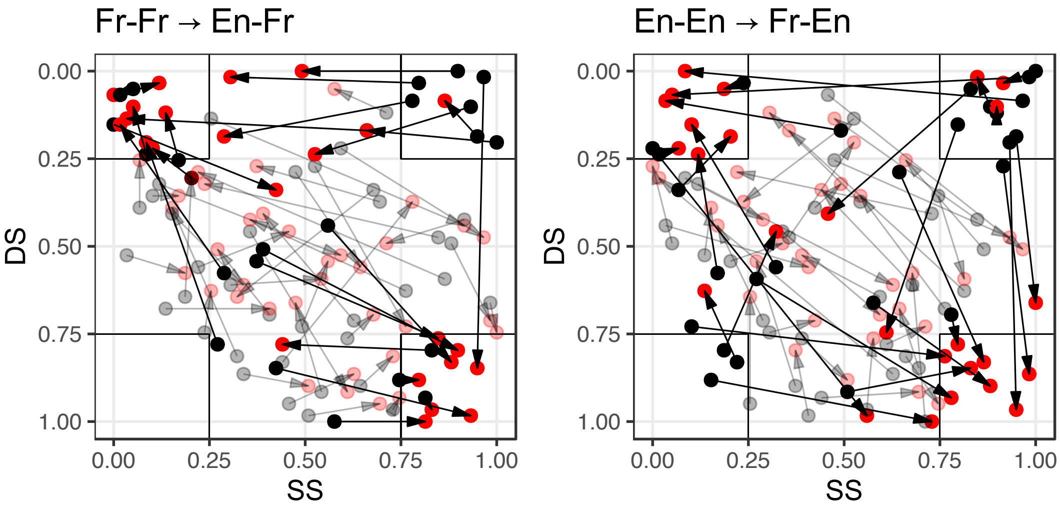

Figure 8.25: Normalised zooplots for systems with LTF2 as input in same-language (black) and cross-language (red) comparisons.

8.2.2.3 LTF3

Zooplots for cross-language comparisons with LTF3, shown in Figure 8.26, depict subtle differences in the distribution of speakers between En–Fr and Fr–En comparisons. In the former set, speakers were clustered near SS-LLR $= 0$, with a high proportion of speakers (17, or 28%) producing negative mean SS-LLR. Mean DS-LLR for all but three speakers was between 0 and -1, suggesting that the strength of evidence provided was weak for the most part. While no worm speakers were found, seven speakers were identified as doves, with particularly strong performance in both SS and DS comparisons. At the same time, the upper left and lower right corners were only occupied by a small number of phantoms and doves (Table 8.6), such that a moderately strong correlation was found between performance in SS and DS comparisons ($r = -.56,\ p < .0001$). In comparison, speakers in Fr–En comparisons were less densely clustered near SS-LLR $= 0$. Non-outlying speakers (those in the middle 50%) produced a similarly narrow range of mean SS-LLR, as the 25th and 75th percentiles along the SS axis were at similar positions in both sets of comparisons, but within this range speakers showed a tendency to have higher mean SS-LLR. With the exception of one particular outlier, all speakers also produced a small mean DS-LLR of between 0 and -1. In this set of comparisons, a slightly higher number of phantoms and lower number of chameleons were found, coinciding with a weaker correlation between SS and DS performance ($r = -.38,\ p = .0025$).

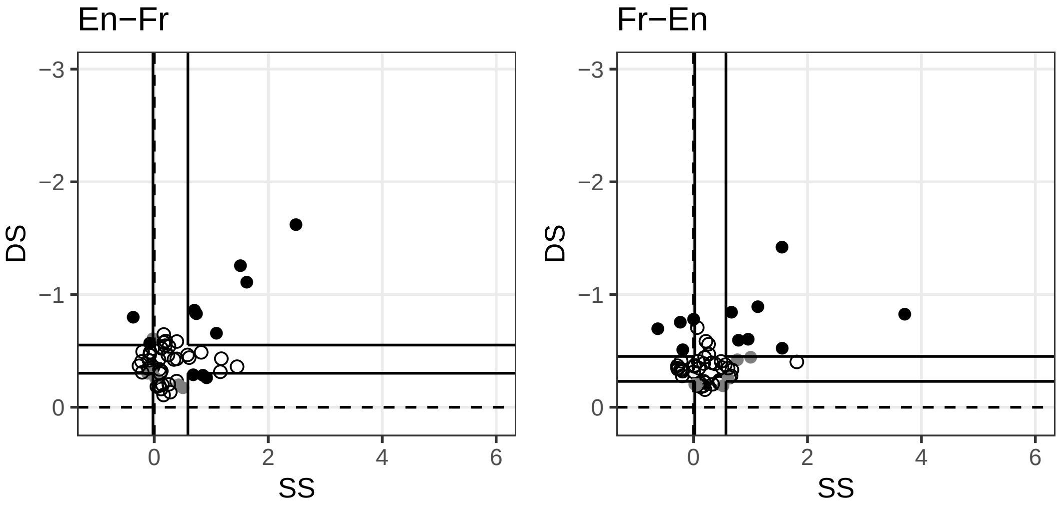

Figure 8.26: Zooplots for systems with LTF3 as input in cross-language En–Fr (left) and Fr–En (right) comparisons.

Table 8.6: Number of speakers in each relational group for systems with LTF3 as input in cross-language comparisons.

Comparison	Dove	Worm	Phantom	Chameleon
En–Fr	7	0	2	3
Fr–En	7	0	4	0

Figure 8.27 shows the normalised zooplots for LTF3. In both plots, there is no clear pattern of movement for speakers in the central portion of the plot, though notably speakers primarily displayed greater movement along the SS dimension than the DS dimension. The distribution of outlying speakers in same- and cross-language compar-

isons, however, showed different patterns of movement. Worms in SS comparisons typically displayed some movement along either axis, so that the level of their performance was no longer poor relative to the population in both SS and DS comparisons, but was still relatively poor in at least one type of comparison. Their movement, however, was not so substantial that they fell into a different relational group in cross-language comparisons. Indeed, the phantoms and chameleons found in cross-language comparisons were mostly speakers who were classified as doves in same-language comparisons. Unlike other doves in the top right corner who were capable of maintaining a relatively strong level of performance in both types of comparisons, these speakers were among the most adversely affected by language mismatch in SS or DS comparisons.

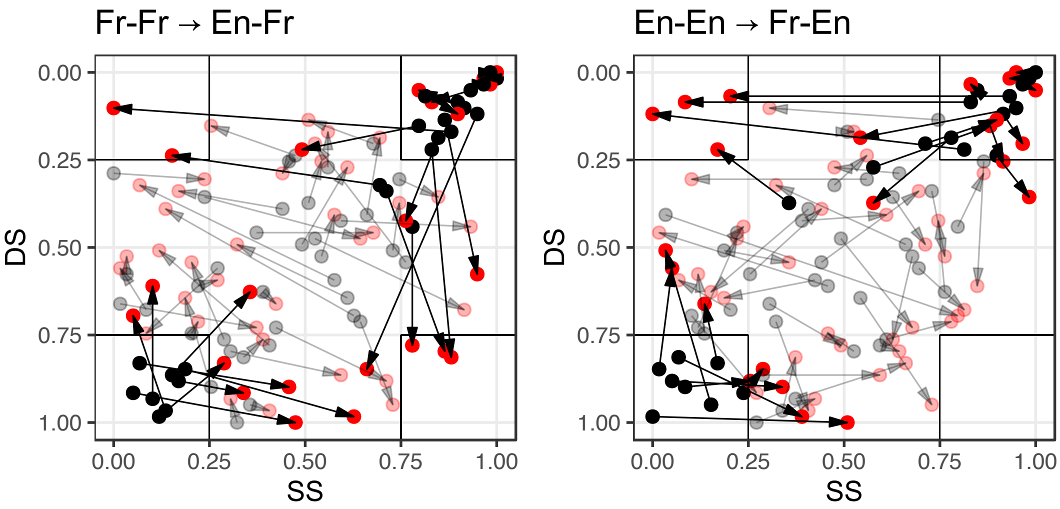

Figure 8.27: Normalised zooplots for systems with LTF3 as input in same-language (black) and cross-language (red) comparisons.

8.2.2.4 LTF4

In the case of LTF4, the distribution of speakers in the zooplots in Figure 8.28 is broadly similar to that for LTF1. In both En–Fr and Fr–En comparisons, speakers were largely concentrated near the lower left corner, although very few speakers were actually classified as worms (Table 8.7). Individual performance in SS comparisons was particularly poor for approximately a quarter of the speakers (14 in En–Fr; 13 in Fr–En) who produced negative mean SS-LLR. As in the other LTFs, mean DS-LLR of speakers did not span a wide range, with only up to two speakers producing a mean DS-LLR beyond -1, whereas their mean SS-LLR had a greater spread. The highest mean SS-LLR was just above 2 (En–Fr: 2.35; Fr–En: 2.15), while the lowest mean SS-LLR reached -0.73 in

En–Fr comparisons and an even greater -1.25 in Fr–En comparisons. Nevertheless, performance remained relatively strong for a number of exceptional speakers classified as doves, who were set apart from the rest of the speakers in the zooplots, especially for Fr–En comparisons. Performance between SS and DS comparisons was weakly correlated in Fr–En comparisons ($r = -.37,\ p = .0035$) but not significantly correlated in En–Fr comparisons ($r = -.15,\ p = .2465$).

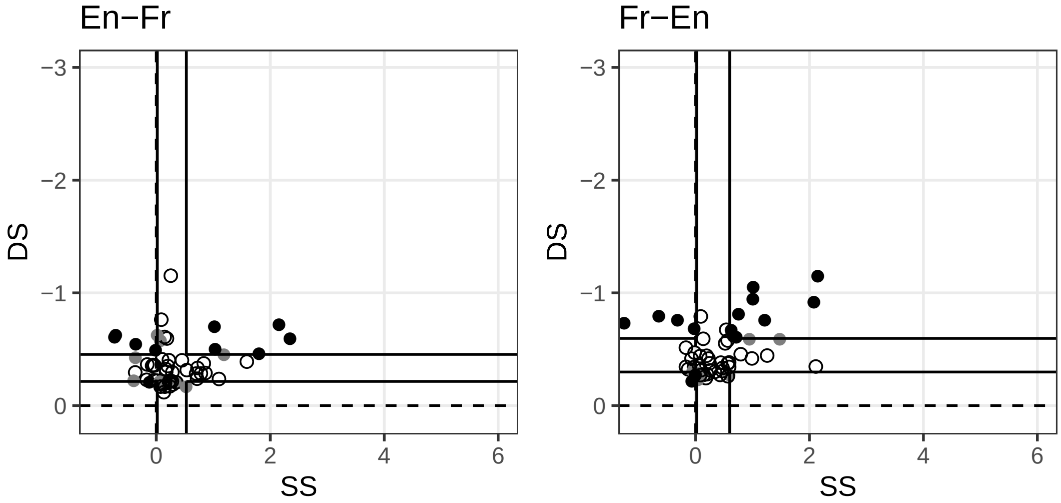

Figure 8.28: Zooplots for systems with LTF4 as input in cross-language En–Fr (left) and Fr–En (right) comparisons.

Table 8.7: Number of speakers in each relational group for systems with LTF4 as input in cross-language comparisons.

Comparison	Dove	Worm	Phantom	Chameleon
En–Fr	5	1	4	0
Fr–En	9	4	4	0

Normalised zooplots for LTF4 shown in Figure 8.29 demonstrate a pattern of movement between same- and cross-language comparisons resembling that for LTF1, in that speakers remained more stable in their relative position along the DS axis while displaying more marked variation in SS comparisons. For KS in both English and French, speakers classified as doves generally exhibited two main types of behaviour. For one group, their performance in both SS and DS comparisons was not as substantially affected as other speakers by language mismatch, such that they remained within or close to the dove group. The other group of speakers were relatively more impacted in SS comparisons and moved towards the top left corner, with a number of speakers being reclassified as phantoms in cross-language comparisons. Movement in the other direction, where a speaker was identified as a dove in cross-language but not same-language comparisons, was decidedly rare. On the other hand, speakers identified as worms in same-language comparisons, found in the bottom left corner of the zooplots, shifted towards the right as a group. When the KS was in French (left panel), the degree of movement of most worm speakers was more consistent, such that none of them were similarly classified in cross-language comparisons. There was more variation in the degree of movement when the KS was in English (right panel), as only some speakers remained worms in cross-language comparisons.

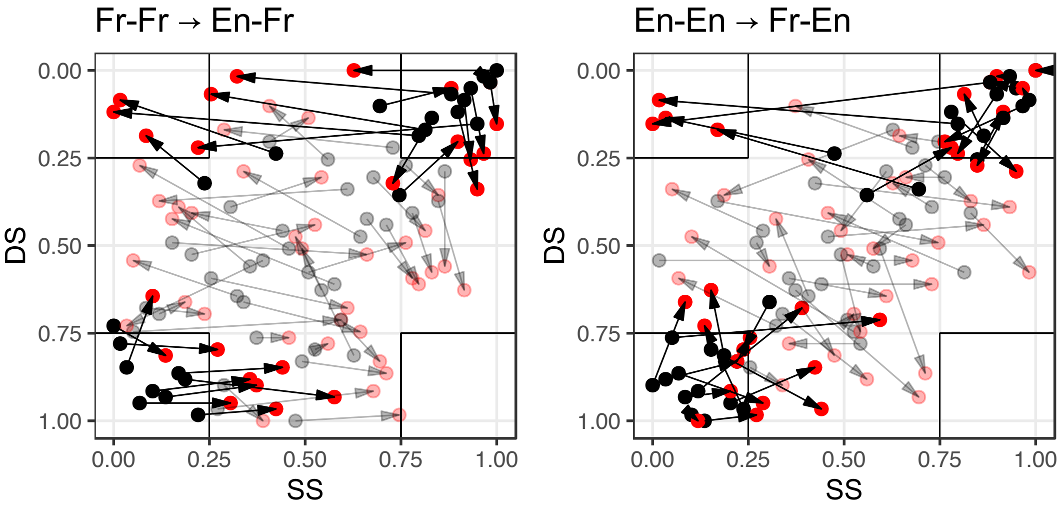

Figure 8.29: Normalised zooplots for systems with LTF4 as input in same-language (black) and cross-language (red) comparisons.

8.2.2.5 All LTFs combined

As shown in Figure 8.30, when all LTFs were combined in cross-language comparisons, the whole distribution of speakers shifted upwards in the direction of more negative DS-LLR. The degree of movement was especially sizeable for Fr–En comparisons, where all speakers produced mean DS-LLR exceeding -1. Nevertheless, in both sets of comparisons, mean DS-LLR of individual speakers was capped at -3. Along the SS dimension, the majority of speakers were more evenly distributed across a wider range, as evidenced by the clear rightward shift of the 75th percentiles that marked the boundaries for the dove and chameleon groups. A comparatively larger number of speakers produced stronger SS-LLR $>$ 2 in both En–Fr and Fr–En comparisons, though only a small number of speakers in Fr–En comparisons yielded much more extreme levels of performance. Better SS performance was also evident from the lower number of speakers

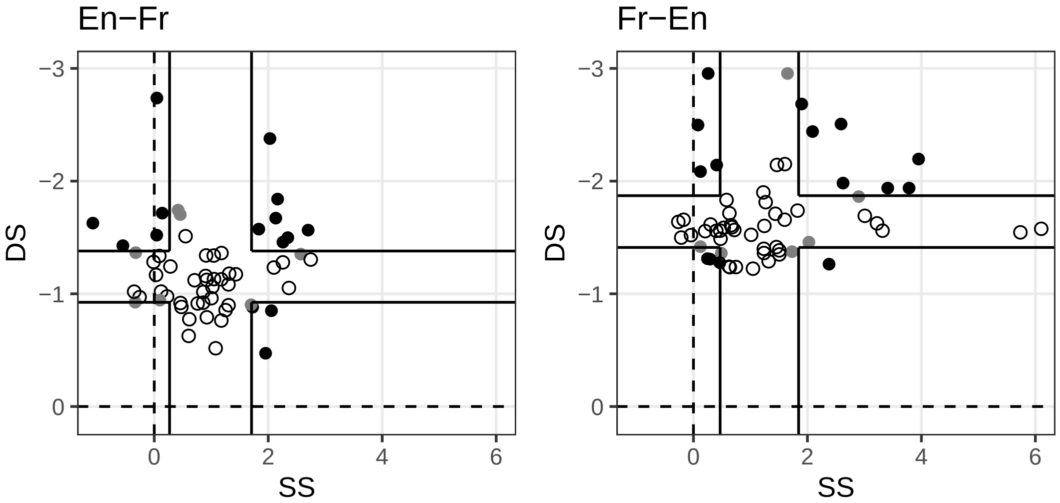

Figure 8.30: Zooplots for systems with all four LTFs as input in cross-language En−Fr (left) and Fr−En (right) comparisons.

producing DS-LLR < 0 (En−Fr: 7; Fr−En: 4). Despite the overall improvement of individual performance in both SS and DS comparisons, the size of the group of phantoms (and chameleons, to some extent) shows that this combination of parameters remained problematic for some speakers, such that they could not be effectively matched against any speaker. Overall, the speaker distribution in Figure 8.30 was generally balanced, and no significant correlation between SS and DS performance could be found for either SS ($r = -.19,\ p = .8839$) or DS ($r = -.14,\ p = .3028$) comparisons.

Table 8.8: Number of speakers in each relational group for systems with all four LTFs as input in cross-language comparisons.

Comparison	Dove	Worm	Phantom	Chameleon
En−Fr	7	0	5	3
Fr−En	7	3	4	1

The comparison of cross-language comparisons with same-language comparisons in the normalised zooplots in Figure 8.31 reveals extensive variation between speakers in how their individual performance was impacted by language mismatch. For KS in French (left panel), there was much movement along both SS and DS dimensions between same- and cross-language comparisons. While a small number of doves in same-language comparisons still produced relatively strong performance in cross-language comparisons and remained doves, other speakers were relatively more adversely affected in either SS or DS comparisons, with the speakers exhibiting the most movement falling into the group of phantoms in cross-language comparisons. Not all phantoms in cross-language comparisons, however, originated as doves in SS comparisons. Indeed, other phantoms were derived from speakers with relatively poor SS performance in same-language comparisons, but whose DS performance improved relative to the rest of the population. Meanwhile, DS performance of speakers identified as worms in same-language comparisons was overall stable, meaning that in cross-language comparisons they remained the speakers who were not well distinguished from other speakers. In opposition to the doves whose SS performance was more adversely affected, their performance in SS comparisons generally improved relative to the other speakers, as evidenced by a rightward shift of their positions on the zooplot.

For KS in English (right panel), the overall trends displayed by the speakers were broadly similar, particularly among those classified as doves in same-language comparisons. While a number of speakers were relatively unaffected and performed well in cross-language comparisons, other doves constituted the group of speakers who displayed the most movement away from the top right corner, to as far as being classified as phantoms in cross-language comparisons. Other phantoms in cross-language comparisons who were not doves were nonetheless similarly speakers who shifted horizontally and thus had relatively good DS performance in both same- and cross-language comparisons. Speakers near the lower left corner were more variable in the direction of their movement, but the distance between their positions in same- and cross-language comparisons was generally quite small, such that speakers with relatively poor SS and DS performance in same-language comparisons did not become speakers who produced particularly strong SS or DS performance in cross-language comparisons.

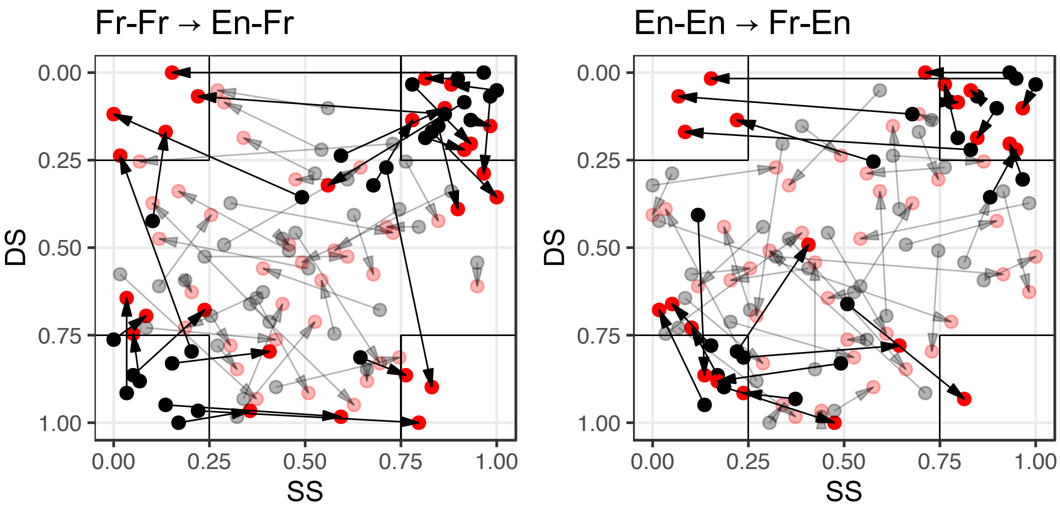

Figure 8.31: Normalised zooplots for systems with all four LTFs as input in same-language (black) and cross-language (red) comparisons.

8.2.3 Acoustic analysis

The individual-level analysis above demonstrates a general decrease in the number of doves and worms across systems using LTFDs in cross-language comparisons, to the point where worms were wholly absent for LTF2 and LTF3, as well as for the combined system in En–Fr comparisons. While doves typically remained the largest relational group, phantoms and chameleons were regularly found, even overtaking the doves as the largest groups in systems using LTF2. This section goes on to explore the acoustic characteristics of the LTFDs of these speakers who were classified as members of relational groups in cross-language comparisons. Following the previous section, only F-only systems with the language of the background data matched with the KS language are considered. Figures 8.32 to 8.35 reproduce the individual LTFDs of such speakers in each of the systems using LTF1–4, plotted alongside the pooled distribution from all 60 speakers in the language of the KS (and hence of the background data). In the interest of clarity, only full members of relational groups are presented and discussed, and near-members and speakers not classified in each system are not included.

Figure 8.32 shows LTF1 distributions for speakers with outlying performance in En–Fr and Fr–En comparisons using LTF1 only. Doves, who performed best in both SS and DS comparisons, showed multiple forms of distinctiveness in their distributions, such as bimodality (040, 100), low peak locations (165, 201) and especially sharp peaks (443, 470). When compared to the KS language, LTF1 showed clear indications of cross-linguistic shifts in its shape for most doves, but differences between the two languages were typically small. There was no major shift in the location of peak frequency for any dove speaker, and each speaker’s distribution either retained their distinctive shape in the other language, or moved in a direction divergent from the group norm of the reference data, such that they became even more distinctive. Speaker 201, one of the four speakers identified as doves in both En–Fr and Fr–En comparisons, displayed the strongest similarities with almost overlapping distributions. The other three speakers similarly classified in both sets of comparisons displayed greater within-speaker differences, but their LTF1 was nonetheless highly distinctive in both languages.

By comparison, chameleons, who also performed relatively well in SS but not in DS comparisons, showed distributions in the KS that were much less distinctive than doves. In En–Fr comparisons, 253 and 369 had LTF1 with a slightly sharper peak in French than the pooled distribution, whereas in Fr–En comparisons the distribution for 229 in English had a heavier shoulder at around 600 Hz than the population. Nevertheless, it remains the case that these distributions were largely in line with the rest of the reference data. In the QS, their distributions shifted in a direction that accentuated those characteristics: The peaks for 253 and 369 in English grew even sharper, while 229 produced a higher concentration of formants around 600 Hz to develop a more prominent secondary peak in French. These differences served to distinguish them further from the population, such that they performed relatively well in SS comparisons. The proximity of their distributions in the KS language with the rest of the population, however, means that their QS was not much more different from their own KS than from the background population. As such, there could be no strong evidence that the acoustic features in the QS were derived from the same speaker rather than any other speaker in DS comparisons.

For speakers classified as worms, with poor performance in both SS and DS comparisons, their LTF1 distributions in the KS language were all uncharacteristic and extremely close to the group norm. Like chameleons, their distributions also showed cross-linguistic dissimilarities, in a number of different forms. For example, 250 produced a sharper peak in English, whereas 406 produced a somewhat bimodal distribution with both peaks at lower frequencies in French. In most cases, such distinctions did not result in highly distinctive peak locations or shapes (cf. 250). Indeed, 092, who was classified as a worm in both En–Fr and Fr–En comparisons, had an LTF1 distribution in each language that resembled the respective overall distribution of the background data. The characteristics shared by worms are highly similar to those that defined the chameleons above, with subtle distinctions in the typicality of the distribution from the QS with respect to the other language.

Phantoms, who performed relatively well in DS but not in SS comparisons, were indeed found to have markedly different distributions in English and French. Their distributions in the KS language were somewhat distinctive, such as through a display of bimodality (307, 459) or sharp peaks (143, 230). Their distributions in the other language, however, did not exhibit any particular trends, bearing variable resemblances to the respective population norm. The distribution for 385 in English, for example, was not far from the group distribution in French, whereas 438 and 459 had LTF1 with a sharp peak in English.

For LTF2, cross-language comparisons found only a small number of doves but a much larger group of phantoms and chameleons. As Figure 8.33 illustrates, all speakers with outlying performance, regardless of the group they belonged to, showed a shift towards higher LTF2 frequencies in French in line with the findings from Section 5.2, though the extent of the movement was highly variable. Notwithstanding these differences, the few doves all showed LTF2 distributions with highly distinctive shapes. Distributions in the QS language either amplified the differences between the KS and the overall distribution of the group (470), or became highly distinctive in their own way,

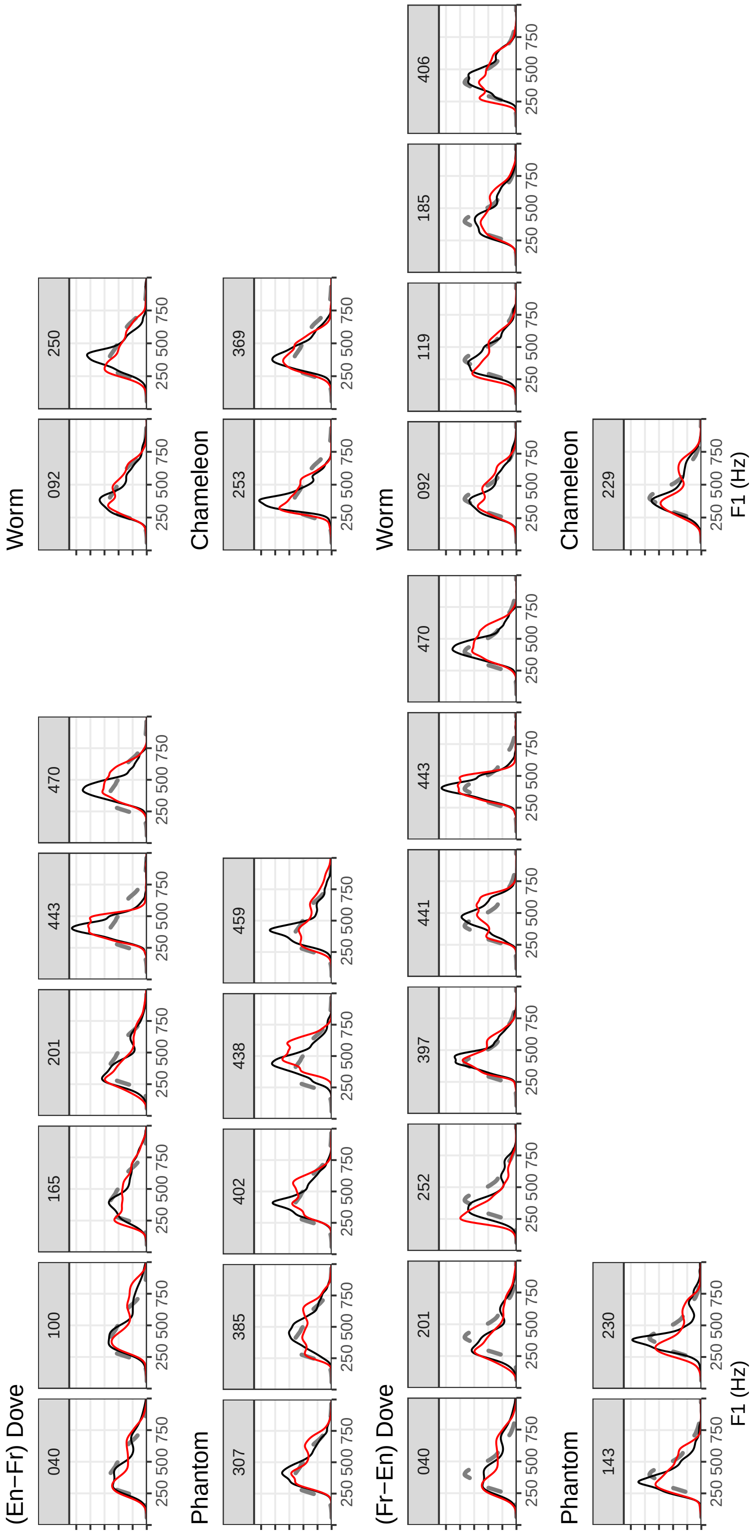

Figure 8.32: Distributions of LTF1 in English (black) and French (red) for all members of relational groups in En–Fr and Fr–En comparisons. Grey dashed line indicates group distribution pooled from all 60 speakers in the KS language.

such as through bimodality (288) or peaks at higher frequencies (106, 452). These patterns can be contrasted with those of the phantoms, many of whom demonstrated similar cross-linguistic shifts. The chief difference between the two groups is that phantom speakers mostly had a much less distinctive LTF2 in the KS language. Notably, for the three speakers (059, 119 and 239) who were identified as phantoms in both sets of comparisons, their LTF2 strongly conformed to the respective aggregate distribution in each language. LTF2 for chameleons similarly showed strong typicality in the KS language. Between English and French, however, chameleons exhibited only small within-speaker differences, such as in the peak location, when compared to the clear differences found among doves and phantoms, such that they could perform relatively strongly in SS comparisons.

A number of speakers had outlying individual performance in both En–Fr and Fr–En comparisons but in different ways. 253 and 420, for example, were both classified as phantom and chameleon in different language pairings. 253, who was a phantom in En–Fr comparisons but a chameleon in Fr–En comparisons, produced very different LTF2 in the two languages, with his distributions showing very little distinctiveness in English but a relatively sharp peak in French. 420, who was classified the other way round, accordingly produced LTF2 that was indistinctive in French but more sharply peaked than the rest of the speakers in English. Also of interest is 470, who was classified a dove in Fr–En comparisons but a chameleon in En–Fr comparisons. The difference in his relative DS performance is likely due to the language-induced shift of the population: Although his distribution in English had a sharper peak than other speakers and could be considered relatively distinctive for Fr–En comparisons, the shift of the reference population to higher frequencies in French meant that his distributions were relatively less distinctive for En–Fr comparisons.

As in the case of LTF2, LTF3 as illustrated in Figure 8.34 showed a consistent upward shift in French for most speakers, albeit on a smaller scale. Across both sets of comparisons, LTF3 for doves all had peaks at particularly high frequencies (e.g., 040, 173) or particularly low frequencies (e.g., 438, 441), although the distributions came in various shapes. In spite of minor shifts, their distributions remained cross-linguistically similar, which may explain why there was much overlap between the sets of doves for En–Fr and Fr–En comparisons.

Members of other relational groups generally demonstrated greater within-speaker differences between English and French, either in the peak location or in the shape of the distribution. The patterning of the distributions from chameleons, all appearing in En–Fr comparisons, was in fact similar to that observed in the corresponding group of doves,

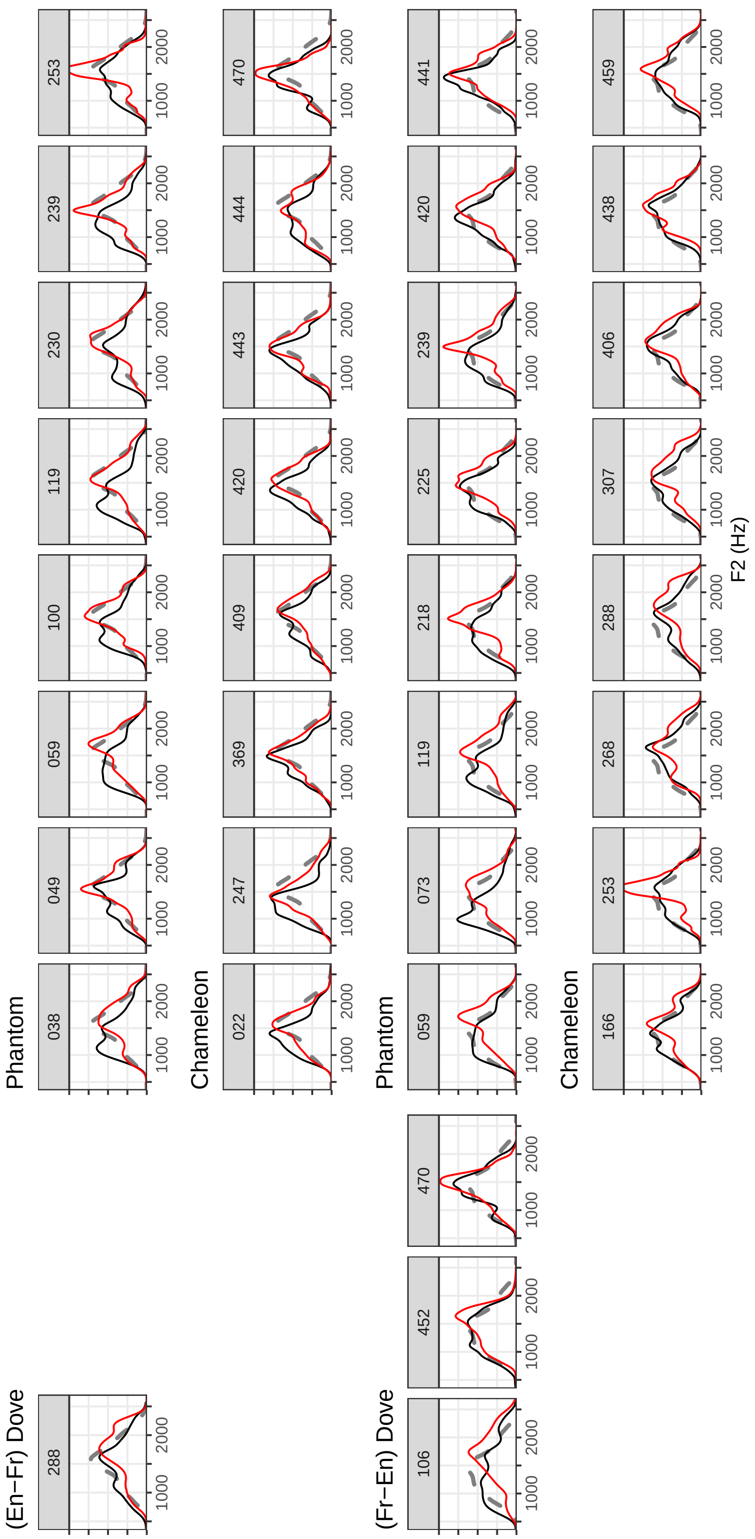

Figure 8.33: Distributions of LTF2 in English (black) and French (red) for all members of relational groups in En–Fr and Fr–En comparisons. Grey dashed line indicates group distribution pooled from all 60 speakers in the KS language.

but their distributions, especially in French, showed neither the extreme peak frequencies nor sharpness found in the doves and were as such comparatively less distinctive. In the case of the phantoms, their distributions did not follow a single clear pattern. For one group of speakers (250, 268 and, to a lesser extent, 201), their LTF3 was highly distinctive in the KS language. Their distributions in the QS language, however, showed clear differences and were much closer to the overall distribution of the reference population. Other phantoms (401, 409 and 420), all in Fr–En comparisons, produced LTF3 with much lower distinctiveness, with their distribution in French somewhat shifted in the direction of the population distribution in English.

Finally, for LTF4, Figure 8.35 shows that there is great acoustic variability across speakers with outlying individual performance. In these systems, doves commonly showed distinctively sharp peaks at either particularly high (e.g., 165, 413) or low (e.g., 236, 406) frequencies. However, the degree of within-speaker similarity across languages did vary. Some individuals showed almost completely overlapping distributions (e.g., 165, 236), although most doves did demonstrate a minor shift to higher frequencies in French, in line with the overall language-specific patterns found in Chapter 5. Their LTF4 distributions nevertheless remained largely similar and retained their distinctive shapes. 253 displayed the strongest differences between English and French. In particular, his distribution of LTF4 showed strong bimodality in both languages, with a peak located at around 2500 Hz and another at around 3700 Hz, but the relative heights of the peaks were exchanged, resulting in distributions that were highly distinctive in both languages.

When compared to the reference population, speakers classified as worms showed a highly typical, very similar distribution in the KS language. For the sole worm in En–Fr comparisons, his distributions showed little cross-linguistic resemblance, with a sharper peak at a lower frequency in English. Worms in Fr–En comparisons, on the other hand, were remarkably consistent across languages, with strongly similar distributions all very close to the population norm in English. As such, their relatively poor SS performance was not simply a result of highly dissimilar distributions within the same speaker, but the absence of any strong evidence due to their low distinctiveness, particularly with respect to the English data. As the norm of the population shifted in French, the same distributions of these speakers could become relatively less indistinctive, which may be the reason the speakers did not perform as poorly in Fr–En comparisons for them to be classified as worms.

In contrast, phantoms all produced an LTF4 distribution in the KS language that could be considered distinctive from the background population in some way, be it a sharp peak (369, 441), a peak at particularly high (106, 470) or low (420) frequencies, or

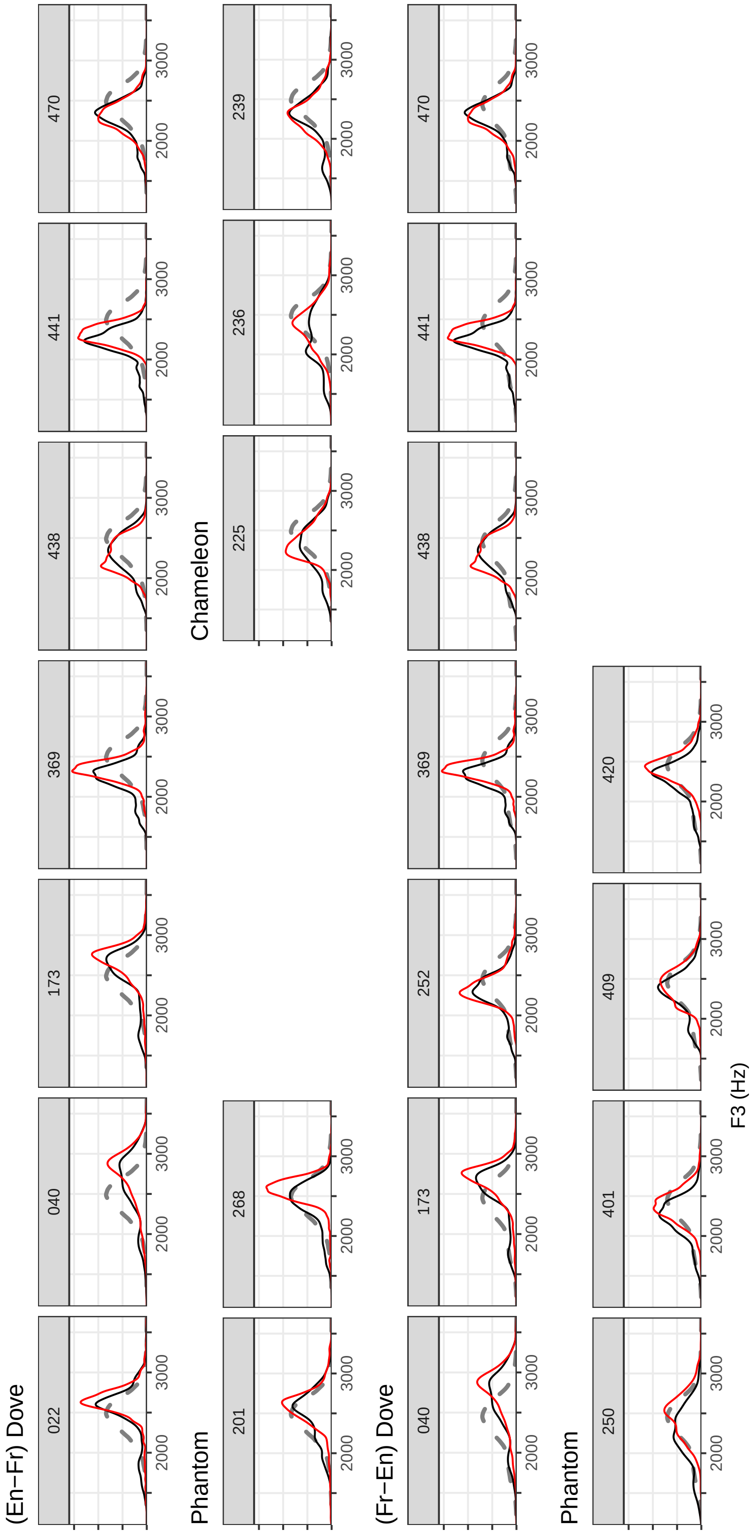

Figure 8.34: Distributions of LTF3 in English (black) and French (red) for all members of relational groups in En–Fr and Fr–En comparisons. Grey dashed line indicates group distribution pooled from all 60 speakers in the KS language.

a highly atypical shape (138). The distribution in the other language (of the QS), on the other hand, all tended to be clearly distinct and much more similar to that of the pooled reference data.

8.2.4 Discussion: Cross-language comparison

The primary issue in Experiment 2 is the impact of language mismatch on the effectiveness of LTFDs as speaker discriminants in cross-language comparisons. A global-level analysis focused on C_{llr} and EER, discussed in greater detail in Section 8.2.1.3, shows that language mismatch indeed dramatically increased both C_{llr} and EER for LTFDs, whether individually or combined in use, such that their discriminatory power was much more restricted, especially in the case of LTF2. In line with previous findings from Becker et al. (2008) and Gold et al. (2013b), combining formant centre frequencies with formant bandwidths or combining multiple formants was found to improve performance considerably, even when the effects of language mismatch were taken into account. Additionally, matching the language of the background data with that of the KS and using training data derived from matching circumstances (i.e., cross- not same-language comparisons) were found to be essential for LTFDs to achieve (relatively) optimal performance. Otherwise, system validity could be severely compromised, rendering any conclusions given on the basis of the evidence insecure.

The zooplot analysis of F-only systems demonstrates that, on an individual level, poorer performance of LTFDs was manifested in a number of ways. In absolute terms, both SS and DS comparisons saw a general shift of mean LLR in the contrary-to-fact direction, with an increase in the number of speakers who could not be matched with themselves on average in SS comparisons. There was also noticeable compression in the range of LLRs produced offered by speakers, particularly in DS comparisons. Overall, when compared with same-language comparisons, LLRs from individual speakers formed less extreme distributions and the evidence could only provide limited to moderate support, even when all four formants were combined.

In relative terms, performance in SS and DS comparisons was no longer strongly correlated as it was in same-language comparisons, as attested by the commonplace occurrence of phantom and chameleon speakers in each tested system. Such weakening was pushed to the extreme in the case of LTF2, when there was a clear *negative* correlation between SS and DS performance and the whole distribution became aligned with the other diagonal, running from the top left to the bottom right of the zooplot, indicating a systemic algorithmic weakness in how speaker-specific information is being captured (Dunstone & Yager, 2009). Further analysis of speakers' relative position in each sys-

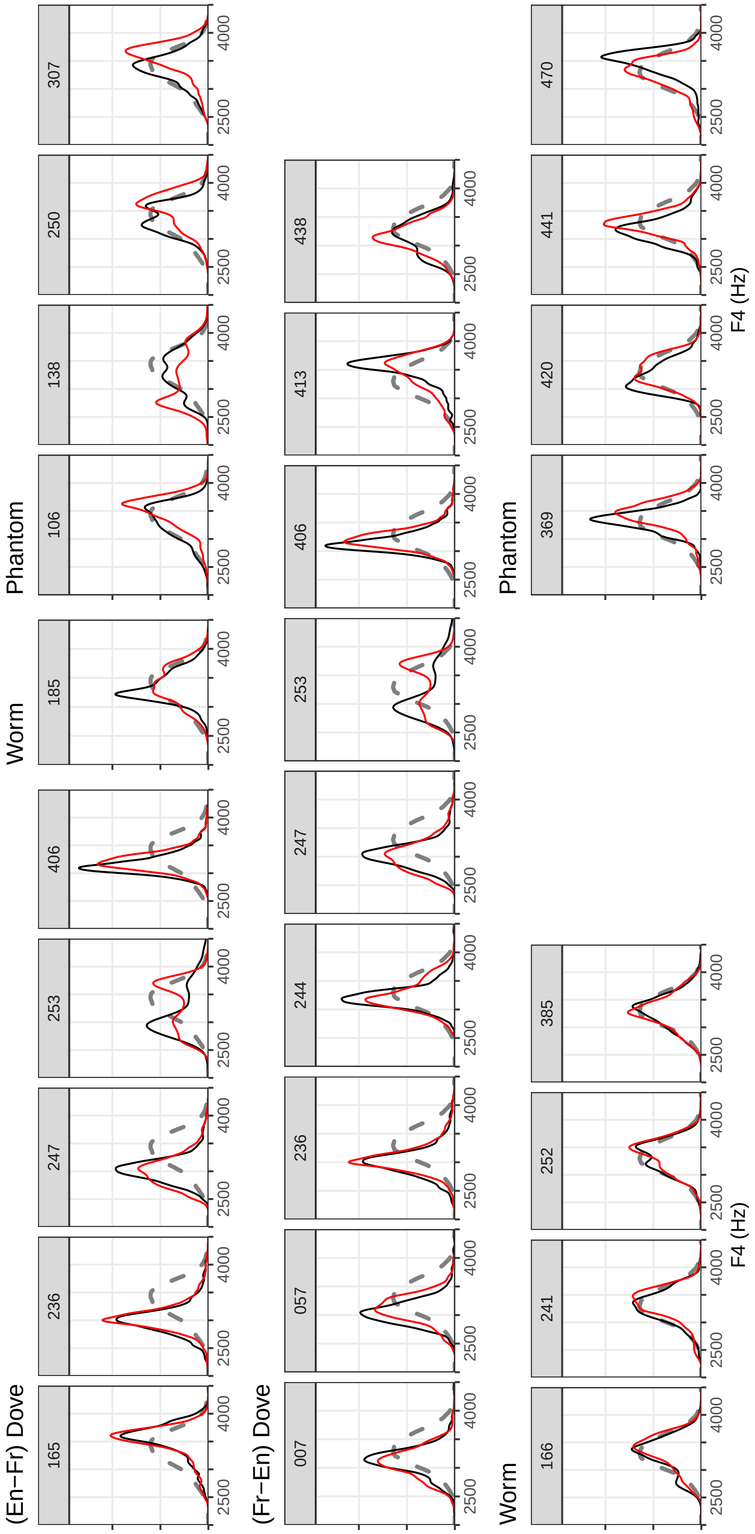

Figure 8.35: Distributions of LTF4 in English (black) and French (red) for all members of relational groups in En–Fr and Fr–En comparisons. Grey dashed line indicates group distribution pooled from all 60 speakers in the KS language.

tem, through the use of "normalised" zooplots, elaborated on these findings, revealing a set of consistent patterns. With the exception of LTF2, across different systems, speakers who performed relatively well (in both SS and DS comparisons) in cross-language comparisons typically also performed well in same-language comparisons, evidenced by the doves who maintained their classification and the short distance travelled by other speakers entering the dove group in cross-language comparisons. These speakers who maintained their level of relative performance were, however, the minority. For other speakers, language mismatch generally impacted individual performance in DS comparisons to a similar degree, as evidenced by the relatively small vertical movements in the plots, although this tendency was weakened in the combined systems. The effect of language mismatch on performance in SS comparisons was less uniform, with speakers in the top half of the normalised zooplots tending to shift towards the left while those in the bottom half typically moving to the right. In other words, speakers with stronger DS performance were more adversely impacted in their SS comparisons than those with weaker DS performance, to the extent that the former often replaced the latter as the worst performers in SS comparisons. Such deterioration was particularly marked for speakers who were doves in same-language comparisons but phantoms in cross-language comparisons. Nevertheless, the relative improvement for speakers performing relatively poorly in same-language comparisons was limited, as they typically remained in the lower left quarter of the zooplot. For LTF2, the worst-performing individual LTF, substantial movement away from the doves (and consequently worms) indicates that speakers who performed well in same-language comparisons were particularly affected by language mismatch in SS and/or DS comparisons. As such, unlike other systems where a small number of speakers could still perform consistently well, almost no speakers were able to maintain good relative performance in the case of LTF2.

In Section 8.1.4, it was shown that the the acoustic characteristics of speakers' individual LTFDs were highly predictive of their level of individual performance in same-language comparisons. The exploration above in Section 8.2.3 demonstrates that while a similar relationship between acoustic and individual performance can be found in cross-language comparisons, the language-specificity of LTFDs entails a more complex interplay between the individual and the population.

In line with findings from Section 8.1.4, speakers who maintained relatively good performance (doves) were indeed acoustically distinctive. Perhaps somewhat surprisingly, their strong performance in SS comparisons was not necessarily rooted in maintaining the same LTFDs in different languages and not participating in cross-linguistic shifts. In fact, for most LTFs, doves did show clear signs of language-specificity in line with the other speakers. Instead, they performed relatively well because, in their QS, they maintained or even strengthened their distinctiveness in relation to the reference population, modelled in the KS language, such as by shifting the peak location from an already distinctively high frequency in the KS to an even higher frequency in the QS. By contrast, speakers with the worst performance (worms) in cross-language comparisons were highly uncharacteristic, at least in the KS language, but did not always show high within-speaker variability across languages. Although there were only very few worms in these cross-language comparisons for strong conclusions to be drawn, worms were frequently among speakers whose distributions were the most similar between English and French, especially in the case of LTF4, even more so than doves and phantoms in the same set of comparisons. The proximity of their distributions in both languages to the overall population norm means that, no matter whether they were acting as the questioned or known speaker, no strong support for either the prosecutor's or the defence's proposition could be obtained, and the resultant strength of evidence would be weak for both SS and DS comparisons.

The complex relationship between the individual and the group is especially prominent when phantoms and chameleons in cross-language comparisons are considered. Much like the doves, across most LTFs, phantoms, who similarly performed relatively well in DS comparisons, had a distinctive distribution in the KS language. Unlike the doves, their distribution in the other language shifted towards, rather than away from, the reference population, which was also modelled on the KS language. As such, performance in SS comparisons of phantoms was not necessarily due to them being disadvantaged by higher intra-speaker variability, but by their relatively indistinctive QS that led to weaker evidence. At the same time, their relatively distinctive distributions in the KS contributed to relatively strong evidence when compared against QS from other speakers. Chameleons, on the other hand, generally did not have distinctive LTFDs in the KS language. While their LTFDs in the other language were largely similar, they could also differ slightly in a direction away from the norm of the reference population. Such a combination, diametrically opposite to that displayed in phantoms, had the effect of rendering their performance in DS comparisons relatively weak while maintaining relatively good performance in SS comparisons.

The findings thus far show that classifications in cross-language comparisons, particularly of phantoms and chameleons, rested on how their QS were positioned with respect to not only their KS but also the reference population in the KS language. What this means, however, is that speaker idiosyncrasy in the form of their observance of cross-linguistic shifts cannot be adequately captured. Speakers who followed the general language-specific patterns to a similar degree may end up with different levels of performance depending on where they were in relation to the reference population. This can be illustrated by speakers 040 and 250, who were respectively identified as a dove and a phantom in Fr–En comparisons with LTF3 (Figure 8.34). Both speakers had a relatively flat distribution in English and a distribution in French with a higher and sharper peak. The difference between the two is that LTF3 in both languages had peaks at particularly high frequencies for 040, but at relatively low frequencies for 250, such that his French distribution converged with the pooled English distribution. At the same time, speakers who diverged from the language-specific shifts could also end up with similar performance as others who did not. This is particularly well exemplified by phantoms for LTF4. In Fr–En comparisons, LTF4 for both 420 and 441 had a below-average peak in English. Their distribution in French moved nearer the population norm as a result of the (small) shift towards higher frequencies shared by most speakers. 470, also a phantom in Fr–En comparisons, had an above-average peak in English, but was one of the few speakers whose French distribution was lower than that in English. Although the direction of his shift could be considered atypical, the relatively unremarkable destination of the move nonetheless resulted in his phantom-like performance.

In addition to diagnosing individuals with outlying performance, the acoustic analysis may further shed light on the particularly poor performance of LTF2 overall in cross-language comparisons. As Figure 8.33 shows, even among speakers with relatively outlying performance, there is limited variability in their LTF2 distributions, particularly in the location of the peak. Out of all four LTFs, the cross-linguistic distance between English and French was shown to be the largest for LTF2 (Section 5.2), and speakers in Figure 8.33 strongly conformed to the overall distributions in both languages. Speakers classified as phantoms in both En–Fr and Fr–En comparisons (059, 119 and 239) exemplified this most strongly. Chameleons behaved similarly in this regard, but their distributions demonstrated a tendency to be closer together than phantoms. The great distance between English and French, together with the general lack of variability in either language, means that performance in either SS or DS comparisons would be weak. As a speaker's distribution in one language deviated from the reference population and moved closer to the other language, as it did for chameleons, it became more probable for evidence from any QS in the other language to be derived from that individual distribution than from any other speaker in the reference population. In other words, that speaker would become more easily matched to not just himself but also any other speaker. As a result, performance in SS comparisons would be improved, while performance in DS comparisons would further deteriorate, culminating in a rightward and downward movement in the zooplot. Contrary to results reported in Heeren et al. (2014), the findings here illustrate that, under the dominating influence of language-specificity in LTF2, within-speaker variability of this particular formant is no less than between-speaker variability, and as such provide support for the argument that F2, at least when measured in the long term, is primarily responsible for encoding linguistic information and shows relatively little potential for speaker discrimination, particular in the context of language mismatch. This of course does not preclude F2 from possessing stronger discriminatory power for specific vowels, such as in the case of filled pauses (Hughes et al., 2016). Recent research into the production of filled pauses in bilinguals provides emerging yet consistent evidence that while the phonetic realisation of filled pauses is similarly susceptible to cross-linguistic influences, there is considerable variability in how speakers shift in response to a language switch (de Boer & Heeren, 2020; García-Amaya & Lang, 2020; Lo, 2020). Whether such a source of speaker-specific information can be effectively harnessed in a forensic setting remains an empirical question for investigation.

Chapter 9

Results: ASR

In this chapter, results from system testing using the ASR software *Phonexia Voice Inspector* (Phonexia) are reported. Findings from both global- and individual-level analysis in Experiment 1, comparing system performance in English and in French, are first presented and discussed in Section 9.1. Results from testing of ASR performance in cross-language comparisons in Experiment 2 then follow in Section 9.2.

9.1 Experiment 1: Same-language comparisons

9.1.1 Global metrics

C_{llr} from ASR testing in English and French is displayed in Figure 9.1. EER is not illustrated, since performance as measured by this metric is at ceiling level (0%) in both languages in all replications. Strong performance of the ASR system is similarly evident from the extremely low C_{llr} obtained in both languages. Mean C_{llr} over 100 replications was 0.0047 in English and 0.012 in French. This difference between English and French was found to be significant, as shown by the significant main effect of language in LMEMs fitted to the C_{llr} obtained ($\chi^2(1) = 90.94,\ p < .0001$).

9.1.2 Individual-level analysis

Displayed in Figure 9.2, the zooplots illustrate the individual LR performance of all 60 speakers in English and French and show strong similarities across languages. In both languages, all speakers produced positive mean SS-LLR and negative mean DS-LLR, suggesting that on average these speakers matched well with themselves but were distinguished from other speakers. The overall ranges of mean LLRs were also similar between

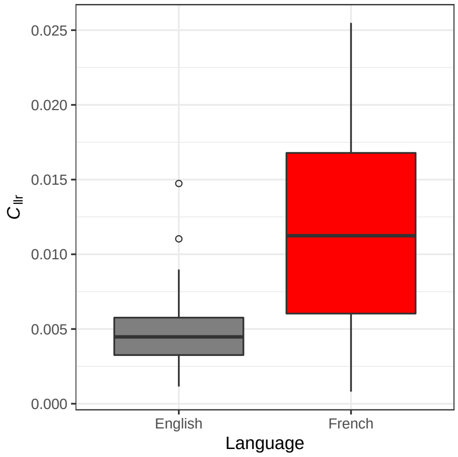

Figure 9.1: C_{llr} from system testing in Phonexia in English (grey) and French (red).

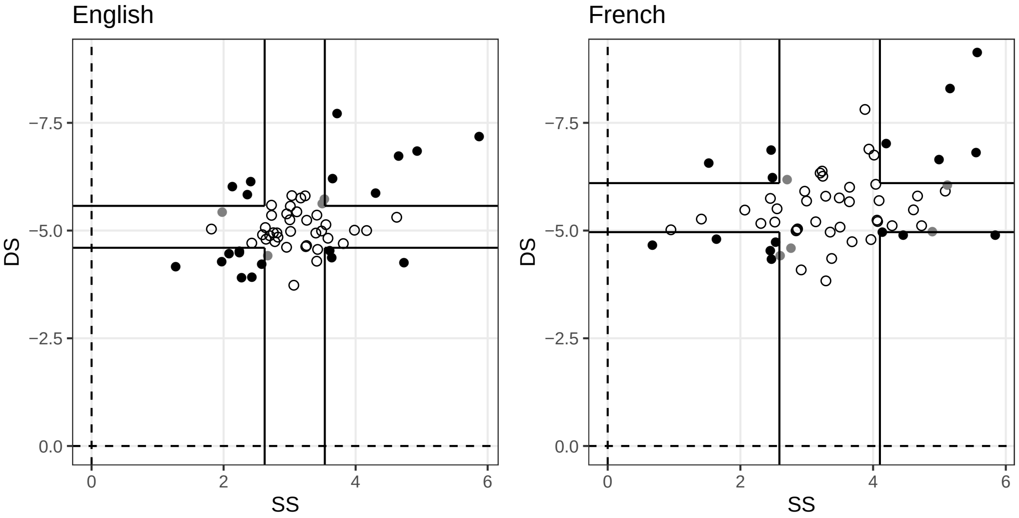

Figure 9.2: Zooplots for systems based on Phonexia in English (left) and French (right).

English and French, with mean SS-LLR between 0 and 6 and mean DS-LLR generally between -3 and -8. Some cross-linguistic differences can nevertheless be discerned in the distribution of speakers. Individual SS-LLR and DS-LLR tended to be of higher magnitude in French than in English, as evident from the upward and rightward shift of the 25th and 75th percentiles demarcating the boundaries of the relational groups. At the same time, the wider spacing between the lines supports the observation that the distribution of speakers in French was more dispersed, with greater separation between speakers, than in English, where speakers were noticeably more clustered towards the centre of the plot.

None of the four relational groups were empty, but there was also not a significant

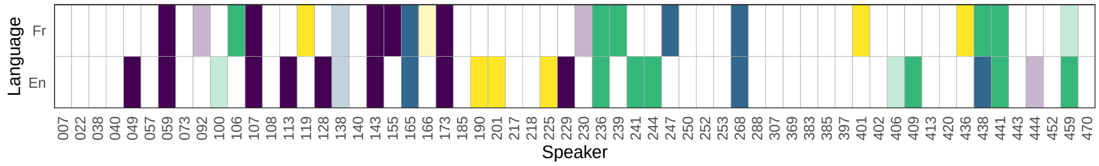

Figure 9.3: Summary of speakers classified in Phonexia as doves (green), worms (purple), phantoms (blue) or chameleons (yellow) in each language (near-members of each group in lighter shade).

presence of any of the groups. Six dove and eight worm speakers were found in English, while five members were identified in French for both groups. The size of the phantom and chameleon groups was smaller than that of the doves and worms, consisting of three members each in both English and French. Animal group classification of speakers, as summarised in Figure 9.3, does not show overall strong correspondence between the two languages. Only two of the doves (236 and 441, alongside 459 who was a dove in English and near-dove in French) and two of the phantoms (165 and 268, as well as one near-phantom 138) were shared between English and French. Five speakers identified as worms were classified as such in both English and French, the only group to have a majority of speakers commonly identified in both languages, while the two sets of chameleons did not show any intersection at all. Only one speaker was classified as a member of different groups in the two languages (438, who was a dove in French but a phantom in English), but no speakers who were among best-performing doves in one language were classified as one of the worst-performing worms in the other.

9.1.3 Discussion: Same-language comparisons

The findings here show that, when faced with high-quality materials with no channel or language mismatch, state-of-the-art ASR software unsurprisingly demonstrated high power of speaker discrimination. Indeed, ceiling performance in EER and extremely low C_{llr} were not limited to English but also found for French. The same level of calibration performance, however, was not reached in different languages. While EER for both English and French was at 0%, indicating that speaker classification was not at issue, differences in C_{llr} suggest that there is nonetheless some discrepancy in the strength of evidence offered, with better calibration achieved in English than in French. Individual-level analysis further demonstrates that the relatively weaker performance of French was not due to a wholesale deterioration of the strength of evidence offered by the population, as speakers in both languages occupied a similar “LLR space”. Contrary to the trend in C_{llr}, French even reported overall stronger DS-LLR for individual speakers than English.

As the range of SS-LLR in French widened in both directions, the (slightly) worse performance of French thus appears to be mainly stemming from the few speakers with the weakest performance in SS comparisons. Indeed, zooplot analysis found generally stronger correspondence between the two languages in worm and phantom membership than in the dove and chameleons groups, suggesting that the same speakers were more consistent in providing relatively weaker performance in SS comparisons. While it is difficult to elucidate the reason behind the cross-linguistic differences in performance, as only limited information about the system is available, it is speculated here that the architecture of the software and the data that went in to train the system may have played a part. Although this version of Phonexia, unlike a previous version, does not rely on a pretrained GMM in the step of feature extraction, it nevertheless makes use of a trained acoustic model to carry out comparison via PLDA (Jessen et al., 2019). If English data accounted for a greater proportion of the data that had been selected to train this model than data in other languages, such as French, discrepancy in its performance in favour of English could ensue, leading to the current set of results.

9.2 Experiment 2: Cross-language comparisons

9.2.1 Global metrics

9.2.1.1 Language mismatch

Figure 9.4 presents C_{llr} and EER obtained from carrying out cross-language comparisons in Phonexia. When compared to same-language comparisons in either language, both En–Fr and Fr–En comparisons yielded C_{llr} that was higher and occupied a wider range, but was nevertheless very low in absolute value. Mean C_{llr} was 0.040 when English QS were compared with French KS, and slightly lower at 0.032 when French QS were compared with English KS. EER in cross-language comparisons did not reach overall ceiling performance but was still very low, at 0.25% for En–Fr comparisons and 0.09% for Fr–En comparisons. Out of all 100 replications, 31 produced an EER of 0% for En–Fr comparisons and 54 produced the same for Fr–En comparisons. Therefore, even in cases of language mismatch, systems still performed at strong levels and remained well calibrated.

9.2.1.2 Training mismatch

The effect of training mismatch on C_{llr} is illustrated in Figure 9.5, which compares C_{llr} from systems calibrated using training scores from English or French same-language

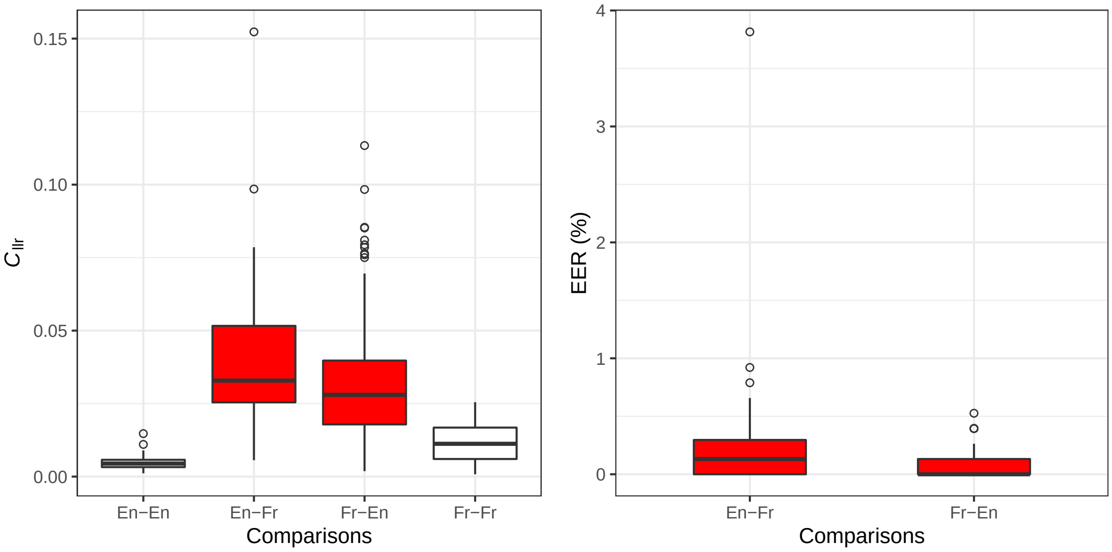

Figure 9.4: C_{llr} (left) and EER (right) from system testing of cross-language comparisons (En–Fr and Fr–En) in Phonexia with reference to same-language comparisons in English (En–En) and French (Fr–Fr).

comparisons with those calibrated using scores from matching cross-language comparisons. Training mismatch had a considerable impact on C_{llr}, especially in En–Fr comparisons, where mean C_{llr} reached 0.80 when calibrated using Fr–Fr training scores and exceeded 1 when calibrated using En–En scores (1.06). Fr–En comparisons produced generally lower C_{llr}, with only a small number of replications producing $C_{llr} > 1$. In this case, higher C_{llr} was instead obtained from calibrating using Fr–Fr scores (0.54) rather than when using En–En scores (0.44). Model comparison found a significant effect of calibration condition for both En–Fr ($\chi^2(2) = 441.63,\ p < .0001$) and Fr–En ($\chi^2(2) = 338.60,\ p < .0001$) comparisons, and post-hoc comparisons found the differences between calibration using En–En scores and Fr–Fr scores to be significant in both cases ($p < .0001$).

9.2.2 Individual-level analysis

Figure 9.6 shows the zooplots for En–Fr and Fr–En cross-language comparisons. As per Sections 7.2.2 and 8.2.2, only systems without training mismatch were analysed. Both plots depict a generally even distribution of speakers with little clustering, indicating that speakers did not produce highly similar levels of performance but were generally well separated in terms of level of performance. In both cases, the separation of speakers was primarily along the SS dimension, rather than the DS dimension. Mean SS-LLR spanned from near 0 (but still positive) up to approximately 8 in both En–Fr and Fr–En

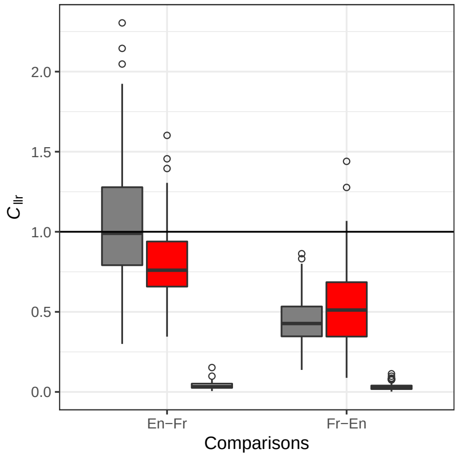

Figure 9.5: C_{llr} from all systems of cross-language comparisons tested in Phonexia, calibrated using En−En scores (grey), Fr−Fr scores (red) and matching scores (white).

comparisons, while mean DS-LLR was not weaker than −3 for any speaker. Notably, when compared with the same-language comparisons in Section 9.1.2, the magnitude of mean DS-LLR increased overall, while mean SS-LLR extended in both directions, with the best-performing speakers having higher SS-LLR and the worst-performing speakers producing SS-LLR much closer to 0. The 25th and 75th percentiles demarcating the boundaries of relational groups further attest to the relatively narrow range of DS-LLR, as half of the speakers produced mean DS-LLR between −5.43 and −6.75 in En−Fr comparisons and between −5.41 and −6.63 in Fr−En comparisons, whereas the corresponding SS-LLR range was 2.69–5.23 and 3.04–4.89 in En−Fr and Fr−En comparisons respectively. While the speaker distribution was broadly similar across the two sets of comparisons, the outlying relational groups showed some minor differences. None of the relational groups were empty in either set of comparisons, but whereas the phantoms and chameleons constituted the largest groups in En−Fr comparisons (with five and four speakers respectively), in Fr−En comparisons there were only three phantoms and one chameleon, making the doves (four speakers) and worms (six speakers) the largest groups.

A normalised version of the zooplots, constructed as per Section 7.2.2, is presented in Figure 9.7 to compare individual performance in same- and cross-language comparisons. Particularly notable in both plots is the absence of systematic patterns in the movement of speakers between same- and cross-language comparisons. Only a small number of speakers remained stable in their relative position, while most speakers displayed considerable movement along either SS or DS dimension, or indeed both. Greater movement was more commonly found along the SS dimension, as evidenced by the numerous

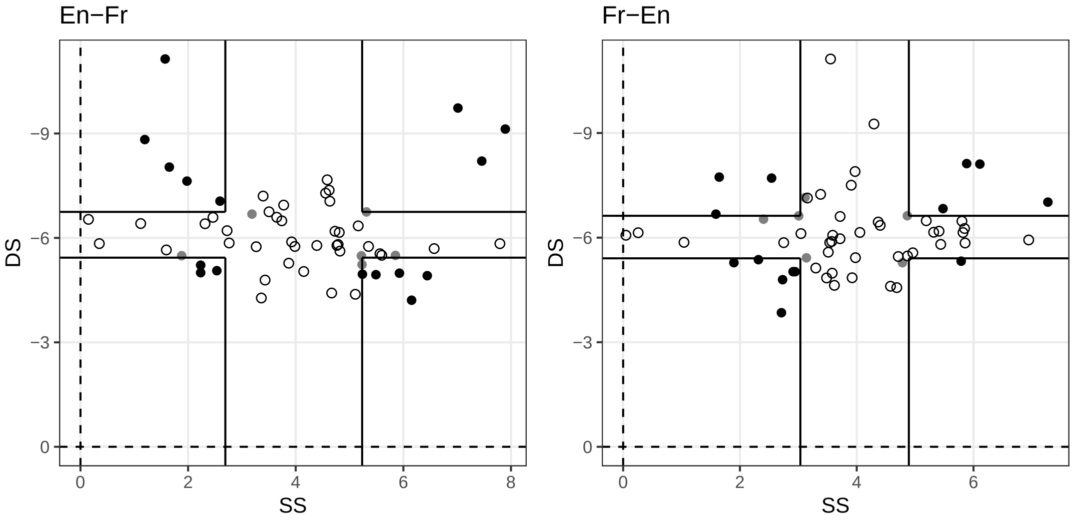

Figure 9.6: Zooplots for systems based on Phonexia in cross-language En–Fr (left) and Fr–En (right) comparisons.

speakers who were situated near the left edge of the plot for one set of comparisons but near the right edge for the other set, and vice versa, but there was also much vertical movement exhibited by many speakers. When KS language was maintained and QS language was varied, outlying speakers only rarely stayed in the same relational group for both same- and cross-language comparisons. Indeed, speakers exhibiting dove-like behaviour in same-language comparisons often shifted to the top left corner and became more phantom-like in cross-language comparisons, while others maintained their performance in SS comparisons but were more adversely affected in DS comparisons, resulting in a downward move to become more chameleon-like. With individual exceptions, speakers classified as worms in same-language comparisons typically exhibited less movement than those found in other relational groups, suggesting that their relatively poor performance was maintained regardless of language mismatch.

9.2.3 Discussion: Cross-language comparisons

The current findings show that, in line with findings from Künzel (2013), language mismatch does have a quantitative impact on the performance of ASR systems. In both En–Fr and Fr–En comparisons, both C_{llr} and EER indicated a level of performance that was indeed not far from ceiling performance. The high quality of the recordings, as well as the lack of any channel mismatch, is likely to have contributed, at least in part, to the excellent performance here. Interestingly, the individual-level analysis shows that cross-language comparisons led not to a general reduction in the speakers' performance in

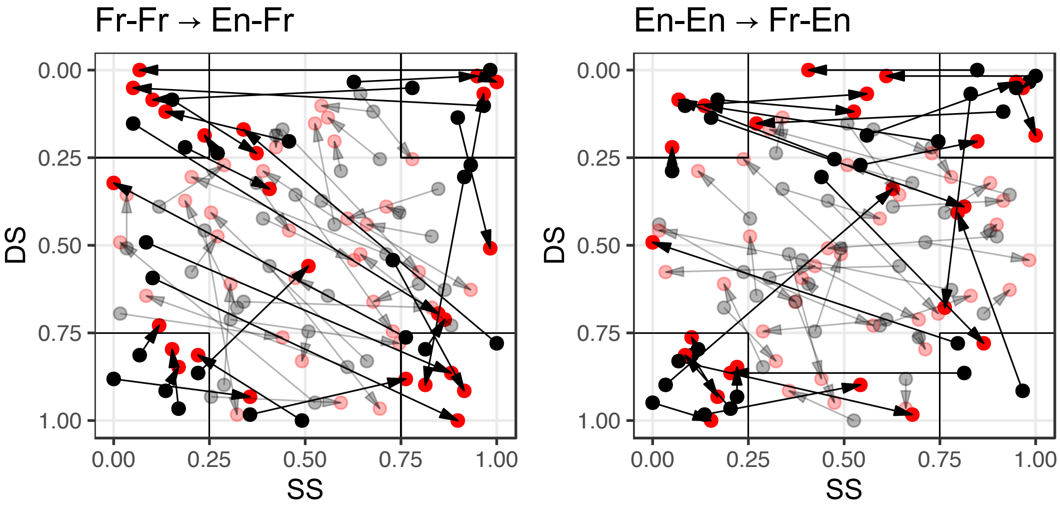

Figure 9.7: Normalised zooplots for systems based on Phonexia in same-language (black) and cross-language (red) comparisons. Arrows connect speakers from same-language to cross-language comparisons.

SS comparisons, but rather a wider range of their SS-LLR, meaning that some speakers might then actually produce even better performance than when there was no language mismatch. A comparative analysis of the speakers' relative position on the zooplots in same- and cross-language comparisons further showed a general lack of correspondence of their individual performance in the two sets of comparisons, where only speakers with the worst performance in same-language comparisons continued to also perform relatively poorly in cross-language comparisons. Even though performance from MFCCs, at least in the way they are utilised in Phonexia, from English and French was highly comparable, there was considerable variation in the extent to which speakers' individual performance was affected by language mismatch. The instability of the speakers' relative performance suggests that, although MFCCs are capable of providing excellent discriminatory performance across languages under the conditions of the present experiment, they are nevertheless not fully immune to the effects of language mismatch. In particular, the relatively worse overall performance in cross-language comparisons did not appear to be tied to the a specific portion of speakers who performed poorly in both types of comparisons. As Figures 9.6 and 9.7 show, speakers who produced the lowest mean SS-LLR in cross-language comparisons were rarely those who did so in same-language comparisons, but were in fact commonly speakers who produced relatively good SS performance in same-language comparisons.

Nevertheless, the results from Section 9.2.1.2 show that calibration with appropriate training scores reflecting the case-specific conditions (i.e., cross-language comparisons) is necessary to achieve the high level of performance. If the analyst simply performed calibration using reference data derived from a single language (e.g., due to lack of access to a bilingual database), then system validity and reliability would deteriorate substantially, as evidenced by the values and ranges of C_{llr}, to the extent that the system may become poorly calibrated, depending on the composition of the sample speaker sets. This would then lead to a conclusion with a misestimated strength of evidence. In line with the recommendation from Morrison (2018) and Morrison et al. (2021), these findings thus underscore the significance of using training data that are representative of the conditions of the case when ASR systems are used.

Chapter 10

General Discussion & Conclusion

This chapter returns to address the main research questions (RQs) outlined in Section 2.4, in light of the results from Chapters 7 to 9. Section 10.1 unites results from Experiment 1 across the three sets of features and considers RQ1, while Section 10.2 goes on to discuss RQ2 and RQ2.1 in light of the results from Experiment 2. The implications of the findings from the current study for FVC are then discussed in 10.3. Directions for future research are considered in Section 10.4, before Section 10.5 concludes the thesis.

10.1 Effect of language

The first RQ that the current study seeks to address is the effect of language on a feature's discriminatory potential in FVC, which was examined in Experiment 1 by using the same bilingual speakers in both languages to isolate the factor of language.

Results from system testing of /s/ spectral moments (Section 7.1) show that the same discriminatory power was not maintained between English and French. Across different individual spectral moments and modes of input tested, EER and C_{llr} were consistently lower in French than in English. As more spectral moments were combined, results showed that EER and C_{llr} remained lower for French than for English. These findings indicate that, overall, /s/ in French produces stronger discriminatory performance on the global level than /s/ in English. Individual-level analysis using zooplots (Section 7.1.3) confirms that stronger performance is not limited to specific outliers but holds more or less generally for all speakers.

Results from system testing of LTFDs (Section 8.1), on the other hand, show that their discriminatory power was generally similar in English and French, but systems showed slightly different patterns depending on whether formant bandwidths were included. In F-only systems of individual LTFs, where no formant bandwidths were included, there were only small cross-linguistic differences in both C_{llr} and EER for LTF2, where stronger performance was found in English. No significant differences between English and French were found for the other LTFs, except for LTF1, where EER was lower in French. The same pattern was maintained when multiple formants were included, where no significant differences in EER or C_{llr} were found between the two languages. Zooplot analysis of F-only systems (Section 8.1.3) shows that cross-linguistic stability of discriminatory performance was maintained on the individual level, with speakers showing highly similar distributions of LLRs for each individual LTF as well as for the combination of all four LTFs. In F+BW systems, lower EER and C_{llr} were generally found in French than in English, although some variation was observed when only individual LTFs were included, with LTF2 and LTF4 producing lower C_{llr} and LTF1 and LTF4 yielding higher EER in French than in English. The differences between English and French F+BW systems were nevertheless small: When all four LTFs were included, C_{llr} was 0.19 and 0.25 in French and English respectively, while EER was 4.3% and 6.2% respectively in each language.

Testing of the ASR software *Phonexia Voice Inspector* shows only very minor effects of language. On the global level (Section 9.2), EER was at ceiling performance regardless of language, suggesting that there are no issues with errors in discrimination. C_{llr} was similarly extremely low for both languages, although C_{llr} was significantly higher in French than in English, suggesting that the system is better calibrated for English materials. Individual-level analysis (Section 9.2.2) shows that the main difference between English and French appears to lie in SS comparisons, where a small number of speakers produced relatively weaker performance, in the form of smaller but still positive mean SS-LLR. Overall, these results suggest that, when tested with high-quality materials, the performance of ASR is cross-linguistically highly similar.

The summary of the findings above suggests that the effect of language on the discriminatory potential of features used in FVC is highly dependent on the choice of feature itself. For ASR, at least for the architecture of this specific system, the effect of language is minimal. Of the two linguistic-phonetic variables, LTFDs can be considered to show more language-independent speaker-specificity, whereas the discriminatory power of /s/ is more language-dependent.

When considered in light of results from the acoustic analysis in Chapter 5, it is argued that the differential effect of language on /s/ and LTFDs can at least in part be linguistically motivated. Although the bilingual speakers in the current study largely did not make acoustic distinctions between English /s/ and French /s/ (with only minor differences in the shape of the CoG and skewness trajectories), they were found to exhibit greater within-speaker variability in English /s/ than in French /s/, whereas the range of between-speaker variation was cross-linguistically consistent. The lower within-speaker variability in French thus enhanced the performance of /s/ as a speaker discriminant. LTFDs, on the other hand, showed consistent cross-linguistic differences in their central tendencies, except in the case of LTF1, but showed only relatively small cross-linguistic differences in their shapes. As such, language-specificity in the realisation of LTFDs did not have a substantial effect on their speaker-specificity. As argued in Section 8.1.2, the slightly stronger performance of F+BW systems in French than in English may have been due to the addition of the degree of nasality in nasal vowels as a source of speaker individuality, but further research is needed to answer this.

In the current study, the French materials were generally longer and contained more instances of /s/ and vowels than the corresponding English materials. While it is possible that differences in the amount of materials also contributed to the results here, a clear advantage of French over English was only found for /s/, but not for the other variables. Any differences in the current results that may be attributed to the amount of materials are thus likely to be small. Hughes and Foulkes (2015b), which tested the stability of systems using English /u:/ for different numbers of tokens per reference speaker, found that while LRs were highly unstable and C_{llr} was poor when there were only very few tokens, LRs and system validity reached a level of stability when seven or more tokens per speaker were included, with the addition of more tokens bringing only marginal improvements. In the current study, the variable that was tested using the same approach (/s/) contained an average of 36 tokens per speaker in English and 43 tokens per speaker in French. In a similar vein, the duration of vocalic materials used for testing the performance of LTFDs (26.2s in English vs. 34.6s in French) is similar to previous studies that have adopted the same approach (Becker et al., 2008). It is therefore considered that any effects due to differences in number of tokens would be minimal, even after taking into account the fact that recordings were divided in half to form the QS and the KS in system testing, and that the main effect in operation here is that of language.

10.2 Language mismatch

The second, and arguably the main, research aim of the current study is the effect of language mismatch between the QS and the KS on discriminatory performance in cases of cross-language comparison.

On the global level, a strong effect of language mismatch was found for both linguistic-phonetic variables, leading to much higher EER and C_{llr}. For systems using individual

LTFs or spectral moments, C_{llr} reached close to or above 1, indicating that there can only be very limited speaker-specific information to be found cross-linguistically within individual parameters. While the combinations of all spectral moments or LTFs were similarly affected by language mismatch, these systems nevertheless yielded C_{llr} that was well below 1 and relatively low EER of 10–15%. In the current high-quality conditions, then, these combined systems could retain some speaker-specific information. It remains to be investigated, however, whether this is still the case when these systems are tested in forensically realistic conditions with poorer quality and/or channel mismatch.

In contrast with the findings for ASR, where both C_{llr} and EER only increased very slightly, the results for /s/ and LTFDs indicate that, in isolation, the viability of these features in cross-language comparisons may be largely reduced. An additional factor that likely has contributed to the distance between same- and cross-language comparisons in the current findings is the use of contemporaneous materials in same-language comparisons. As the QS and the KS being tested in same-language comparisons were derived from the same audio, while the QS and the KS in cross-language comparisons necessarily came from different recordings, the range of within-speaker variability represented in same-speaker comparisons may, independently of the factor of language, be smaller than that in cross-language comparisons. Future investigations conducted with non-contemporaneous materials may help tease apart the contributions of these factors.

Individual-level analysis shows that, in the cases of /s/ and LTFDs, performance in cross-language comparisons deteriorated in a similar fashion. For both sets of variables, performance in both SS and DS comparisons was weakened, with SS comparisons particularly causing issues for a number of weakly performing speakers. In both cases, performance in SS comparisons was more variably affected among speakers than DS comparisons, especially among speakers who performed exceptionally well in same-language comparisons. Only a small selection of speakers managed to maintain relatively strong performance in cross-language comparisons, while others experienced particularly adverse affects on their performance in SS comparisons. However, for ASR, which experienced only a minor drop in performance in cases of language mismatch, there was no general weakening of individual performance. Instead, speakers' performance in DS comparisons generally improved in cross-language comparisons, producing stronger mean LLRs. Worse performance in SS comparisons was also far from a uniform effect, when the range of mean SS-LLR expanded in both directions, such that a number of speakers produced stronger SS performance in cross-language comparisons than in same-language comparisons. The overall slightly weaker performance in cross-language comparisons may then be possibly attributed to the few speakers with relatively weaker

SS performance.

Even though individual performance deteriorated in similar ways for /s/ and LTFDs, individual-level analysis demonstrates that such degradation appears to have arisen from different sets of individual behaviour. In the case of /s/, which was not cross-linguistically distinct (at least at the midpoint, for which individual-level analysis was conducted), speakers with exceptionally good performance in both SS and DS comparisons were commonly those with highly distinctive spectral moments, while those who show distinctive but cross-linguistically distinct patterns performed more poorly in SS comparisons. Speakers with highly uncharacteristic spectral moments in both languages generally performed worse. In the case of LTFDs, which displayed stronger language-specificity, the relationship between individual performance and their underlying acoustic behaviour proves to be much more complex. Most notably, cross-linguistic shifts within an individual would not be necessarily detrimental to their individual performance, while stability across languages also would not guarantee relatively strong performance in either SS or DS comparisons. In fact, speakers who performed particularly well in both types of comparisons did show clear signs of participating in cross-linguistic shifts of LTFDs, but maintained their distinctiveness in their QS when compared to the KS language. Speakers who produced the worst performance, on the other hand, often showed little change in LTFDs across languages while other speakers shifted. Their weak performance in not only DS comparisons but also SS comparisons, then, stemmed from the typicality of their distributions in both languages when compared to the reference norm in the KS language. The analysis in this study shows that the relationship between acoustic and individual performance, and hence the assessment of typicality, in cases of language mismatch, is by no means straightforward, and demands detailed knowledge of individual and group distributions in both languages.

For the linguistic-phonetic variables, the practical impact of matching the conditions (in the present case, language) of the background data with the QS instead of the KS was also assessed. In the case of /s/, the quantitative impact of such a mismatch was relatively small, with little to no difference from the condition where the language of the background data was (logically correctly) matched with the KS, although there was a more pronounced discrepancy when all spectral moments were combined in En–Fr dynamic systems. In the case of LTFDs, the impact of this factor was also not entirely uniform, but results show that C_{llr} and EER generally suffered as a result of matching the language of the background data with the QS instead of the KS. While the results from /s/ indicate that such impact may possibly be mitigated by strong similarities of the conditions in the QS and the KS, manifested here by the lack of cross-linguistic distinctions in its spectral moments, the current findings overall illustrate the detriment of not adhering to the position that the conditions of the reference materials be matched with those of the KS, and not the QS (Morrison, 2018; Morrison et al., 2021).

The findings above are further contingent upon calibration using appropriate training data that match with the circumstances of the case. For all the variables tested, when training data were derived from same-language comparisons instead of cross-language comparisons, system performance as measured by C_{llr} suffered in two ways. On the one hand, C_{llr} typically increased to near or above 1, meaning that system validity was overall poor. On the other hand, the range of C_{llr} obtained over 100 replications considerably expanded, with many systems reporting extreme values of maximum C_{llr}. Training mismatch thus induced much higher instability in the C_{llr} and generally lowered system reliability. These findings can have severe implications for conducting cross-language comparisons in the LR framework if the analyst does not have access to suitable bilingual reference materials.

Overall, findings from Experiment 2 suggest that there is a clear and consistent impact of language mismatch between the QS and the KS on system performance. Its impact on ASR performance is the smallest, where EER and C_{llr} remained close to 0, thus providing some support for the claim in Künzel (2013) that low-level acoustic features would only be "quantitatively" but not principally affected. The effect of language mismatch on the performance of the linguistic-phonetic variables is more substantial, with much increased C_{llr} and EER in cross-language comparisons.

10.3 Implications for FVC

Findings from the current study have implications for forensic casework, particularly in relation to cross-language comparisons. In the IAFPA Code of Practice, forensic analysts are advised to "exercise particular caution with cross-language comparisons" (3.10; IAFPA, 2020). Results from the current investigation provide strong empirical support for this guidance, showing much weakened system validity (and, to some extent, reliability), regardless of the variable's cross-linguistic acoustic stability. The findings here thus provide caution for the use of variables such as /s/ or LTFDs in isolation in cases of language mismatch. However, it is not the case that linguistic-phonetic features can simply provide no speaker-specific information in cross-language comparisons. Somewhat lower C_{llr} and EER could, in some cases, be achieved by combining multiple parameters from the same variable. It could be the case, then, that fusing a careful selection of variables that have been empirically tested may provide stronger system validity.

In cases of cross-language comparison, the results here underscore the need for appropriate reference materials that can be matched with the conditions of the case circumstances. For a numerical LR approach, if the analyst does not have access to data in both languages in the case materials, the use of reference materials in only a single language cannot be recommended on the basis of the current findings, as the resultant system may be highly unreliable. In such cases, pressing ahead with an FVC analysis can easily result in conclusions that are insecure and cannot be validated. Additionally, individual-level analysis demonstrates the complexity of the relationship between the observed values in the acoustic data and the individual performance of the speaker, thus posing analytical challenges to analysts who do not rely on reference databases to assess typicality in such cases.

In order to make any effective use of a chosen feature in such cases, a reference database that consists of recordings from the same set of speakers in both languages is essential, such that case-appropriate calibration can be applied. The dearth of pre-existing databases that contain such data and are suited to forensic use (cf. Morrison, Rose, et al., 2012), as highlighted in Section 2.2.2.2, makes conducting cross-language comparisons in a numerical LR framework especially challenging. Admittedly, the database used in the present study is also suboptimal in this regard, as it consists of only read speech, but it does contain samples collected from the same speakers in different forensically relevant channels (even though only the high-quality condition is used here). Creating such resources, however, would present considerable challenges in addition to those for comparably sized databases involving a more homogeneous population and only a single language. The need for non-contemporaneous recordings for both languages would significantly increase the time and financial resources required, as the number of sessions required of participants and the risk of incomplete participation from speakers both increase. As time is often of essence in forensic casework (Hughes & Rhodes, 2018), it is difficult to imagine how such kind of intensive data collection can be achieved on a case-specific basis. At the same time, the diversity of the population in society means that the possible number of language pairs is potentially huge, and having readily available databases for even a small proportion of the possible language pairs would itself be highly challenging. Recently, Hughes and Wormald (2020) calls for stronger collaboration and data sharing between sociophonetics and forensic speech science, so that the latter can continue to benefit from up-to-date descriptive detail that can emerge from sociophonetic research. Findings from the current study indicate that such a call should well be extended to include the areas of bilingualism and second language acquisition.

On the methodological side, the current findings highlight the usefulness of the zooplot as a diagnostic tool for individual-level analysis, as well as system testing in general. In FVC, zooplot analysis has been applied mostly in the context of ASR (e.g., Alexander et al., 2014; Nash, 2019). In the current study, zooplot analysis is applied to calibrated LRs in the context of FVC instead of uncalibrated scores. As such, individual performance is interpretable not only relative to other speakers, but also in absolute terms, by assessing the performance of individual speakers against the threshold of LLR $= 0$.

On its own, zooplot analysis may give insights into how individual performance varies across different systems, such as between same- and cross-language comparisons in the current study. In addition to identifying the dimension(s) along which performance may vary, zooplot analysis, together with the relational animal groups from Dunstone and Yager (2009), allows for relatively easy identification of problematic speakers who may be disproportionately responsible for errors in the system. These speakers may then be targeted for diagnostic analysis to determine whether particular technical, physiological or behavioural characteristics of the samples or the speakers may be behind the weaknesses of the system. In this study, cross-referencing speakers with outlying performance with the underlying acoustic data has been particularly illuminating in identifying the patterns and sources of performance degradation in cross-language comparisons. The value of conducting such an analysis thus cannot be underestimated and, echoing the call by Alexander et al. (2014), performing individual-level analysis should become a regular practice when system testing is conducted.

The use of the relational animal groups in this study is not without its drawbacks. In particular, the adoption of the criteria in Dunstone and Yager (2009), that is, best and worst 25% in each type of comparison, imposed an arbitrary cut-off for each group. O'Connor et al. (2015) highlights the possibility of a cliff-edge effect concerning speakers who sit on different sides of the boundary but are nonetheless very similar in terms of actual performance. The inclusion of near-members in this study was designed to mitigate this, and the existence of such individuals in a number of systems demonstrates that there may indeed be a risk of such cliff-edge effects in the course of individual-level analysis. It is clear that, by nature, any such schemes that rely on discrete categorisation cannot eliminate the potential effects of losing fine-grained information in individual performance. As such, animal group classifications should be interpreted with due caution and contextualised within the overall distribution of speakers. Indeed, in some of the systems tested in the present study, especially in cross-language comparisons, most of the speakers performed very similarly but a small number of speakers were considerably more extreme. It is very possible that relational groups would capture not only true outliers but also include portions of a large cluster of speakers whose performance cannot be realistically differentiated from the rest of the cluster. In such cases, an alternative form of individual-level analysis that does not depend on the imposition of arbitrary boundaries, but instead seeks to encode such information quantitatively, may be more helpful. When individual performance is compared across different systems, O'Connor et al. (2015) proposes a Stability Score Index to quantify the movement of individuals on the zooplot, which may supplement the "normalised" zooplots devised in the current study and provide a more fine-grained perspective of relative individual performance.

10.4 Future directions

The findings of the current study strengthen the case for examining issues of bilingualism in FVC, particularly within the LR framework. As an initial LR-based exploration of cross-language comparisons, the current study used only relatively tightly controlled materials of high quality, so as to focus on the effect of language mismatch. Future investigations using non-contemporaneous, forensically realistic materials would not only be necessary to ascertain how forensically relevant features perform in cases of language mismatch, but also be useful for exploring how language mismatch interacts with other factors such as channel mismatch in less favourable conditions. The Voice ID Database used in the current study, containing recordings also in landline, mobile phone and covert bug conditions, provides a welcoming opportunity to replicate the experiments carried out in this study, but the scope to expand beyond the limitations of the features chosen here and of using read speech is considerable. A related avenue of research that should also be explored is the choice of linguistic-phonetic variables that can be potentially of value in cross-language comparisons.

The current study also calls for much more research into issues of bilingualism in FVC more generally. In Chapter 2, a number of challenges related to bilingualism in FVC have been identified, especially in relation to the LR framework. The current project focused only on the issue of cross-language comparisons and the effects of language mismatch and training mismatch in such cases, but investigation of the impact of other issues is no less important. One particular issue that is at the heart of cross-language comparisons but the current study is unable to include in its scope is the definition of the relevant population in such cases. Hughes and Foulkes (2015a) has investigated the effects of varying the relevant population according to age group and class and found that a narrowly but inappropriately defined relevant population would give rise to worse validity than a more broadly defined relevant population. Given the particular difficulties of refining the relevant population by non-native-sounding features, the relevant impact in this regard ought to be investigated further.

10.5 Conclusion

This thesis set out to explore the issues of language and language mismatch on three sets of features used in FVC (/s/, LTFDs and ASR). Through an empirical investigation within the numerical LR framework, the current study has demonstrated that the discriminatory power of these variables may be cross-linguistically similar but language-specific. This study has also demonstrated that language mismatch between the samples compared in FVC has an adverse impact on the discriminatory performance of all features tested, albeit at different levels. The sources of degraded performance in cross-language comparisons were further identified through individual-level analysis. It is hoped that findings from this study will contribute to a broader discussion of the issues of bilingualism in FVC and prompt a greater focus of research into this area. It is also hoped that this research has demonstrated the utility of detailed individual-level analysis in the course of system testing and will encourage its adoption in forensic work.

Appendix A

Reading Materials from *Voice ID Database*

Phonemic transcription is for reference only and does not include alternative pronunciations used by speakers in the Database. Spaces between words are included in phonemic transcription for clarity. In French, word-final consonants that are customarily realised as a result of *liaison* are included in the transcription without the linking mark (e.g., *quand il* /kɑ̃t il/).

A.1 English

A.1.1 Sentences

1. The frosty air passed through the coat
   /ðə fɹɔsti ɛɹ pæst θɹu ðə kot/

2. The crooked maze failed to fool the mouse
   /ðə kɹʊkəd mez feld tə ful ðə maʊs/

3. Adding fast leads to wrong sums
   /ædɪŋ fæst lidz tə ɹɔŋ sʌmz/

4. The show was a flop from the very start
   /ðə ʃo wəz ə flɑp fɹəm ðə vɛɹi stɑɹt/

5. A saw is a tool used for making boards
   /ə sɔ ɪz ə tul juzd fɚ mekɪŋ bɔɹdz/

6. The wagon moved on well oiled wheels
   /ðə wægən muvd ɔn wɛl ɔɪld wilz/

7. March the soldiers past the next hill
   /mɑɹtʃ ðə soldʒɚz pæst ðə nɛkst hɪl/

8. A cup of sugar makes sweet fudge
   /ə kʌp əv ʃʊgɚ meks swit fʌdʒ/

9. Place a rosebush near the porch steps
   /ples ə ɹozbʊʃ nɪɹ ðə pɔɹtʃ stɛps/

10. Both lost their lives in the raging storm
    /boθ lɔst ðɛɹ laɪvz ɪn ðə ɹedʒɪŋ stɔɹm/

11. The small pup gnawed a hole through the sock
    /ðə smɔl pʌp nɔd ə hol θɹu ðə sɑk/

12. The fish twisted and turned on the bent hook
    /ðə fɪʃ twɪstəd ənd tɜ˞nd ɔn ðə bɛnt hʊk/

13. Press the pants and sew a button on the vest
    /pɹɛs ðə pænts ənd so ə bʌtən ɔn ðə vɛst/

14. The swan dive was far short of perfect
    /ðə swɑn daɪv wəz fɑɹ ʃɔɹt əv pɜ˞fɪkt/

15. The beauty of the view stunned the young boy
    /ðə bjuti əv ðə vju stʌnd ðə jʌŋ bɔɪ/

16. Two blue fish swam in the tank
    /tu blu fɪʃ swæm ɪn ðə tæŋk/

17. Her purse was full of useless trash
    /hɚ pɜ˞s wəz fʊl əv jusləs tɹæʃ/

18. The colt reared and threw the tall rider
    /ðə kolt ɹɪɹd ənd θɹu ðə tɔl ɹaɪdɚ/

19. It snowed rained and hailed the same morning
    /ɪt snod ɹend ənd held ðə sem mɔɹnɪŋ/

20. Read verse out loud for pleasure
    /ɹid vɜ˞s aʊt laʊd fɚ plɛʒɚ/

A.1.2 Passage (recorded from 2012 onwards)

When the sunlight strikes raindrops in the air, they act as a prism and form
/wɛn ðə sʌnlaɪt stɹaɪks ɹendɹɑps ɪn ði ɛɹ ðe ækt æz ə pɹɪzəm ənd fɔɹm

a rainbow. The rainbow is a division of white light into many beautiful colours.
ə ɹenbo ðə ɹenbo ɪz ə dɪvɪʒən əv waɪt laɪt ɪntu mɛni bjutəfəl kʌlɚz

These take the shape of a long round arch, with its path high above and its two
ðiz tek ðə ʃep əv ə lɔŋ ɹaʊnd ɑɹtʃ wɪθ ɪts pæθ haɪ əbʌv ənd ɪts tu

ends apparently beyond the horizon. There is, according to legend, a boiling pot
ɛndz əpɛɹəntli bɪɑnd ðə hɚaɪzən ðɛɹ ɪz əkɔɹdɪŋ tə lɛdʒənd ə bɔɪlɪŋ pɑt

of gold at one end. People look, but no one ever finds it. When a man looks
əv gold æt wʌn ɛnd pipəl lʊk bʌt no wʌn ɛvɚ faɪndz ɪt wɛn ə mæn lʊks

for something beyond his reach, his friends say he is looking for the pot of gold
fɚ sʌmθɪŋ bɪɑnd hɪz ɹitʃ hɪz fɹɛndz se hi ɪz lʊkɪŋ fɚɹ ðə pɑt əv gold

at the end of the rainbow. Throughout the centuries, people have explained the
æt ði ɛnd əv ðə ɹenbo θɹuaʊt ðə sɛntɹiz pipəl hæv ɛksplend ðə

rainbow in various ways. Some have accepted it as a miracle without physical
ɹenbo ɪn vɛɹiəs wez sʌm hæv əksɛptəd ɪt æz ə mɪɹəkəl wɪθaʊt fɪzɪkəl

explanation. To the Hebrews, it was a token that there would be no more
ɛksplənеʃən tə ðə hibɹuz ɪt wəz ə tokən ðæt ðɛɹ wʊd bi no mɔɹ

universal floods. The Greeks used to imagine that it was a sign from the gods
junəvɝsəl flʌdz ðə gɹiks juzd tə ɪmædʒɪn ðæt ɪt wəz ə saɪn fɹəm ðə gɑdz

to foretell war or heavy rain. The Norsemen considered the rainbow as a bridge
tə fɔɹtɛl wɔɹ ɔɹ hɛvi ɹen ðə nɔɹsmɪn kənsɪdɚd ðə ɹenbo æz ə bɹɪdʒ

over which the gods passed from Earth to their home in the sky.
ovɚ wɪtʃ ðə gɑdz pæst fɹəm ɝθ tə ðɛɹ hom ɪn ðə skaɪ/

A.2 French

A.2.1 Sentences

1. Elle est très patiente quand il écoute sa radio
   /ɛl e tʁɛ pasjɑ̃t kɑ̃d il ekut sa ʁadjo/

2. Ce musicien joue du piano donc l'écureuil grimpe dans l'arbre
   /sə myzisjɛ̃ ʒu dy pjano dɔ̃k lekʁyʁœj gʁɛ̃p dɑ̃ laʁbʁ/

3. Prennent-elles une marche si le ciel est nuageux
   /pʁɛntɛl yn maʁʃ si lə sjɛl e nɥaʒø/

4. Elle à caché sa plume car il lui tire les cheveux
   /ɛl a kaʃe sa plym kaʁ il lɥi tiʁ le ʃəvø/

5. Il va perdre son temps avec ce sac de billes
   /il va pɛʁdʁ sɔ̃ tɑ̃ avɛk sə sak də bij/

6. Ma fille cherche sa bague lorsque ta mère range sa vaisselle
   /ma fij ʃɛʁʃ sa bag lɔʁskə ta mɛʁ ʁɑ̃ʒ sa vɛsɛl/

7. Il regarde par sa fenêtre parce que les chiens jappent très fort
   /il ʁəgɑʁd paʁ sa fənɛtʁ paʁs kə le ʃjɛ̃ ʒap tʁɛ fɔʁ/

8. Elle écrit avec un stylo et il porte ses lunettes
   /ɛl ekʁi avɛk œ̃ stilo e il pɔʁt se lynɛt/

9. Je bois du chocolat chaud quand la cloche sonne à midi
   /ʒə bwa dy ʃokola ʃo kɑ̃ la klɔʃ sɔn a midi/

10. L'oiseau sort de sa cage parce que la lune brille dans le ciel
    /lwazo sɔʁ də sa kaʒ paʁs kə la lyn bʁij dɑ̃ lə sjɛl/

11. Il prend un bain chaud avec ces grenouilles qui sont vertes
    /il pʁɑ̃ œ̃ bɛ̃ ʃo avɛk se gʁənuj ki sɔ̃ vɛʁt/

12. Tout le monde est en classe quand l'éléphant a une longue trompe
    /tu lə mɔ̃d et ɑ̃ klas kɑ̃ lelefɑ̃ a yn lɔ̃g tʁɔ̃p/

13. Elle a compté jusqu'á dix pendant que sa fille lave ses mains
    /ɛl a kɔ̃te ʒyska dis pɑ̃dɑ̃ kə sa fij lav se mɛ̃/

14. La souris mange du fromage et il boit du jus d’orange
    /la suʁi mɑ̃ʒ dy fʁɔmaʒ e il bwa dy ʒy dɔʁɑ̃ʒ/

15. Maman épluche une orange pour ton jus qui est sur la table
    /mamɑ̃n eplyʃ yn ɔʁɑ̃ʒ puʁ tɔ̃ ʒy ki e syʁ la tabl/

16. Elle nage dans la rivière quand ils vont á la plage
    /ɛl naʒ dɑ̃ la ʁivjɛʁ kɑ̃t il vɔ̃ a la plaʒ/

17. Le serveur apporte la crème mais elle mange avec une fourchette
    /lə sɛʁvœʁ apɔʁt la kʁɛm mɛ ɛl mɑ̃ʒ avɛk yn fuʁʃɛt/

18. Ce veau grossit vite mais il ne faut pas manger vite
    /sə vo ɡʁosi vit mɛ il nə fo pa mɑ̃ʒe vit/

19. Elle joue avec sa poupée quand ils vont jouer au parc
    /ɛl ʒu avɛk sa pupe kɑ̃t il vɔ̃ ʒwe o paʁk/

20. Si le bonbon est très sucré il saute sur la trampoline
    /si lə bɔ̃bɔ̃ e tʁɛ sykʁe il sot syʁ la tʁɑ̃polin/

A.2.2 Passage (recorded from 2012 onwards)

La bise et le soleil se disputaient, chacun assurant qu’il était le plus fort, quand ils
/la biz e lə solɛj sə dispytɛ ʃakœ̃ asyʁɑ̃ kil etɛ lə ply fɔʁ kɑ̃t ilz
ont vu un voyageur qui s’avançait, enveloppè dans son manteau. Ils sont tombé
ɔ̃ vy œ̃ vwajaʒœʁ ki savɑ̃sɛ ɑ̃vəlope dɑ̃ sɔ̃ mɑ̃to il sɔ̃ tɔ̃be
d’accord que celui qui arriverait le premier à faire ôter son manteau au voyageur
dakɔʁ kə səlɥi ki aʁivəʁɛ lə pʁəmje a fɛʁ ote sɔ̃ mɑ̃to o vwajaʒœʁ
serait regardé comme le plus fort. Alors, la bise s’est mise à souffler de toute sa
səʁɛ ʁəgɑʁde kɔm lə ply fɔʁ alɔʁ la biz se miz a sufle də tut sa
force mais plus elle soufflait, plus le voyageur serrait son manteau autour de lui
fɔʁs mɛ ply(z) ɛl suflɛ ply lə vwajaʒœʁ sɛʁɛ sɔ̃ mɑ̃to otuʁ də lɥi
et à la fin, la bise a renoncé á le lui faire ôter. Alors le soleil a commencé à
e a la fɛ̃ la biz a ʁɛnɔ̃se a lə lɥi fɛʁ ote alɔʁ lə solɛj a kɔmɑ̃se a
briller et au bout d’un moment, le voyageur réchauffé a ôté son manteau. Ainsi,
bʁije e o bu dœ̃ momɑ̃ lə vwajaʒœʁ ʁeʃofe a ote sɔ̃ mɑ̃to ɛ̃si
la bise a du reconnaître que le soleil était le plus fort des deux.
la biz a dy ʁɛkɔnɛtʁ kə lə solɛj etɛ lə ply fɔʁ de dø/

Appendix B

Zooplots and Summaries for /s/ Dynamic Systems

B.1 Quadratic

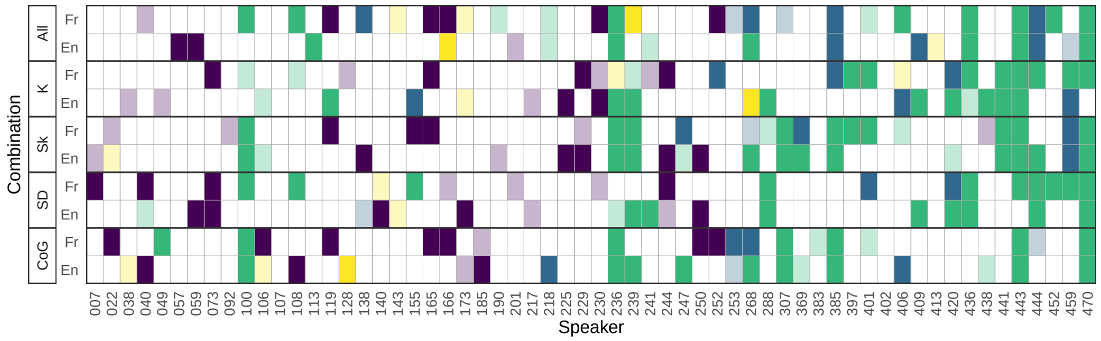

Figure B.1: Summary of speakers classified as doves (green), worms (purple), phantoms (blue) or chameleons (yellow) in quadratic /s/ systems tested (near-members of each group in lighter shade).

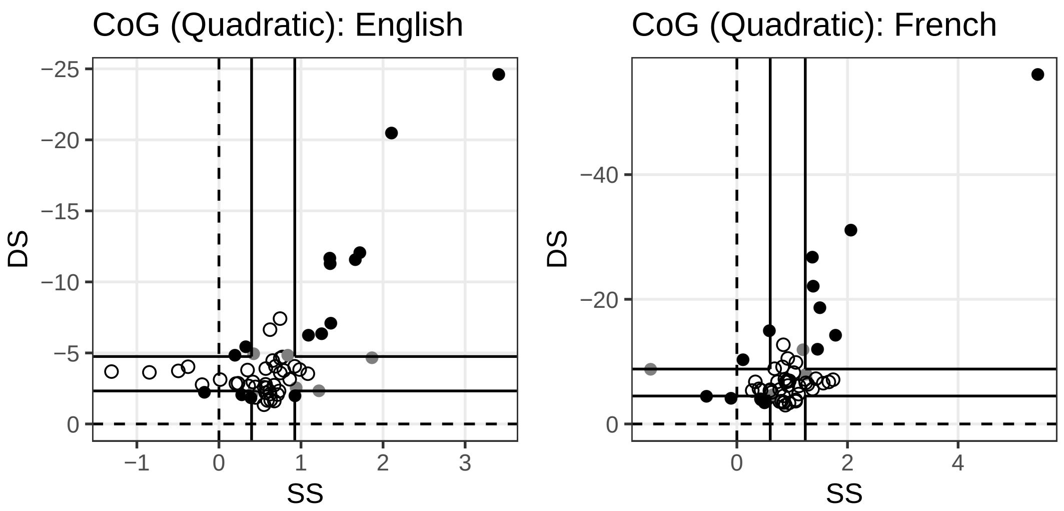

Figure B.2: Zooplots for systems with quadratic CoG as input in English (left) and French (right).

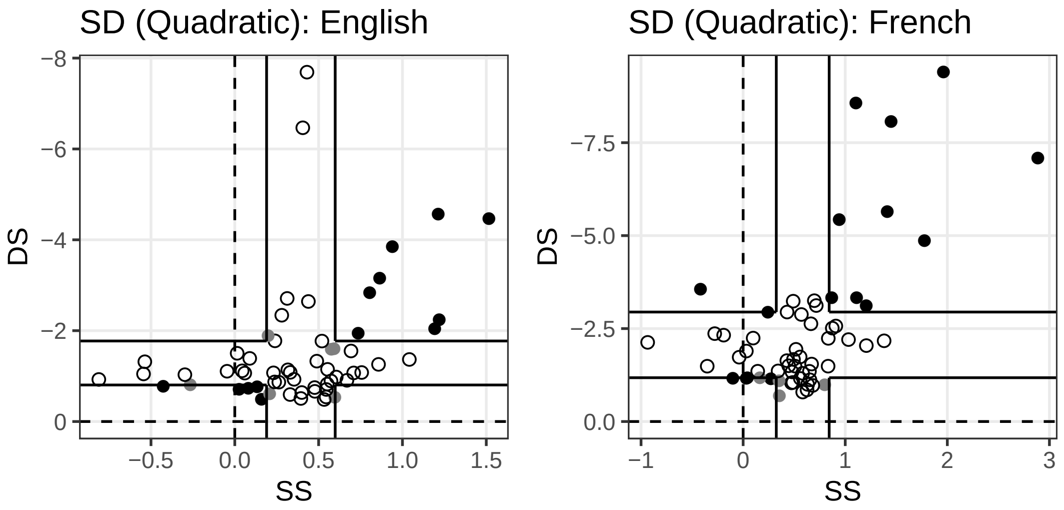

Figure B.3: Zooplots for systems with quadratic SD as input in English (left) and French (right).

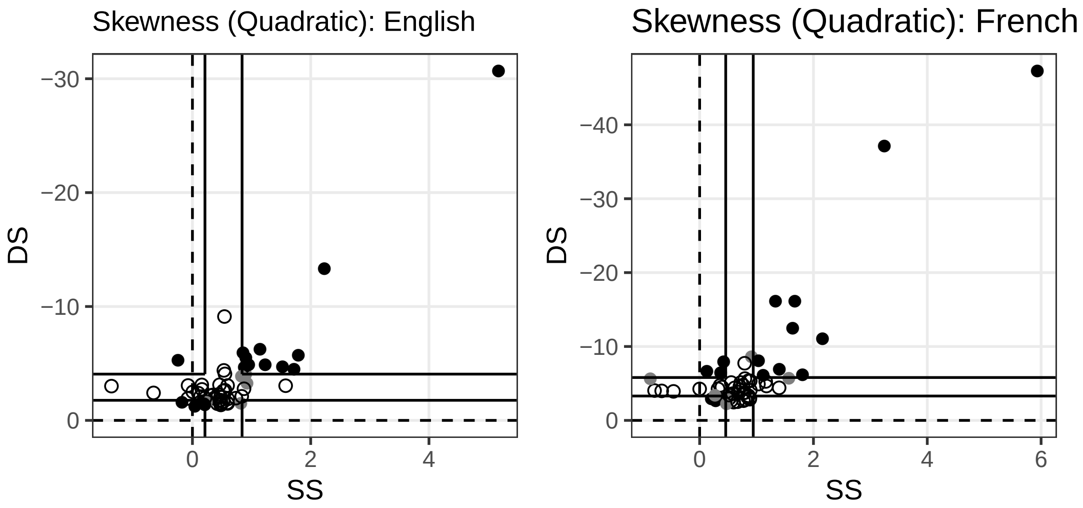

Figure B.4: Zooplots for systems with quadratic skewness as input in English (left) and French (right).

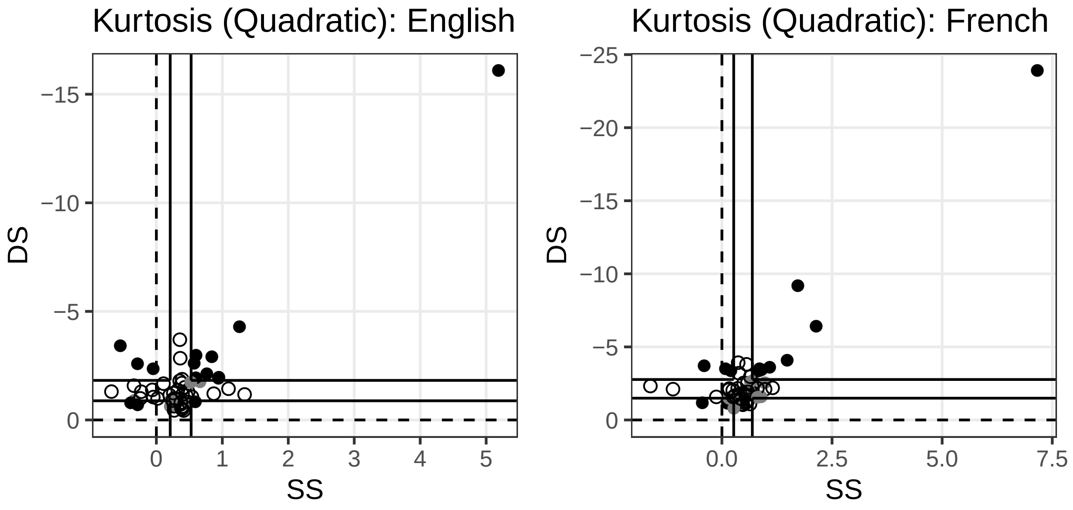

Figure B.5: Zooplots for systems with quadratic log-kurtosis as input in English (left) and French (right).

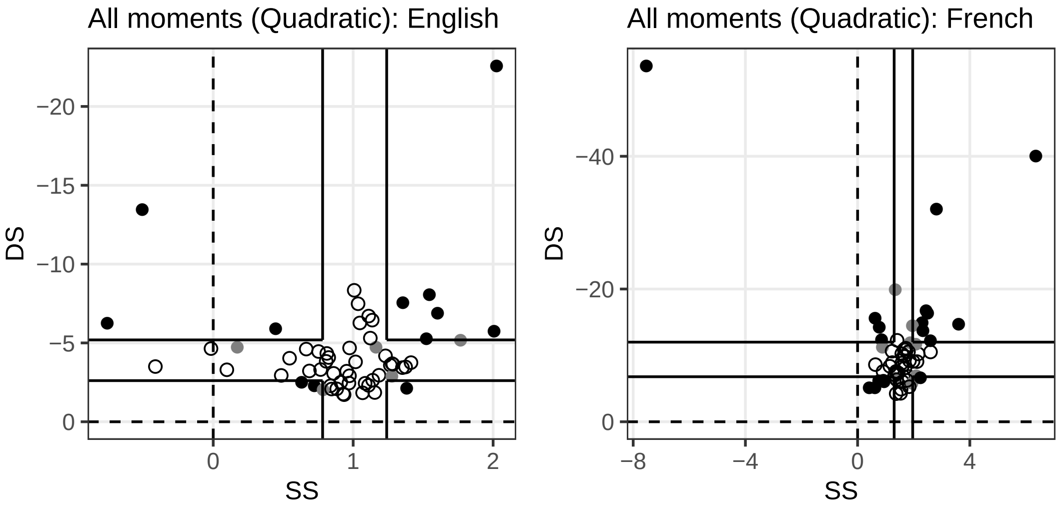

Figure B.6: Zooplots for systems with all four spectral moments in quadratic input in English (left) and French (right).

B.2 Three-point

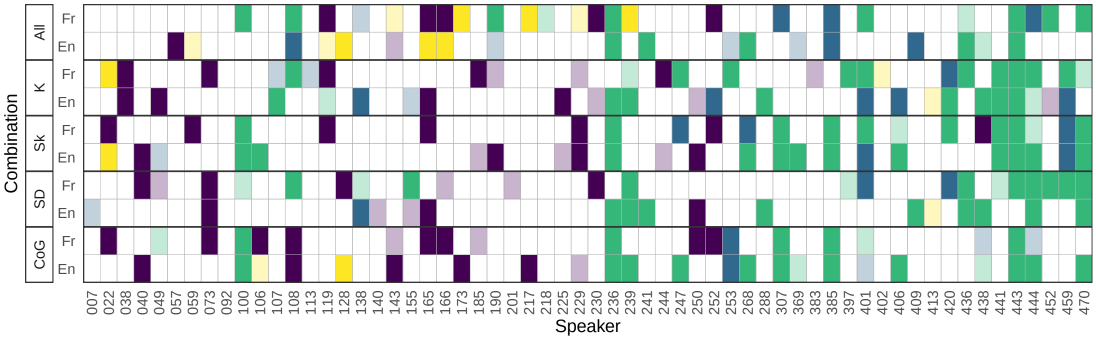

Figure B.7: Summary of speakers classified as doves (green), worms (purple), phantoms (blue) or chameleons (yellow) in three-point /s/ systems tested (near-members of each group in lighter shade).

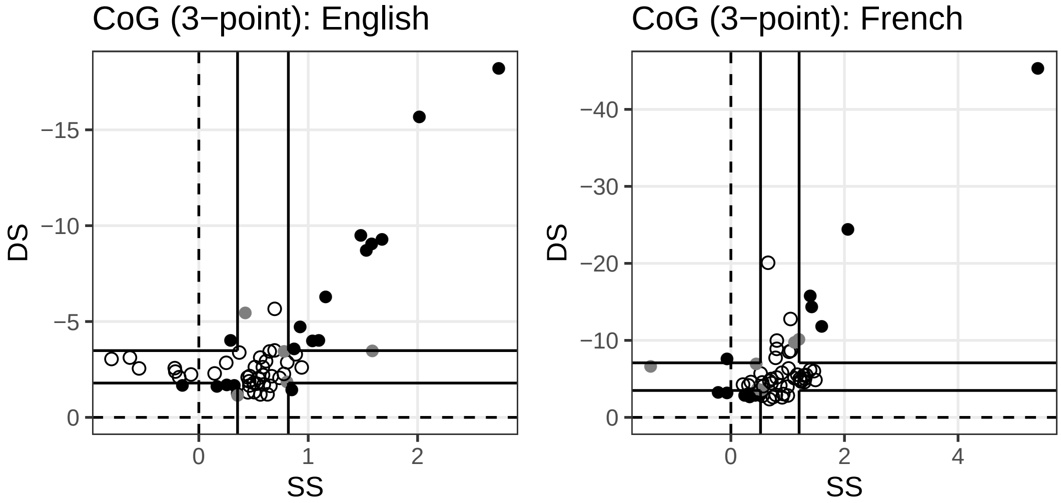

Figure B.8: Zooplots for systems with three-point CoG as input in English (left) and French (right).

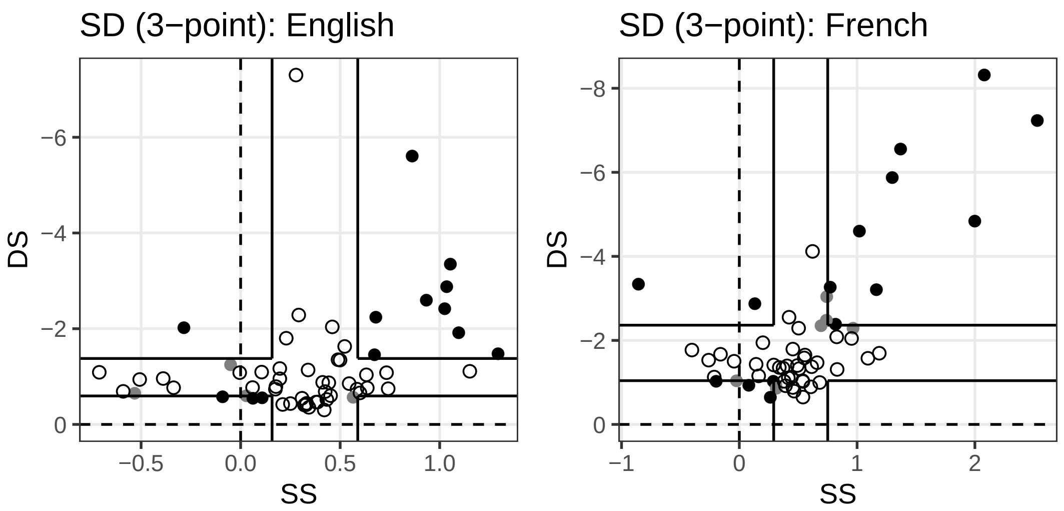

Figure B.9: Zooplots for systems with three-point SD as input in English (left) and French (right).

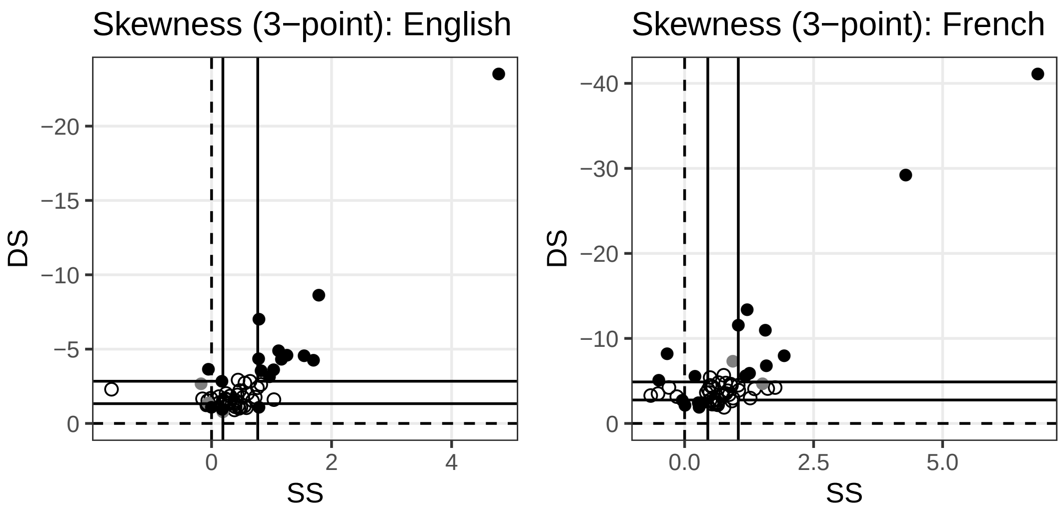

Figure B.10: Zooplots for systems with three-point skewness as input in English (left) and French (right).

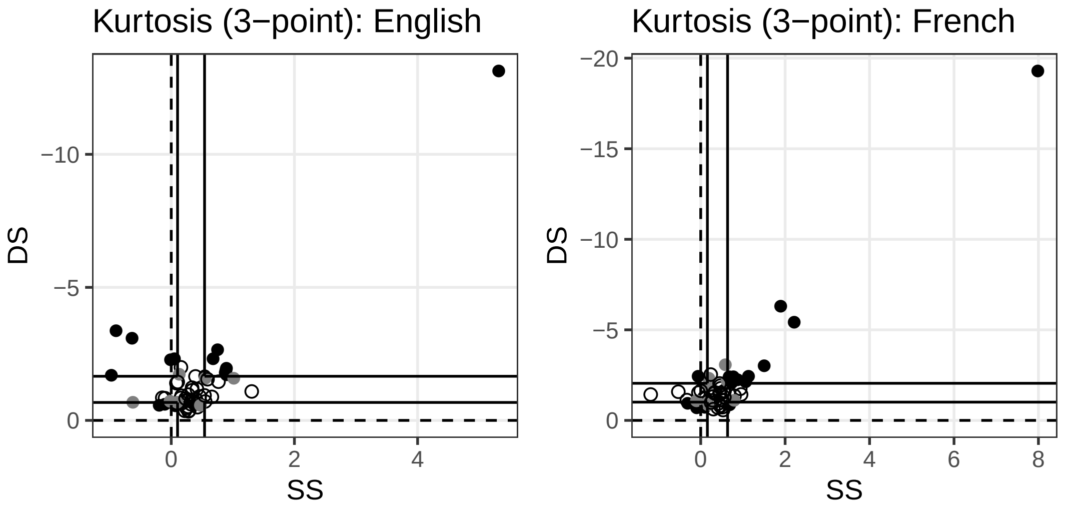

Figure B.11: Zooplots for systems with three-point log-kurtosis as input in English (left) and French (right).

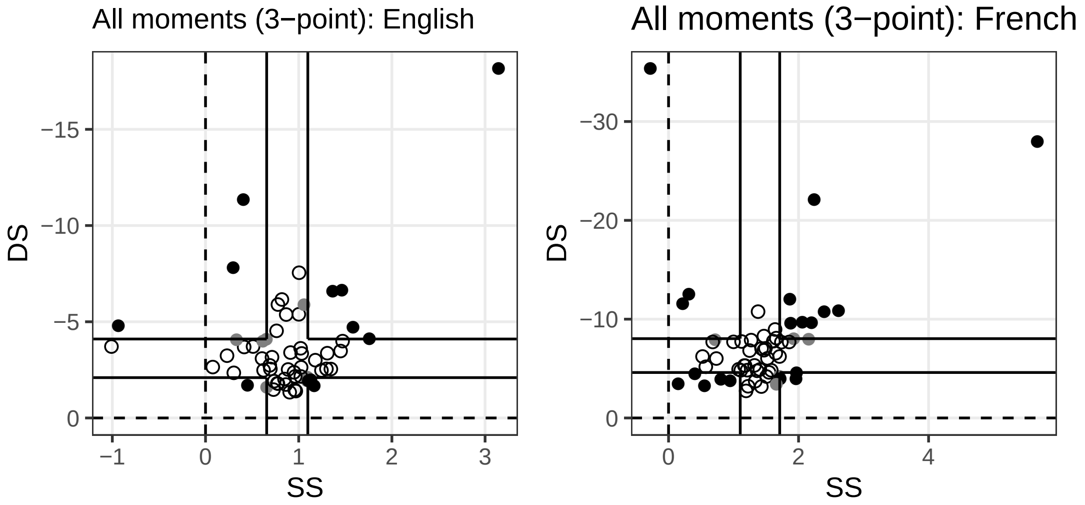

Figure B.12: Zooplots for systems with all four spectral moments in three-point input in English (left) and French (right).

Appendix C

Zooplots and Summary for LTFD F+BW Systems

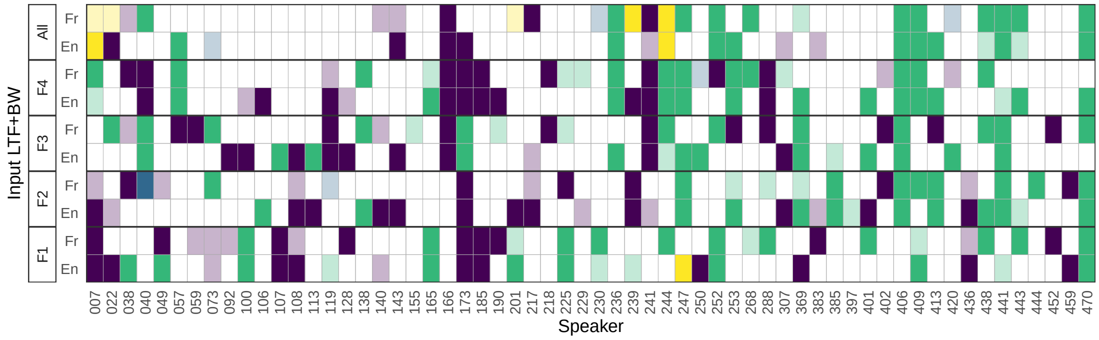

Figure C.1: Summary of speakers classified as doves (green), worms (purple), phantoms (blue) or chameleons (yellow) in F+BW systems tested (near-members of each group in lighter shade).

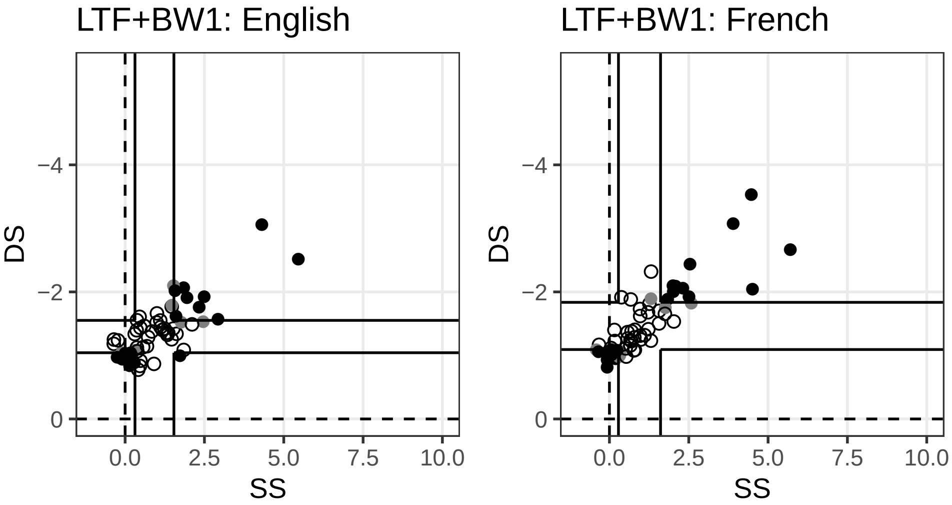

Figure C.2: Zooplots for systems with LTF+BW1 as input in English (left) and French (right).

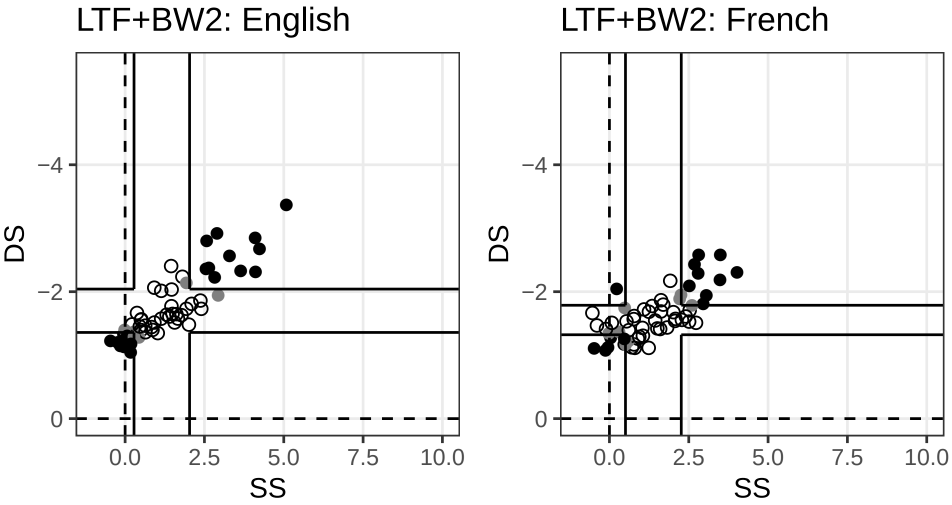

Figure C.3: Zooplots for systems with LTF+BW2 as input in English (left) and French (right).

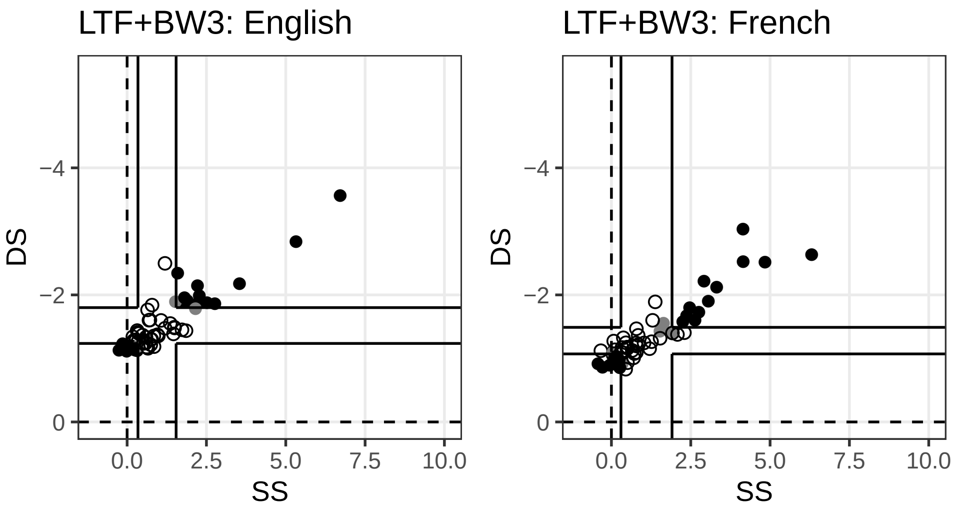

Figure C.4: Zooplots for systems with LTF+BW3 as input in English (left) and French (right).

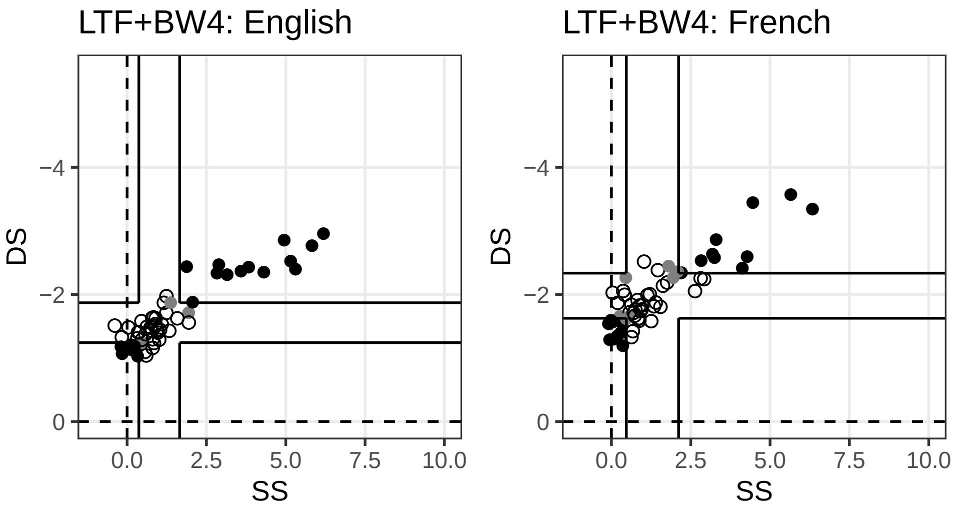

Figure C.5: Zooplots for systems with LTF+BW4 as input in English (left) and French (right).

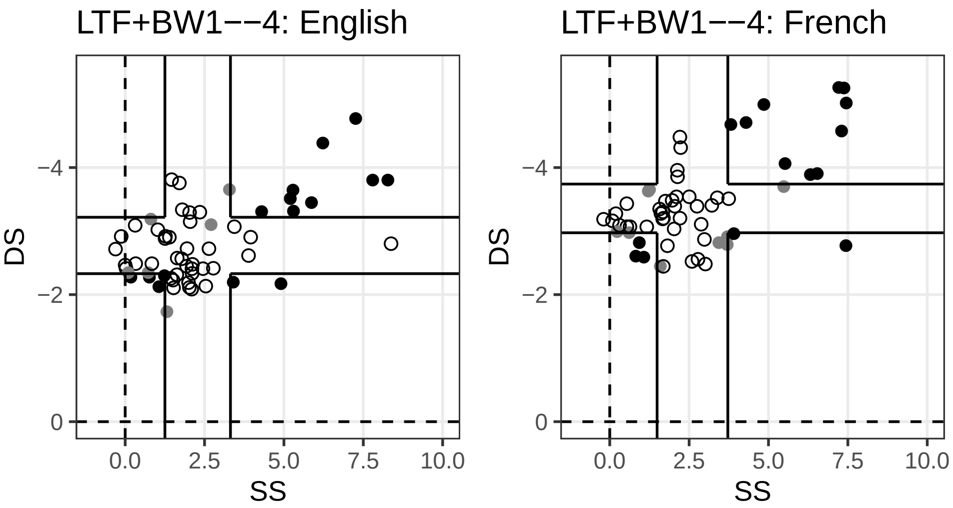

Figure C.6: Zooplots for systems with all four LTFs and BWs as input in English (left) and French (right).

Appendix D

Model comparisons

Table D.1: Results of LRTs from step-down model comparisons for LMEMs fitted to C_{llr} and EER from /s/ systems of same- and cross-language comparisons.

		C_{llr}			EER		
	Predictor	χ^2	df	p	χ^2	df	p
En–Fr							
CoG	Condition × Input	17.47	6	.0077	48.75	6	<.0001
SD	Condition × Input	76.65	6	<.0001	67.09	6	<.0001
Skewness	Condition × Input	61.33	6	<.0001	106.76	6	<.0001
Kurtosis	Condition × Input	30.65	6	<.0001	60.62	6	<.0001
All	Condition × Input	22.19	6	.0011	104.31	6	<.0001
Fr–En							
CoG	Condition × Input	150.87	6	<.0001	162.24	6	<.0001
SD	Condition × Input	76.95	6	<.0001	91.50	6	<.0001
Skewness	Condition × Input	70.21	6	<.0001	38.70	6	<.0001
Kurtosis	Condition × Input	47.58	6	<.0001	65.45	6	<.0001
All	Condition × Input	11.18	6	.0830	35.21	6	<.0001
	Condition	560.39	3	<.0001			
	Input	84.29	2	<.0001			

Table D.2: Results of LRTs from step-down model comparisons for LMEMs fitted to C_{llr} from /s/ systems of cross-language comparisons with and without calibration mismatch.

	Predictor	χ^2	df	p
En–Fr				
CoG	Condition × Mismatch × Input	0.66	2	.7203
	Condition × Mismatch	168.97	1	$<$.0001
	Condition × Input	3.68	2	.1586
	Mismatch × Input	9.74	2	.0008
SD	Condition × Mismatch × Input	0.34	2	.8421
F2	Condition × Mismatch	56.25	1	$<$.0001
	Condition × Input	0.08	2	.9600
	Mismatch × Input	11.46	2	.0033
Skewness	Condition × Mismatch × Input	9.60	2	.0082
Kurtosis	Condition × Mismatch × Input	11.63	2	.0030
All	Condition × Mismatch × Input	7.39	2	.0249
Fr–En				
CoG	Condition × Mismatch × Input	61.72	2	$<$.0001
SD	Condition × Mismatch × Input	1.44	2	.4877
	Condition × Mismatch	96.26	1	$<$.0001
	Condition × Input	0.85	2	.6534
	Mismatch × Input	3.94	2	.1393
	Input	34.13	2	$<$.0001
Skewness	Condition × Mismatch × Input	4.82	2	.0900
	Condition × Mismatch	133.86	1	$<$.0001
	Condition × Input	4.16	2	.1252
	Mismatch × Input	8.37	2	.0152
Kurtosis	Condition × Mismatch × Input	6.12	2	.0468
All	Condition × Mismatch × Input	26.27	2	$<$.0001

Table D.3: Results of LRTs from step-down model comparisons for LMEMs fitted to C_{llr} and EER from LTFD systems of same- and cross-language comparisons.

		C_{llr}			EER		
	Predictor	χ^2	df	p	χ^2	df	p
En–Fr							
F1	Condition × Input	24.94	3	$<.0001$	93.36	3	$<.0001$
F2	Condition × Input	106.78	3	$<.0001$	271.89	3	$<.0001$
F3	Condition × Input	62.84	3	$<.0001$	2.23	3	.5253
	Condition				864.55	3	$<.0001$
	Input				904.41	1	$<.0001$
F4	Condition × Input	62.14	3	$<.0001$	50.54	3	$<.0001$
All	Condition × Input	20.62	3	$<.0001$	25.37	3	$<.0001$
Fr–En							
F1	Condition × Input	89.28	3	$<.0001$	79.29	3	$<.0001$
F2	Condition × Input	180.68	3	$<.0001$	71.15	3	$<.0001$
F3	Condition × Input	188.64	3	$<.0001$	19.29	3	.0002
F4	Condition × Input	81.49	3	$<.0001$	46.47	3	$<.0001$
All	Condition × Input	98.26	3	$<.0001$	28.33	3	$<.0001$

Table D.4: Results of LRTs from step-down model comparisons for LMEMs fitted to C_{llr} from LTFD systems of cross-language comparisons with and without calibration mismatch.

	Predictor	χ^2	df	p
En–Fr				
F1	Condition × Mismatch × Input	0.03	1	.8543
	Condition × Mismatch	53.05	1	<.0001
	Condition × Input	18.88	1	<.0001
	Mismatch × Input	0.03	1	.8604
F2	Condition × Mismatch × Input	78.28	1	<.0001
F3	Condition × Mismatch × Input	0.12	1	.7290
	Condition × Mismatch	47.52	1	<.0001
	Condition × Input	0.31	1	.5762
	Mismatch × Input	66.12	1	<.0001
F4	Condition × Mismatch × Input	25.19	1	<.0001
All	Condition × Mismatch × Input	6.27	1	.0123
Fr–En				
F1	Condition × Mismatch × Input	2.64	1	.1041
	Condition × Mismatch	13.88	1	.0002
	Condition × Input	0.13	1	.7205
	Mismatch × Input	40.41	1	<.0001
F2	Condition × Mismatch × Input	0.79	1	.3743
	Condition × Mismatch	20.42	1	<.0001
	Condition × Input	1.97	1	.1602
	Mismatch × Input	6.02	1	.0141
F3	Condition × Mismatch × Input	4.48	1	.0344
F4	Condition × Mismatch × Input	18.18	1	<.0001
All	Condition × Mismatch × Input	28.08	1	<.0001

References

Abdelli-Beruh, N. B. (2012). Voicing and devoicing assimilation of French /s/ and /z/. *Journal of Psycholinguistic Research*, *41*(5), 371–386. https://doi.org/10.1007/s10936-011-9187-x

Adamson, H. D., & Regan, V. M. (1991). The acquisition of community speech norms by Asian immigrants learning English as a second language: A preliminary study. *Studies in Second Language Acquisition*, *13*(1), 1–22. https://doi.org/10.1017/S0272263100009694

Ahlers, W., & Meer, P. (2019). Sibilant variation in New Englishes: A comparative sociophonetic study of Trinidadian and American English /s(tr)/-retraction. *Proceedings of Interspeech 2019*, 291–295. https://doi.org/10.21437/Interspeech.2019-1821

Aitken, C. G. G., & Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data: *Evaluation of Multivariate Data. Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *53*(1), 109–122. https://doi.org/10.1046/j.0035-9254.2003.05271.x

Aitken, C., & Gold, E. (2013). Evidence evaluation for discrete data. *Forensic Science International*, *230*(1–3), 147–155. https://doi.org/10.1016/j.forsciint.2013.02.042

Akbacak, M., & Hansen, J. H. L. (2007). Language normalization for bilingual speaker recognition systems. *Proceedings of 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, IV.257–IV.260. https://doi.org/10.1109/ICASSP.2007.366898

Akpanglo-Nartey, J. N. (1982). *On fricative phones and phonemes: Measuring the phonetic differences within and between languages*. UCLA Department of Linguistics. https://escholarship.org/uc/item/7ft9b3x0

Alexander, A., Forth, O., Nash, J., & Yager, N. (2014). *Zoo plots for speaker recognition with tall and fat animals* [Oral presentation]. 23rd Annual Conference of the International Association for Forensic Phonetics and Acoustics, Zurich, Switzerland.

Almbark, R., Bouchhioua, N., & Hellmuth, S. (2014). Acquiring the phonetics and phonology of English word stress: Comparing learners from different L1 backgrounds. *Concordia Working Papers in Applied Linguistics*, *5*, 19–35.

Altenberg, E. P., & Ferrand, C. T. (2006). Fundamental frequency in monolingual English, bilingual English/Russian, and bilingual English/Cantonese young adult women. *Journal of Voice*, *20*(1), 89–96. https://doi.org/10.1016/j.jvoice.2005.01.005

Altendorf, U., MacDonald, R., & Thielking, N. (2021). S Retraction in the south-east of England. *Anglistik*, *32*(1), 45–64. https://doi.org/10.33675/ANGL/2021/1/7

Anderson-Hsieh, J., Johnson, R., & Koehler, K. (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentais, prosody, and syllable structure. *Language Learning*, *42*(4), 529–555. https://doi.org/10.1111/j.1467-1770.1992.tb01043.x

Antoniou, M., Best, C. T., Tyler, M. D., & Kroos, C. (2011). Inter-language interference in VOT production by L2-dominant bilinguals: Asymmetries in phonetic code-switching. *Journal of Phonetics*, *39*(4), 558–570. https://doi.org/10.1016/j.wocn.2011.03.001

Asadi, H., Nourbakhsh, M., Sasani, F., & Dellwo, V. (2018). Examining long-term formant frequency as a forensic cue for speaker identification: An experiment on Persian. In M. Nourbakhsh, H. Asadi, & M. Asiaee (Eds.), *Proceedings of the First International Conference on Laboratory Phonetics and Phonology* (pp. 21–28). Neveesh Parsi Publications.

Auckenthaler, R., Carey, M., & Mason, J. (2001). Language dependency in text-independent speaker verification. *Proceedings of 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing*, I.441–I.444. https://doi.org/10.1109/ICASSP.2001.940862

Baese-Berk, M. M., & Morrill, T. H. (2015). Speaking rate consistency in native and non-native speakers of English. *The Journal of the Acoustical Society of America*, *138*(3), EL223–EL228. https://doi.org/10.1121/1.4929622

Bailey, G., Nicholas, S., Baranowski, M., & Turton, D. (2019). *A [ʃ]triking change in Manchester English* [Oral presentation]. UK Language Variation and Change 12, London, UK.

Baird, B. O. (2019). Language-specific pitch ranges among simultaneous K'ichee'-Spanish bilinguals. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019* (pp. 2675–2679). Australasian Speech Science & Technology Association Inc.

Baker, A., Archangeli, D., & Mielke, J. (2011). Variability in American English s-retraction suggests a solution to the actuation problem. *Language Variation and Change*, *23*(3), 347–374. https://doi.org/10.1017/S0954394511000135

Balukas, C., & Koops, C. (2015). Spanish-English bilingual voice onset time in spontaneous code-switching. *International Journal of Bilingualism*, *19*(4), 423–443. https://doi.org/10.1177/1367006913516035

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using *lme4*. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Bayley, R., & Regan, V. (2004). Introduction: The acquisition of sociolinguistic competence. *Journal of Sociolinguistics*, *8*(3), 323–338. https://doi.org/10.1111/j.1467-9841.2004.00263.x

Becker, T., Jessen, M., & Grigoras, C. (2008). Forensic speaker verification using formant features and Gaussian mixture models. *Proceedings of Interspeech 2008*, 1505–1508.

Becker, T., Jessen, M., & Grigoras, C. (2009). Speaker verification based on formants using Gaussian mixture models. *Proceedings of the International Conference on Acoustics NAG/DAGA 2009*, 1640–1643.

Beddor, P. S., Harnsberger, J. D., & Lindemann, S. (2002). Language-specific patterns of vowel-to-vowel coarticulation: Acoustic structures and their perceptual correlates. *Journal of Phonetics*, *30*(4), 591–627. https://doi.org/10.1006/jpho.2002.0177

Beebe, L. M. (1980). Sociolinguistic variation and style shifting in second language acquisition. *Language Learning*, *30*(2), 433–445. https://doi.org/10.1111/j.1467-1770.1980.tb00327.x

Bekker, I., & Levon, E. (2017). The embedded indexical value of /s/-fronting in Afrikaans and South African English. *Linguistics*, *55*(5). https://doi.org/10.1515/ling-2017-0022

Bent, T., Atagi, E., Akbik, A., & Bonifield, E. (2016). Classification of regional dialects, international dialects, and nonnative accents. *Journal of Phonetics*, *58*, 104–117. https://doi.org/10.1016/j.wocn.2016.08.004

Beristain, A. (2021). Spectral properties of anterior sibilant fricatives in Northern Peninsular Spanish and sibilant-merging and non-merging varieties of Basque. *Journal of the International Phonetic Association*, Advance online publication. https://doi.org/10.1017/S0025100320000274

Bessett, R. M. (2017). Exploring the phonological integration of lone other-language nouns in the Spanish of Southern Arizona. *University of Pennsylvania Working Papers in Linguistics*, *23*(2), 31–39. https://repository.upenn.edu/pwpl/vol23/iss2/5

Best, C. T. (1995). A direct realist perspective on cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 171–204). York Press.

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In O.-S. Bohn & M. J. Munro (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 13–34). John Benjamins. https://doi.org/10.1075/lllt.17.07bes

Birdsong, D. (2007). Nativelike pronunciation among late learners of French as a second language. In O.-S. Bohn & M. J. Munro (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 99–116). John Benjamins. https://doi.org/10.1075/lllt.17.12bir

Bladon, R. A. W., & Nolan, F. (1977). A video-fluorographic investigation of tip and blade alveolars in English. *Journal of Phonetics*, *5*(2), 185–193. https://doi.org/10.1016/S0095-4470(19)31128-3

Boberg, C. (2011). Reshaping the vowel system: An Index of Phonetic Innovation in Canadian English. *University of Pennsylvania Working Papers in Linguistics*, *17*(2), Article 4. https://repository.upenn.edu/pwpl/vol17/iss2/4

Boersma, P., & Weenink, D. (2016). *Praat: Doing phonetics by computer* (Version 6.0.19) [Computer software]. http://www.praat.org/

Bolck, A., & Stamouli, A. (2017). Likelihood Ratios for categorical evidence; Comparison of LR models applied to gunshot residue data. *Law, Probability and Risk*, *16*(2-3), 71–90. https://doi.org/10.1093/lpr/mgx005

Bombien, L., & Hoole, P. (2013). Articulatory overlap as a function of voicing in French and German consonant clusters. *The Journal of the Acoustical Society of America*, *134*(1), 539–550. https://doi.org/10.1121/1.4807510

Bonastre, J.-F., & Méloni, H. (1994). Inter- and intra-speaker variability of French phonemes - Advantages of an explicit knowledge based approach. *Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification, and Verification*, 157–160.

Bonastre, J.-F., Méloni, H., & Langlais, P. (1991). Analytical strategy for speaker identification. *Proceedings of the Second European Conference on Speech Communication and Technology (EUROSPEECH 91)*, 435–438.

Boula de Mareüil, P., & Vieru-Dimulescu, B. (2006). The contribution of prosody to the perception of foreign accent. *Phonetica*, *63*(4), 247–267. https://doi.org/10.1159/000097308

Boyd, Z. (2018). *Cross-linguistic variation of /s/ as an index of non-normative sexual orientation and masculinity in French and German men* [Doctoral thesis, University of Edinburgh]. Edinburgh Research Archive. http://hdl.handle.net/1842/33201

Bradlow, A. R. (n.d.). *ALLSSTAR: Archive of L1 and L2 Scripted and Spontaneous Transcripts And Recordings*. https://oscaar3.ling.northwestern.edu/ALLSSTARcentral/#!/recordings

Brajot, F.-X., Mollaei, F., Callahan, M., Klein, D., Baum, S. R., & Gracco, V. L. (2013). Articulatory phonetics of coronal stops in monolingual and simultaneous bilingual speakers of Canadian French and English. *Proceedings of Meetings on Acoustics*, *19*, 060064. https://doi.org/10.1121/1.4799468

Braun, A. (1995). Fundamental frequency: How speaker-specific is it? In A. Braun & J.-P. Köster (Eds.), *Studies in forensic phonetics* (pp. 9–23). Wissenschaftlicher Verlag.

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage*. Cambridge University Press.

Brümmer, N., & du Preez, J. (2006). Application-independent evaluation of speaker detection. *Computer Speech & Language*, *20*(2-3), 230–275. https://doi.org/10.1016/j.csl.2005.08.001

Bullock, B. E., & Toribio, A. J. (2009). Trying to hit a moving target: On the sociophonetics of code-switching. In L. Isurin, D. Winford, & K. de Bot (Eds.), *Multidisciplinary approaches to code switching* (pp. 189–206). John Benjamins. https://doi.org/10.1075/sibil.41.12bul

Burris, C., Vorperian, H. K., Fourakis, M., Kent, R. D., & Bolt, D. M. (2014). Quantitative and descriptive comparison of four acoustic analysis systems: Vowel measurements. *Journal of Speech, Language, and Hearing Research*, *57*(1), 26–45. https://doi.org/10.1044/1092-4388(2013/12-0103)

Byrne, C., & Foulkes, P. (2004). The 'mobile phone effect' on vowel formants. *International Journal of Speech, Language and the Law*, *11*(1), 83–102. https://doi.org/10.1558/ijsll.v11i1.83

Campbell, J. P., Nakasone, H., Cieri, C., Miller, D., Walker, K., Martin, A. F., & Przybocki, M. A. (2004). The MMSR bilingual and crosschannel corpora for speaker recognition research and evaluation. *Proceedings of ODYSSEY 2004 - The Speaker and Language Recognition Workshop*, 29–32.

Cao, W. (2015). Phonetic convergence of Mandarin L2 English speakers towards Australian English. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences* (Paper 1002). University of Glasgow.

Cao, W. (2018). *Short-term accommodation of Hong Kong English speakers towards native English accents and the effect of language attitudes* [Doctoral thesis, University of York]. White Rose eTheses Online. https://etheses.whiterose.ac.uk/23588/

Carne, M., & Ishihara, S. (2019). Disyllabic parameterisation of Vietnamese tonal F0 trajectories in likelihood ratio-based forensic voice comparison. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019* (pp. 780–784). Australasian Speech Science & Technology Association Inc.

Carrie, E. (2017). 'British is professional, American is urban': Attitudes towards English reference accents in Spain. *International Journal of Applied Linguistics*, *27*(2), 427–447. https://doi.org/10.1111/ijal.12139

CECL. (2019). *Learner corpora around the world.* Centre for English Corpus Linguistics, Université Catholique de Louvain. https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html

Champod, C., & Meuwly, D. (2000). The inference of identity in forensic speaker recognition. *Speech Communication*, *31*(2–3), 193–203. https://doi.org/10.1016/S0167-6393(99)00078-3

Chan, A. Y. (2006). Strategies used by Cantonese speakers in pronouncing English initial consonant clusters: Insights into the interlanguage phonology of Cantonese ESL learners in Hong Kong. *International Review of Applied Linguistics in Language Teaching*, *44*(4). https://doi.org/10.1515/IRAL.2006.015

Chan, R. K. W. (2016). Speaker variability in the realisation of lexical tones. *International Journal of Speech, Language and the Law*, *23*(2), 195–214. https://doi.org/10.1558/ijsll.v23i2.30908

Chang, C. B. (2012). Rapid and multifaceted effects of second-language learning on first-language speech production. *Journal of Phonetics*, *40*(2), 249–268. https://doi.org/10.1016/j.wocn.2011.10.007

Chang, C. B. (2019). Phonetic drift. In M. S. Schmid & B. Köpke (Eds.), *The Oxford handbook of language attrition* (pp. 190–203). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198793595.013.16

Chen, M. Y. (1997). Acoustic correlates of English and French nasalized vowels. *The Journal of the Acoustical Society of America*, *102*(4), 2360–2370. https://doi.org/10.1121/1.419620

Cheng, A. (2020). Cross-linguistic $f0$ differences in bilingual speakers of English and Korean. *The Journal of the Acoustical Society of America*, *147*(2), EL67–EL73. https://doi.org/10.1121/10.0000498

Cho, S., & Munro, M. J. (2017). F0, long-term formants and LTAS in Korean-English bilinguals. *Proceedings of the 31st General Meeting of the Phonetic Society of Japan*, 188–193.

Cho, T., & Ladefoged, P. (1999). Variation and universals in VOT: Evidence from 18 languages. *Journal of Phonetics*, *27*(2), 207–229. https://doi.org/10.1006/jpho.1999.0094

Chodroff, E., Golden, A., & Wilson, C. (2019). Covariation of stop voice onset time across languages: Evidence for a universal constraint on phonetic realization. *The Journal of the Acoustical Society of America*, *145*(1), EL109–EL115. https://doi.org/10.1121/1.5088035

Cicres, J. (2011). Los sonidos fricativos sordos y sus implicaciones forenses. *Estudios filológicos*, *48*, 33–48. https://doi.org/10.4067/S0071-17132011000200003

Clarke, S., Elms, F., & Youssef, A. (1995). The third dialect of English: Some Canadian evidence. *Language Variation and Change*, *7*(2), 209–228. https://doi.org/10.1017/S0954394500000995

Cohn, A. C. (1993). Nasalisation in English: Phonology or phonetics. *Phonology*, *10*(1), 43–81. https://doi.org/10.1017/S0952675700001731

Crystal, D. (2003). *English as a global language* (2nd ed.). Cambridge University Press.

Daily Graphic. (2014). *Limping Man jailed 22 years for importing cocaine with MV Benjamin*. https://www.graphic.com.gh/news/general-news/limping-man-jailed-22-years-for-importing-cocaine-with-mv-benjamin.html

Dalton-Puffer, C., Kaltenboeck, G., & Smit, U. (1997). Learner attitudes and L2 pronunciation in Austria. *World Englishes*, *16*(1), 115–128. https://doi.org/10.1111/1467-971X.00052

Dart, S. N. (1991). *Articulatory and acoustic properties of apical and laminal articulations*. UCLA Department of Linguistics. https://escholarship.org/uc/item/52f5v2x2

Dart, S. N. (1998). Comparing French and English coronal consonant articulation. *Journal of Phonetics*, *26*(1), 71–94. https://doi.org/10.1006/jpho.1997.0060

Davidson, L. (2011). Phonetic and phonological factors in the second language production of phonemes and phonotactics: Second language speech production. *Language and Linguistics Compass*, *5*(3), 126–139. https://doi.org/10.1111/j.1749-818X.2010.00266.x

de Boer, M., & Heeren, W. (2019). The speaker-specificity of filled pauses: A cross-linguistic study. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019* (pp. 607–611). Australasian Speech Science & Technology Association Inc.

de Boer, M., & Heeren, W. (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. https://doi.org/10.1121/10.0002871

de Boer, M., & Heeren, W. (2021). *Language-dependency of /m/ in L1 Dutch and L2 English* [Poster presentation]. XVII Associazione Italiana Scienza della Voce (AISV) Annual Conference, Zurich, Switzerland.

de Bruin, A. (2019). Not all bilinguals are the same: A call for more detailed assessments and descriptions of bilingual experiences. *Behavioral Sciences*, *9*(3), 33. https://doi.org/10.3390/bs9030033

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2011). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(4), 788–798. https://doi.org/10.1109/TASL.2010.2064307

de Jong, G., McDougall, K., Hudson, T., & Nolan, F. (2007). The speaker discriminating power of sounds undergoing historical change: A formant-based study. In J. Trouvain & W. J. Barry (Eds.), *Proceedings of the 16th International Congress of Phonetic Sciences ICPhS XVI, 6-10 August 2007, Saarbrücken, Germany* (pp. 1813–1816). Saarland University.

de Jong, N. H., Groenhout, R., Schoonen, R., & Hulstijn, J. H. (2015). Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics*, *36*(2), 223–243. https://doi.org/10.1017/S0142716413000210

de Leeuw, E., Mennen, I., & Scobbie, J. M. (2013). Dynamic systems, maturational constraints and L1 phonetic attrition. *International Journal of Bilingualism*, *17*(6), 683–700. https://doi.org/10.1177/1367006912454620

de Leeuw, E., Schmid, M. S., & Mennen, I. (2010). The effects of contact on native language pronunciation in an L2 migrant setting. *Bilingualism: Language and Cognition*, *13*(1), 33–40. https://doi.org/10.1017/S1366728909990289

de Leeuw, E., Tusha, A., & Schmid, M. S. (2018). Individual phonological attrition in Albanian–English late bilinguals. *Bilingualism: Language and Cognition*, *21*(2), 278–295. https://doi.org/10.1017/S1366728917000025

Dellwo, V., & Schmid, S. (2016). Speaker-individual rhythmic characteristics in read speech of German-Italian bilinguals. In A. Leemann, M.-J. Kolly, S. Schmid, & V. Dellwo (Eds.), *Trends in phonetics and phonology* (pp. 349–362). Peter Lang CH. https://doi.org/10.3726/978-3-0351-0869-9

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, *19*(1), 1–16. https://doi.org/10.1017/S0272263197001010

Deterding, D. (2006). The pronunciation of English by speakers from China. *English World-Wide*, *27*(2), 175–198. https://doi.org/10.1075/eww.27.2.04det

Deterding, D., & Nolan, F. (2007). Aspiration and voicing of Chinese and English plosives. In J. Trouvain & W. J. Barry (Eds.), *Proceedings of the 16th International Congress of Phonetic Sciences ICPhS XVI, 6-10 August 2007, Saarbrücken, Germany* (pp. 385–388). Saarland University.

Deterding, D., Wong, J., & Kirkpatrick, A. (2008). The pronunciation of Hong Kong English. *English World-Wide*, *29*(2), 148–175. https://doi.org/10.1075/eww.29.2.03det

Deuchar, M. (2020). Code-switching in linguistics: A position paper. *Languages*, *5*(2), 22. https://doi.org/10.3390/languages5020022

Deuchar, M., Webb-Davies, P., & Donnelly, K. (2018). *Building and using the Siarad Corpus: Bilingual conversations in Welsh and English*. John Benjamins. https://doi.org/10.1075/scl.81

Doddington, G., Liggett, W., Martin, A., Przybocki, M., & Reynolds, D. (1998). SHEEP, GOATS, LAMBS and WOLVES: A statistical analysis of speaker performance in the NIST 1998 Speaker Recognition Evaluation. *Proceedings of the 5th International Conference on Spoken Language Processing*, Paper 0608.

Dror, I. E. (2020). Cognitive and human factors in expert decision making: Six fallacies and the eight sources of bias. *Analytical Chemistry*, *92*(12), 7998–8004. https://doi.org/10.1021/acs.analchem.0c00704

Dror, I. E., Thompson, W. C., Meissner, C. A., Kornfield, I., Krane, D., Saks, M., & Risinger, M. (2015). Context management toolbox: A linear sequential unmasking (LSU) approach for minimizing cognitive bias in forensic decision making. *Journal of Forensic Sciences*, *60*(4), 1111–1112. https://doi.org/10.1111/1556-4029.12805

Dunstone, T., & Yager, N. (2009). *Biometric system and data analysis*. Springer US. https://doi.org/10.1007/978-0-387-77627-9

Durian, D. (2007). Getting [ʃ]tronger every day?: More on urbanization and the sociogeographic diffusion of (str) in Columbus, OH. *University of Pennsylvania Working Papers in Linguistics*, *13*(2), Article 6. https://repository.upenn.edu/pwpl/vol13/iss2/6

Earnshaw, K. (2014). *Assessing the discriminatory power of /t/ and /k/ for forensic speaker comparison using a likelihood ratio approach* [Unpublished MSc dissertation]. University of York.

Earnshaw, K. (2020). *A forensic phonetic investigation of regional variation and accommodation in West Yorkshire* [Doctoral thesis, University of Huddersfield]. University of Huddersfield Repository. http://eprints.hud.ac.uk/id/eprint/35555/

Edwards, J. (2006). Foundations of bilingualism. In T. K. Bhatia & W. C. Ritchie (Eds.), *The handbook of bilingualism* (1st ed., pp. 7–31). Blackwell Publishing. https://doi.org/10.1002/9780470756997.ch1

Edwards, J. (2012). Bilingualism and multilingualism: Some central concepts. In T. K. Bhatia & W. C. Ritchie (Eds.), *The handbook of bilingualism and multilingualism* (2nd ed., pp. 5–25). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118332382.ch1

Emlen, N. Q. (2017). Multilingualism in the Andes and Amazonia: A view from in-between. *The Journal of Latin American and Caribbean Anthropology*, *22*(3), 556–577. https://doi.org/10.1111/jlca.12250

ENFSI. (2016). *ENFSI guideline for evaluative reporting in forensic science* [European Network of Forensic Science Institutes]. https://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf

European Commission. (2018). Flash Eurobarometer 466: The European education area. https://ec.europa.eu/commfrontoffice/publicopinion/index.cfm/Survey/getSurveyDetail/instruments/FLASH/surveyKy/2186

Evans, S. (2009). The evolution of the English-language speech community in Hong Kong. *English World-Wide*, *30*(3), 278–301. https://doi.org/10.1075/eww.30.3.03eva

Fairbanks, G. (1969). *Voice and articulation drillbook* (2nd ed.). Harper & Row.

Fecher, N. (2014). *Effects of forensically-relevant facial concealment on acoustic and perceptual properties of consonants* [Doctoral thesis, University of York]. White Rose eTheses Online. http://etheses.whiterose.ac.uk/id/eprint/7397

Fehringer, C., & Fry, C. (2007). Hesitation phenomena in the language production of bilingual speakers: The role of working memory. *Folia Linguistica*, *41*(1–2). https://doi.org/10.1515/flin.41.1-2.37

Fejlová, D., Lukeš, D., & Skarnitzl, R. (2013). Formant contours in Czech vowels: Speaker-discriminating potential. *Proceedings of Interspeech 2013*, 3182–3186. https://doi.org/10.21437/Interspeech.2013-706

Fernández Gallardo, L., & Möller, S. (2015). Phoneme intelligibility in narrowband and in wideband channels. In S. Becker (Ed.), *Fortschritte der Akustik - DAGA 2015* (pp. 121–124). Deutsche Gesellschaft für Akustik e.V.

Fitzmaurice, S. (2019). Transnational languages, multilinguals and the challenges for LADO. In P. L. Patrick, M. S. Schmid, & K. Zwaan (Eds.), *Language analysis for* *the determination of origin* (pp. 193–209). Springer International Publishing. https://doi.org/10.1007/978-3-319-79003-9_11

Flege, J. E. (1984). The detection of French accent by American listeners. *The Journal of the Acoustical Society of America*, *76*(3), 692–707. https://doi.org/10.1121/1.391256

Flege, J. E. (1987). The production of "new" and "similar" phones in a foreign language: Evidence for the effect of equivalence classification. *Journal of Phonetics*, *15*(1), 47–65. https://doi.org/10.1016/S0095-4470(19)30537-6

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 233–277). York Press.

Flege, J. E. (2007). Language contact in bilingualism: Phonetic system interactions. In J. Cole & J. I. Hualde (Eds.), *Laboratory phonology 9* (pp. 353–381). Mouton de Gruyter.

Flege, J. E. (2019). A non-critical period for second-language learning. In A. M. Nyvad, M. Hejná, A. Højen, A. B. Jespersen, & M. H. Sørensen (Eds.), *A sound approach to language matters: In honor of Ocke-Schwen Bohn* (pp. 501–541). Aarhus University Library. https://doi.org/10.7146/aul.322.218

Flege, J. E., & Bohn, O.-S. (2021). The revised Speech Learning Model (SLM-r). In R. Wayland (Ed.), *Second language speech learning* (pp. 3–83). Cambridge University Press. https://doi.org/10.1017/9781108886901.002

Flege, J. E., & Eefting, W. (1987). Production and perception of English stops by native Spanish speakers. *Journal of Phonetics*, *15*(1), 67–83. https://doi.org/10.1016/S0095-4470(19)30538-8

Fletcher, S. G., & Newman, D. G. (1991). [s] and [ʃ] as a function of linguapalatal contact place and sibilant groove width. *The Journal of the Acoustical Society of America*, *89*(2), 850–858. https://doi.org/10.1121/1.1894646

Forensic Science Regulator. (2021). *Codes of practice and conduct: Development of evaluative opinions (FSR-C-118)*. https://www.gov.uk/government/publications/development-of-evaluative-opinions

Forrest, K., Weismer, G., Milenkovic, P., & Dougall, R. N. (1988). Statistical analysis of word-initial voiceless obstruents: Preliminary data. *The Journal of the Acoustical Society of America*, *84*(1), 115–123. https://doi.org/10.1121/1.396977

Foulkes, P., & French, P. (2012). Forensic speaker comparison: A linguistic–acoustic perspective. In P. M. Tiersma & L. M. Solan (Eds.), *The Oxford handbook of language and law* (pp. 557–572). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199572120.013.0041

Foulkes, P., French, P., & Wilson, K. (2019). LADO as forensic speaker profiling. In P. L. Patrick, M. S. Schmid, & K. Zwaan (Eds.), *Language analysis for the determination of origin* (pp. 91–116). Springer International Publishing. https://doi.org/10.1007/978-3-319-79003-9_6

Foulkes, P., & Wilson, K. (2011). Language analysis for the determination of origin: An empirical study. In W. S. Lee & E. Zee (Eds.), *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS XVII): August 17-21, 2011* (pp. 691–694).

Fowler, C. A., & Saltzman, E. (1993). Coordination and coarticulation in speech production. *Language and Speech*, *36*(2–3), 171–195. https://doi.org/10.1177/002383099303600304

Fowler, C. A., Sramko, V., Ostry, D. J., Rowland, S. A., & Hallé, P. (2008). Cross language phonetic influences on the speech of French–English bilinguals. *Journal of Phonetics*, *36*(4), 649–663. https://doi.org/10.1016/j.wocn.2008.04.001

French, P. (1998). Mr. Akbar's nearest ear versus the Lombard reflex: A case study in forensic phonetics. *International Journal of Speech, Language and the Law*, *5*(1), 58–68. https://doi.org/10.1558/ijsll.v5i1.58

French, P. (2017). A developmental history of forensic speaker comparison in the UK. *English Phonetics*, *21*, 255–270.

French, P., Foulkes, P., Harrison, P., Hughes, V., San Segundo, E., & Stevens, L. (2015). The vocal tract as a biometric: Output measures, interrelationships, and efficacy. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences* (Paper 0817). University of Glasgow.

French, P., & Fraser, H. (2018). Why "ad hoc experts" should not provide transcripts of indistinct forensic audio, and a proposal for a better approach. *Criminal Law Journal*, *42*(5), 298–302.

French, P., & Harrison, P. (2007). Position Statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases. *International Journal of Speech, Language and the Law*, *14*(1), 3959. https://doi.org/10.1558/ijsll.v14i1.137

French, P., Nolan, F., Foulkes, P., Harrison, P., & McDougall, K. (2010). The UK position statement on forensic speaker comparison: A rejoinder to Rose and Morrison. *International Journal of Speech, Language and the Law*, *17*(1), 143–152. https://doi.org/10.1558/ijsll.v17i1.143

Fricke, M., Kroll, J. F., & Dussias, P. E. (2016). Phonetic variation in bilingual speech: A lens for studying the production–comprehension link. *Journal of Memory and Language*, *89*, 110–137. https://doi.org/10.1016/j.jml.2015.10.001

Frost, D., & Ishihara, S. (2015). Likelihood ratio-based forensic voice comparison on L2 speakers: A case of Hong Kong native male production of English vowels. *Proceedings of the Australasian Language Technology Association Workshop 2015*, 39–47. Retrieved November 22, 2019, from https://www.aclweb.org/anthology/U15-1005

Fuchs, S., & Toda, M. (2010). Do differences in male versus female /s/ reflect biological or sociophonetic factors? In S. Fuchs, M. Toda, & M. Żygis (Eds.), *Turbulent sounds: An interdisciplinary guide* (pp. 281–302). De Gruyter Mouton. https://doi.org/10.1515/9783110226584.281

Gafter, R. J., & Horesh, U. (2020). Two languages, one variable? Pharyngeal realizations among Arabic–Hebrew bilinguals. *Journal of Sociolinguistics*, *24*(3), 369–387. https://doi.org/10.1111/josl.12373

García-Amaya, L., & Lang, S. (2020). Filled pauses are susceptible to cross-language phonetic influence: Evidence from Afrikaans-Spanish bilinguals. *Studies in Second Language Acquisition*, *42*(5), 1077–1105. https://doi.org/10.1017/S0272263120000169

Gardner-Chloros, P. (2009). Sociolinguistic factors in code-switching. In B. E. Bullock & A. J. Toribio (Eds.), *The Cambridge handbook of linguistic code-switching* (pp. 97–113). Cambridge University Press. https://doi.org/10.1017/CBO9780511576331.007

Garibova, J. (2017). Linguistic landscape in Azerbaijan: Policy, attitudes and choices. In L. A'Beckett & T. du Plessis (Eds.), *In pursuit of societal harmony: Reviewing the experiences and approaches in officially monolingual and officially multilingual countries* (pp. 107–145). Conference RAP.

Gavaldà, N. (2016). Individual variation in allophonic processes of /t/ in Standard Southern British English. *International Journal of Speech, Language and the Law*, *23*(1), 43–69. https://doi.org/10.1558/ijsll.v23i1.26870

Gibb-Reid, B. (2018). *Comparing like with like. Using interactional features 'yeah' and 'like' as potential speaker discriminants* [Unpublished MSc dissertation]. University of York.

Gick, B., Wilson, I., Koch, K., & Cook, C. (2004). Language-specific articulatory settings: Evidence from inter-utterance rest position. *Phonetica*, *61*(4), 220–233. https://doi.org/10.1159/000084159

Giot, R., Bourqui, R., & El-Abed, M. (2016). Zoo graph: A new visualisation for biometric system evaluation. *2016 20th International Conference Information Visualisation (IV)*, 190–195. https://doi.org/10.1109/IV.2016.21

Gnevsheva, K. (2018). Variation in foreign accent identification. *Journal of Multilingual and Multicultural Development*, *39*(8), 688–702. https://doi.org/10.1080/01434632.2018.1427756

Gnevsheva, K., & Bürkle, D. (2020). Age estimation in foreign-accented speech by native and non-native speakers. *Language and Speech*, *63*(1), 166–183. https://doi.org/10.1177/0023830919827621

Gold, E. (2018). Articulation rate as a speaker discriminant in British English. *Proceedings of Interspeech 2018*, 1828–1832. https://doi.org/10.21437/Interspeech.2018-1384

Gold, E., & French, P. (2011). International practices in forensic speaker comparison. *International Journal of Speech, Language and the Law*, *18*(2), 293–307. https://doi.org/10.1558/ijsll.v18i2.293

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. https://doi.org/10.1558/ijsll.38028

Gold, E., French, P., & Harrison, P. (2013a). Clicking behavior as a possible speaker discriminant in English. *Journal of the International Phonetic Association*, *43*(3), 339–349. https://doi.org/10.1017/S0025100313000248

Gold, E., French, P., & Harrison, P. (2013b). Examining long-term formant distributions as a discriminant in forensic speaker comparisons under a likelihood ratio framework. *Proceedings of Meetings on Acoustics*, *19*, 060041. https://doi.org/10.1121/1.4800285

Gold, E., & Hughes, V. (2014). Issues and opportunities: The application of the numerical likelihood ratio framework to forensic speaker comparison. *Science & Justice*, *54*(4), 292–299. https://doi.org/10.1016/j.scijus.2014.04.003

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. *Proceedings of Interspeech 2018*, 2748–2752. https://doi.org/10.21437/Interspeech.2018-65

Gombe Muhammad, U., French, P., Chodroff, E., & Brown, G. (2021). *A comparative analysis of Nigerian linguist native speakers, untrained native speakers, UK phoneticians and Y-ACCDIST categorising four accents of Nigerian English* [Poster presentation]. 29th Annual Conference of the International Association for Forensic Phonetics and Acoustics, Marburg, Germany.

Gordon, M., Barthmaier, P., & Sands, K. (2002). A cross-linguistic acoustic study of voiceless fricatives. *Journal of the International Phonetic Association*, *32*(2), 141–174. https://doi.org/10.1017/S0025100302001020

Gorman, K., Howell, J., & Wagner, M. (2011). Prosodylab-aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics*, *39*(3), 192–193.

Graham, C., & Post, B. (2018). Second language acquisition of intonation: Peak alignment in American English. *Journal of Phonetics*, *66*, 1–14. https://doi.org/10.1016/j.wocn.2017.08.002

Grosjean, F. (2012). *Bilingual: Life and reality*. Harvard University Press.

Guion, S. G. (2003). The vowel systems of Quichua-Spanish bilinguals. *Phonetica*, *60*(2), 98–128. https://doi.org/10.1159/000071449

Guirao, M., & García Jurado, M. A. (2009). Frequency of occurence of phonemes in American Spanish. *Revue québécoise de linguistique*, *19*(2), 135–149. https://doi.org/10.7202/602680ar

Gumperz, J. J. (1977). The sociolinguistic significance of conversational code-switching. *RELC Journal*, *8*(2), 1–34. https://doi.org/10.1177/003368827700800201

Gut, U. (2009). *Non-native speech: A corpus-based analysis of phonological and phonetic properties of L2 English and German*. Peter Lang. https://doi.org/10.3726/978-3-653-01155-5

Hagiwara, R. E. (2006). Vowel production in Winnipeg. *The Canadian Journal of Linguistics / La revue canadienne de linguistique*, *51*(2/3), 127–141. https://doi.org/10.1353/cjl.2008.0022

Hallé, P. A., & Adda-Decker, M. (2011). Voice assimilation in French obstruents: Categorical or gradient? In J. A. Goldsmith, E. Hume, & L. Wetzels (Eds.), *Tones and features: Phonetic and phonological perspectives* (pp. 149–175). De Gruyter Mouton. https://doi.org/10.1515/9783110246223.149

Hancin-Bhatt, B. (1994). Segment transfer: A consequence of a dynamic system. *Second Language Research*, *10*(3), 241–269. https://doi.org/10.1177/026765839401000304

Hansen, J. H., & Hasan, T. (2015). Speaker recognition by machines and humans: A tutorial review. *IEEE Signal Processing Magazine*, *32*(6), 74–99. https://doi.org/10.1109/MSP.2015.2462851

Hansen Edwards, J. G. (2015). Hong Kong English: Attitudes, identity, and use. *Asian Englishes*, *17*(3), 184–208. https://doi.org/10.1080/13488678.2015.1049840

Hansen Edwards, J. G. (2016). The politics of language and identity: Attitudes towards Hong Kong English pre and post the Umbrella Movement. *Asian Englishes*, *18*(2), 157–164. https://doi.org/10.1080/13488678.2016.1139937

Harrington, J. (2010). Acoustic phonetics. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), *The handbook of phonetic sciences* (2nd ed., pp. 81–129). Wiley. https://doi.org/10.1002/9781444317251.ch3

Harris, F. (1978). On the use of windows for harmonic analysis with the discrete Fourier transform. *Proceedings of the IEEE*, *66*(1), 51–83. https://doi.org/10.1109/PROC.1978.10837

Harrison, P. (2013). *Making accurate formant measurements: An empirical investigation of the influence of the measurement tool, analysis settings and speaker on formant measurements* [Doctoral thesis, University of York]. White Rose eTheses Online. https://etheses.whiterose.ac.uk/7393/

Hautamäki, R. G., & Kinnunen, T. (2020). Why did the x-vector system miss a target speaker? Impact of acoustic mismatch upon target score on VoxCeleb data. *Proceedings of Interspeech 2020*, 4313–4317. https://doi.org/10.21437/Interspeech.2020-2715

Hayden, R. E. (1950). The relative frequency of phonemes in General-American English. *WORD*, *6*(3), 217–223. https://doi.org/10.1080/00437956.1950.11659381

Heeren, W. (2020). The contribution of dynamic versus static formant information in conversational speech. *International Journal of Speech, Language and the Law*, *27*(1), 75–98. https://doi.org/10.1558/ijsll.41058

Heeren, W., van der Vloed, D., & Vermeulen, J. (2014). *Exploring long-term formants in bilingual speakers* [Oral presentation]. 23rd Annual Conference of the International Association for Forensic Phonetics and Acoustics, Zurich, Switzerland.

Hlavac, J. (2011). Hesitation and monitoring phenomena in bilingual speech: A consequence of code-switching or a strategy to facilitate its incorporation? *Journal of Pragmatics*, *43*(15), 3793–3806. https://doi.org/10.1016/j.pragma.2011.09.008

Holmes-Elliott, S., & Levon, E. (2017). The substance of style: Gender, social class and interactional stance in /s/-fronting in southeast England. *Linguistics*, *55*(5). https://doi.org/10.1515/ling-2017-0020

Holst, T., & Nolan, F. (1995). The influence of syntactic structure on [ʃ] to [ʃ] assimilation. In B. Connell & A. Arvaniti (Eds.), *Phonology and phonetic evidence: Papers in Laboratory Phonology IV* (pp. 315–333). Cambridge University Press. https://doi.org/10.1017/CBO9780511554315.022

Hoole, P., Nguyen-Trong, N., & Hardcastle, W. (1993). A comparative investigation of coarticulation in fricatives: Electropalatographic, electromagnetic, and acoustic Data. *Language and Speech*, *36*(2–3), 235–260. https://doi.org/10.1177/002383099303600307

Hopp, H., & Schmid, M. S. (2013). Perceived foreign accent in first language attrition and second language acquisition: The impact of age of acquisition and bilingualism. *Applied Psycholinguistics*, *34*(2), 361–394. https://doi.org/10.1017/S0142716411000737

Horner, K., & Weber, J. J. (2008). The language situation in Luxembourg. *Current Issues in Language Planning*, *9*(1), 69–128. https://doi.org/10.2167/cilp130.0

Howard, M., Lemée, I., & Regan, V. (2006). The L2 acquisition of a phonological variable: The case of /l/ deletion in French. *Journal of French Language Studies*, *16*(1), 1–24. https://doi.org/10.1017/S0959269506002298

Hudson, T., de Jong, G., McDougall, K., Harrison, P., & Nolan, F. (2007). F0 statistics for 100 young male speakers of Standard Southern British English. In J. Trouvain & W. J. Barry (Eds.), *Proceedings of the 16th International Congress of Phonetic Sciences ICPhS XVI, 6-10 August 2007, Saarbrücken, Germany* (pp. 1809–1812). Saarland University.

Hughes, V. (2014). *The definition of the relevant population and the collection of data for likelihood ratio-based forensic voice comparison* [Doctoral thesis, University of York]. White Rose eTheses Online. http://etheses.whiterose.ac.uk/id/eprint/8309

Hughes, V. (2017). Sample size and the multivariate kernel density likelihood ratio: How many speakers are enough? *Speech Communication*, *94*, 15–29. https://doi.org/10.1016/j.specom.2017.08.005

Hughes, V., Cardoso, A., Harrison, P., Foulkes, P., French, P., & Gully, A. J. (2019). Forensic voice comparison using long-term acoustic measures of laryngeal voice quality. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019* (pp. 1455–1459). Australasian Speech Science & Technology Association Inc.

Hughes, V., Clermont, F., & Harrison, P. (2020). Correlating cepstra with formant frequencies: Implications for phonetically-Informed forensic voice comparison. *Proceedings of Interspeech 2020*, 1858–1862. https://doi.org/10.21437/Interspeech.2020-2216

Hughes, V., & Foulkes, P. (2015a). The relevant population in forensic voice comparison: Effects of varying delimitations of social class and age. *Speech Communication*, *66*, 218–230. https://doi.org/10.1016/j.specom.2014.10.006

Hughes, V., & Foulkes, P. (2015b). Variability in analyst decisions during the computation of numerical likelihood ratios. *International Journal of Speech, Language and the Law*, *21*(2), 279–315. https://doi.org/10.1558/ijsll.v21i2.279

Hughes, V., Harrison, P., Foulkes, P., French, P., Kavanagh, C., & San Segundo, E. (2017). Mapping across feature spaces in forensic voice comparison: The contribution of auditory-based voice quality to (semi-)automatic system testing. *Proceedings of Interspeech 2017*, 3892–3896. https://doi.org/10.21437/Interspeech.2017-1508

Hughes, V., Harrison, P., Foulkes, P., French, P., Kavanagh, C., & San Segundo, E. (2018). The individual and the system: Assessing the stability of the output of a semi-automatic forensic voice comparison system. *Proceedings of Interspeech 2018*, 227–231. https://doi.org/10.21437/Interspeech.2018-1649

Hughes, V., & Rhodes, R. (2018). Questions, propositions and assessing different levels of evidence: Forensic voice comparison in practice. *Science & Justice*, *58*(4), 250–257. https://doi.org/10.1016/j.scijus.2018.03.007

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. https://doi.org/10.1558/ijsll.v23i1.29874

Hughes, V., & Wormald, J. (2020). Sharing innovative methods, data and knowledge across sociophonetics and forensic speech science. *Linguistics Vanguard*, *6*(s1), 20180062. https://doi.org/10.1515/lingvan-2018-0062

IAFPA. (2020). *IAFPA Code of Practice* [International Association for Forensic Phonetics and Acoustics]. https://www.iafpa.net/the-association/code-of-practice/

IEEE. (1969). IEEE recommended practice for speech quality measurements [Institute of Electrical and Electronics Engineers]. *IEEE Transactions on Audio and Electroacoustics*, *17*(3), 225–246. https://doi.org/10.1109/TAU.1969.1162058

Iskarous, K., Mooshammer, C., Hoole, P., Recasens, D., Shadle, C. H., Saltzman, E., & Whalen, D. H. (2013). The coarticulation/invariance scale: Mutual information as a measure of coarticulation resistance, motor synergy, and articulatory invariance. *The Journal of the Acoustical Society of America*, *134*(2), 1271–1282. https://doi.org/10.1121/1.4812855

Iskarous, K., Shadle, C. H., & Proctor, M. (2008). Evidence for the dynamic nature of fricative production: American English /s/. In R. Sock, S. Fuchs, & Y. Laprie (Eds.), *Proceedings of the 8th International Seminar on Speech Production* (pp. 409–412). INRIA.

Iskarous, K., Shadle, C. H., & Proctor, M. I. (2011). Articulatory–acoustic kinematics: The production of American English /s/. *The Journal of the Acoustical Society of America*, *129*(2), 944–954. https://doi.org/10.1121/1.3514537

Jaggers, Z., & Baese-Berk, M. (2019). Moments of moments: Acoustic phonetic character and within-category variability of the Basque three-sibilant contrast. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th Inter-*

national Congress of Phonetic Sciences, Melbourne, Australia 2019 (pp. 1828–1832). Australasian Speech Science & Technology Association Inc.

Jannedy, S., & Weirich, M. (2017). Spectral moments vs discrete cosine transformation coefficients: Evaluation of acoustic measures distinguishing two merging German fricatives. *The Journal of the Acoustical Society of America*, *142*(1), 395–405. https://doi.org/10.1121/1.4991347

Jarvella, R. J., Bang, E., Jakobsen, A. L., & Mees, I. M. (2001). Of mouths and men: Non-native listeners' identification and evaluation of varieties of English. *International Journal of Applied Linguistics*, *11*(1), 37–56. https://doi.org/10.1111/1473-4192.00003

Järvinen, K., Laukkanen, A.-M., & Aaltonen, O. (2013). Speaking a foreign language and its effect on F0. *Logopedics Phoniatrics Vocology*, *38*(2), 47–51. https://doi.org/10.3109/14015439.2012.687764

Jessen, M. (1997). Speaker-specific information in voice quality parameters. *International Journal of Speech, Language and the Law*, *4*(1), 84–103. https://doi.org/10.1558/ijsll.v4i1.84

Jessen, M. (2008). Forensic phonetics. *Language and Linguistics Compass*, *2*(4), 671–711. https://doi.org/10.1111/j.1749-818X.2008.00066.x

Jessen, M. (2020). Speaker profiling and forensic voice comparison. In M. Coulthard, A. May, & R. Sousa-Silva (Eds.), *The Routledge handbook of forensic linguistics* (2nd ed., pp. 382–399). Routledge. https://doi.org/10.4324/9780429030581-31

Jessen, M., & Becker, T. (2010). *Long-Term Formant Distribution as a forensic-phonetic feature* [Oral presentation]. ASA 2nd Pan-American/Iberian Meeting on Acoustics, Cancun, Mexico.

Jessen, M., Bortlík, J., Schwarz, P., & Solewicz, Y. A. (2019). Evaluation of Phonexia automatic speaker recognition software under conditions reflecting those of a real forensic voice comparison case (*forensic_eval_01*). *Speech Communication*, *111*, 22–28. https://doi.org/10.1016/j.specom.2019.05.002

Jessen, M., Köster, 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. https://doi.org/10.1558/sll.2005.12.2.174

Jesus, L. M., & Shadle, C. H. (2002). A parametric study of the spectral characteristics of European Portuguese fricatives. *Journal of Phonetics*, *30*(3), 437–464. https://doi.org/10.1006/jpho.2002.0169

Jiao, D., Watson, V., Wong, S. G.-J., Gnevsheva, K., & Nixon, J. S. (2019). Age estimation in foreign-accented speech by non-native speakers of English. *Speech Communication*, *106*, 118–126. https://doi.org/10.1016/j.specom.2018.12.005

Johnson, K. A. (2021a). Leveraging the uniformity framework to examine crosslinguistic similarity for long-lag stops in spontaneous Cantonese-English bilingual speech. *Proceedings of Interspeech 2021*, 2671–2675. https://doi.org/10.21437/Interspeech.2021-1780

Johnson, K. A. (2021b). *SpiCE: Speech in Cantonese and English* [Data set]. https://doi.org/10.5683/SP2/MJOXP3

Johnson, K. A., Babel, M., & Fuhrman, R. A. (2020). Bilingual acoustic voice variation is similarly structured across languages. *Proceedings of Interspeech 2020*, 2387–2391. https://doi.org/10.21437/Interspeech.2020-3095

Johnson, K. A., & Babel, M. E. (2019). Bilingual sibilant acoustics in conversational Cantonese-English speech. *The Journal of the Acoustical Society of America*, *146*(4), 2839. https://doi.org/10.1121/1.5136840

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America*, *108*(3), 1252–1263. https://doi.org/10.1121/1.1288413

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (2nd ed.). Pearson Prentice Hall.

Kavanagh, C. (2012). *New consonantal acoustic parameters in forensic speaker comparison* [Doctoral thesis, University of York]. White Rose eTheses Online. https://etheses.whiterose.ac.uk/3980/

Kavanagh, C. (2013). Exploring duration and spectral parameters of English /m/ for forensic speaker comparison. *Proceedings of Meetings on Acoustics*, *19*, 060040. https://doi.org/10.1121/1.4798992

Kavanagh, C. (2014). *A cross-linguistic, cross-dialectal investigation of the speaker-specificity of acoustic parameters of [m]* [Unpublished manuscript].

Kent, R. D., & Read, C. (2002). *The acoustic analysis of speech* (2nd ed.). Singular/Thomson Learning.

Kent, R. D., & Vorperian, H. K. (2018). Static measurements of vowel formant frequencies and bandwidths: A review. *Journal of Communication Disorders*, *74*, 74–97. https://doi.org/10.1016/j.jcomdis.2018.05.004

Khattab, G. (2013). Phonetic convergence and divergence strategies in English-Arabic bilingual children. *Linguistics*, *51*(2), 439–472. https://doi.org/10.1515/ling-2013-0017

Kim, J. Y., & Soto-Corominas, A. (2017). Voice quality transfer in the production of Spanish heritage speakers and English L2 learners of Spanish. In S. Perpiñán, D. Heap, & I. Moreno-Villamar (Eds.), *Romance languages and linguistic theory 11: Selected papers from the 44th Linguistic Symposium on Romance Languages (LSRL), London, Ontario* (pp. 191–207). John Benjamins. https://doi.org/10.1075/rllt.11.09kim

Kim, J.-Y. (2019). Heritage speakers' use of prosodic strategies in focus marking in Spanish. *International Journal of Bilingualism*, *23*(5), 986–1004. https://doi.org/10.1177/1367006918763139

King, J., Maclagan, M., Harlow, R., Keegan, P., & Watson, C. (2010). The MAONZE Corpus: Establishing a corpus of Maori speech. *New Zealand Studies in Applied Linguistics*, *16*(2), 1–16. https://www.alanz.org.nz/journal/complete-issue-vol-162-2010/

King, R. (1966). On preferred phonemicizations for statistical studies: Phoneme frequencies in German. *Phonetica*, *15*(1), 22–31. https://doi.org/10.1159/000258535

Kinoshita, Y. (2001). *Testing realistic forensic speaker identification in Japanese: A likelihood ratio-based approach using formants* [Doctoral thesis, Australian National University]. ANU Open Research Library. https://doi.org/10.25911/5d7638f8d32f9

Kinoshita, Y. (2005). Does Lindley's LR estimation formula work for speech data? Investigation using long-term f0. *International Journal of Speech, Language and the Law*, *12*(2), 235–254. https://doi.org/10.1558/sll.2005.12.2.235

Kinoshita, Y., & Ishihara, S. (2014). Background population: How does it affect LR based forensic voice comparison? *International Journal of Speech, Language and the Law*, *21*(2), 191–224. https://doi.org/10.1558/ijsll.v21i2.191

Kitikanan, P., Al-Tamimi, J., & Khattab, G. (2015). An acoustic investigation of the production of English /s/ by L2 Thai learners. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences* (Paper 0655). University of Glasgow.

Klatt, D. (1974). The duration of [s] in English words. *Journal of Speech and Hearing Research*, *17*(1), 51–63. https://doi.org/10.1044/jshr.1701.51

Koenig, L. L., Shadle, C. H., Preston, J. L., & Mooshammer, C. R. (2013). Toward improved spectral measures of /s/: Results From adolescents. *Journal of Speech, Language, and Hearing Research*, *56*(4), 1175–1189. https://doi.org/10.1044/1092-4388(2012/12-0038)

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. https://doi.org/10.1016/j.wocn.2013.11.004

Kolly, M.-J., Leemann, A., Boula de Mareüil, P., & Dellwo, V. (2015). Speaker-idiosyncrasy in pausing behavior: Evidence from a cross-linguistic study. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences* (Paper 0294). University of Glasgow.

Krämer, M. (2019). Is vowel nasalisation phonological in English? A systematic review. *English Language and Linguistics*, *23*(2), 405–437. https://doi.org/10.1017/S1360674317000442

Krebs, P., & Braun, A. (2015). Long Term Formant measurements in bilingual speakers [Oral presentation]. 24th Annual Conference of the International Association for Forensic Phonetics and Acoustics, Leiden, The Netherlands.

Künzel, H. J. (1995). Field procedures in forensic speaker recognition. In J. Windsor Lewis (Ed.), *Studies in general and English phonetics: Essays in honour of Professor J. D. O'Connor* (pp. 68–84). Routledge.

Künzel, H. J. (1997). Some general phonetic and forensic aspects of speaking tempo. *International Journal of Speech, Language and the Law*, *4*(1), 48–83. https://doi.org/10.1558/ijsll.v4i1.48

Künzel, H. J. (2001). Beware of the 'telephone effect': The influence of telephone transmission on the measurement of formant frequencies. *International Journal of Speech, Language and the Law*, *8*(1), 80–99. https://doi.org/10.1558/ijsll.v8i1.80

Künzel, H. J. (2013). Automatic speaker recognition with crosslanguage speech material. *International Journal of Speech, Language and the Law*, *20*(1), 21–44. https://doi.org/10.1558/ijsll.v20i1.21

Kupisch, T., Barton, D., Bianchi, G., & Stangen, I. (2012). The HABLA-corpus (German-French and German-Italian). In T. Schmidt & K. Wörner (Eds.), *Multilingual corpora and multilingual corpus analysis* (pp. 163–179). John Benjamins. https://doi.org/10.1075/hsm.14.11kup

Kupisch, T., Barton, D., Hailer, K., Klaschik, E., Stangen, I., Lein, T., & van de Weijer, J. (2014). Foreign accent in adult simultaneous bilinguals. *Heritage Language Journal*, *11*(2), 123–150. https://doi.org/10.46538/hlj.11.2.2

Ladefoged, P., & Johnson, K. (2015). *A course in phonetics* (7th ed.). Cengage Learning.

Ladefoged, P., & Maddieson, I. (1996). *The sounds of the world's languages*. Blackwell Publishers.

Language and National Origin Group. (2004). Guidelines for the use of language analysis in relation to questions of national origin in refugee cases. *International Journal of Speech, Language and the Law*, *11*(2), 261–266. https://doi.org/10.1558/ijsll.v11i2.261

Laver, J. (1980). *The phonetic description of voice quality*. Cambridge University Press.

Lawrence, W. P. (2000). /str/ ⟶ /ʃtr/: Assimilation at a distance? *American Speech*, *75*(1), 82–87. https://doi.org/10.1215/00031283-75-1-82

Lee, W. S. (1999). An articulatory and acoustical analysis of the syllable-initial sibilants and approximant in Beijing Mandarin. In J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, & A. C. Bailey (Eds.), *Proceedings of the XIVth International Congress of Phonetic Sciences 1-7 August 1999* (pp. 413–416). University of California, Berkeley.

Lenneberg, E. H. (1967). *Biological foundations of language*. Wiley.

Lenth, R. (2020). *emmeans: Estimated marginal means, aka least-squares means* (Version 1.5.0) [Computer software]. https://CRAN.R-project.org/package=emmeans

Li, A., & Post, B. (2014). L2 acquisition of prosodic properties of speech rhythm: Evidence from L1 Mandarin and German learners of English. *Studies in Second Language Acquisition*, *36*(2), 223–255. https://doi.org/10.1017/S0272263113000752

Li, F., Edwards, J., & Beckman, M. E. (2009). Contrast and covert contrast: The phonetic development of voiceless sibilant fricatives in English and Japanese toddlers. *Journal of Phonetics*, *37*(1), 111–124. https://doi.org/10.1016/j.wocn.2008.10.001

Li, J., & Rose, P. (2012). Likelihood ratio-based forensic voice comparison with F-pattern and tonal F0 from the Cantonese /ɔy/ diphthong. In F. Cox, K. Demuth, S. Lin, K. Miles, S. Palethorpe, J. Shaw, & I. Yuen (Eds.), *Proceedings of the 14th Australasian International Conference on Speech Science and Technology* (pp. 201–204). Australasian Speech Science & Technology Association Inc.

Lin, Y.-H. (2003). Interphonology variability: Sociolinguistic factors affecting L2 simplification strategies. *Applied Linguistics*, *24*(4), 439–464. https://doi.org/10.1093/applin/24.4.439

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *WORD*, *20*(3), 384–422. https://doi.org/10.1080/00437956.1964.11659830

Lo, J. J. H. (2020). Between *äh(m)* and *euh(m)* : The distribution and realization of filled pauses in the speech of German-French simultaneous bilinguals. *Language and Speech*, *63*(4), 746–768. https://doi.org/10.1177/0023830919890068

Lo, J. J. H. (2021). *fvclrr: Likelihood ratio calculation and testing in forensic voice comparison* [Computer software]. https://github.com/justinjhlo/fvclrr

Loakes, D. (2004). Front vowels as speaker-specific: Some evidence from Australian English. In S. Cassidy, F. Cox, R. Mannell, & S. Palethorpe (Eds.), *Proceedings of the 10th Australian International Conference on Speech Science & Technology: Macquarie University, 8-10 December, 2004* (pp. 289–294). Australian Speech Science & Technology Association Inc.

Loewen, S., & Hui, B. (2021). Small samples in instructed second language acquisition research. *The Modern Language Journal*, *105*(1), 187–193. https://doi.org/10.1111/modl.12700

Lombardi, L. (2003). Second language data and constraints on manner: Explaining substitutions for the English interdentals. *Second Language Research*, *19*(3), 225–250. https://doi.org/10.1177/026765830301900304

Lyu, D.-C., Tan, T.-P., Chng, E.-S., & Li, H. (2015). Mandarin—English code-switching speech corpus in South-East Asia: SEAME. *Language Resources and Evaluation*, *49*(3), 581–600. https://doi.org/10.1007/s10579-015-9303-x

Ma, B., Meng, H. M., & Mak, M.-W. (2007). Effects of device mismatch, language mismatch and environmental mismatch on speaker verification. *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, *4*, IV–301–IV–304. https://doi.org/10.1109/ICASSP.2007.366909

Magen, H. S. (1998). The perception of foreign-accented speech. *Journal of Phonetics*, *26*(4), 381–400. https://doi.org/10.1006/jpho.1998.0081

Magrin-Chagnolleau, I., Bonastre, J.-F., & Bimbot, F. (1995). Effect of utterance duration and phonetic content on speaker identification using second-order statistical methods. *Proceedings of the 4th European Conference on Speech Communication and Technology (EUROSPEECH 95)*, 337–340.

Mair, V. (1991). What is a Chinese "dialect/topolect"? Reflections on some key Sino-English linguistic terms. *Sino-Platonic Papers*, *29*, 1–31.

Major, R. C. (2007). Identifying a foreign accent in an unfamiliar language. *Studies in Second Language Acquisition*, *29*(4), 539–556. https://doi.org/10.1017/S0272263107070428

Malécot, A. (1974). Frequency of occurrence of French phonemes and consonant clusters. *Phonetica*, *29*(3), 158–170. https://doi.org/10.1159/000259468

Marquina Zarauza, M. (2016). *Estudio fonético-acústico de la variación inter e intrahablante de hablantes bilingües de catalán y de castellano* [Doctoral dissertation, Universitat Pompeu Fabra]. TDX. http://hdl.handle.net/10803/398981

Martire, K., Kemp, R., Sayle, M., & Newell, B. (2014). On the interpretation of likelihood ratios in forensic science evidence: Presentation formats and the weak evidence effect. *Forensic Science International*, *240*, 61–68. https://doi.org/10.1016/j.forsciint.2014.04.005

Mayr, R., Roberts, L., & Morris, J. (2019). Can you tell by their English if they can speak Welsh? Accent perception in a language contact situation. *International Journal of Bilingualism*. https://doi.org/10.1177/1367006919883035

McAuliffe, M. J., Ward, E. C., & Murdoch, B. E. (2001). Tongue-to-palate contact patterns and variability of four English consonants in an /i/ vowel environment. *Logopedics Phoniatrics Vocology*, *26*(4), 165–178. https://doi.org/10.1080/14015430127770

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. *Proceedings of Interspeech 2017*, 498–502. https://doi.org/10.21437/Interspeech.2017-1386

McDougall, K. (2004). Speaker-specific formant dynamics: An experiment on Australian English /aɪ/. *International Journal of Speech, Language and the Law*, *11*(1), 103–130. https://doi.org/10.1558/sll.2004.11.1.103

McDougall, K. (2006). Dynamic features of speech and the characterization of speakers: Toward a new approach using formant frequencies. *International Journal of Speech, Language and the Law*, *13*(1), 89–126. https://doi.org/10.1558/ijsll.v13i1.89

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*(2), 205–230. https://doi.org/10.1558/ijsll.37241

McKenzie, R. M. (2015). The sociolinguistics of variety identification and categorisation: Free classification of varieties of spoken English amongst non-linguist listeners. *Language Awareness*, *24*(2), 150–168. https://doi.org/10.1080/09658416.2014.998232

Mehmedbegovic, D., & Bak, T. H. (2017). Towards an interdisciplinary lifetime approach to multilingualism: From implicit assumptions to current evidence. *European Journal of Language Policy*, *9*(2), 149–167. https://doi.org/10.3828/ejlp.2017.10

Mennen, I. (2007). Phonological and phonetic Influences in non-native intonation. In J. Trouvain & U. Gut (Eds.), *Non-native prosody: Phonetic description and teaching practice* (pp. 53–76). De Gruyter Mouton. https://doi.org/10.1515/9783110198751.1.53

Mennen, I., Scobbie, J. M., de Leeuw, E., Schaeffler, S., & Schaeffler, F. (2010). Measuring language-specific phonetic settings. *Second Language Research*, *26*(1), 13–41. https://doi.org/10.1177/0267658309337617

Mines, M. A., Hanson, B. F., & Shoup, J. E. (1978). Frequency of occurrence of phonemes in conversational English. *Language and Speech*, *21*(3), 221–241. https://doi.org/10.1177/002383097802100302

Modern Ghana. (2007a). *In the matter of the missing 76 parcels of cocaine saga we investigated Manhyia-Prosecution witness tells court*. https://www.modernghana.com/news/121877/in-the-matter-of-the-missing-76-parcels-of-cocaine.html

Modern Ghana. (2007b). *Tagor and Abass jailed 15 years in hard labour*. https://www.modernghana.com/news/148863/tagor-and-abass-jailed-15-years-in-hard-labour.html

Modern Ghana. (2009). *Tagor, Abass acquitted and discharged*. https://www.modernghana.com/news/229591/tagor-abass-acquitted-and-discharged.html

Modern Ghana. (2010). *MV Benjamin cocaine affair resurrected Kofi Boakye on Issa Abass' radar... Another legal battle in the pipeline*. https://www.modernghana.com/news/272560/mv-benjamin-cocaine-affair-resurrected-kofi-boakye.html

Modern Ghana. (2011). *Presidential commission on MV Benjamin scrapped*. https://www.modernghana.com/news/324495/presidential-commission-on-mv-benjamin-scrapped.html

Mok, P. P. K. (2010). Language-specific realizations of syllable structure and vowel-to-vowel coarticulation. *The Journal of the Acoustical Society of America*, *128*(3), 1346–1356. https://doi.org/10.1121/1.3466859

Mok, P. P. K., Xu, R. B., & Zuo, D. (2015). Bilingual speaker identification: Chinese and English. *International Journal of Speech, Language and the Law*, *22*(1), 57–78. https://doi.org/10.1558/ijsll.v22i1.18636

Moos, A. (2008). *Forensische Sprechererkennung mit der Messmethode LTF (long-term formant distribution)* [Unpublished Master's dissertation]. Saarland University.

Moos, A. (2010). Long-term formant distribution as a measure of speaker characteristics in read and spontaneous speech. *The Phonetician*, *101/102*, 7–24.

Mooshammer, C., Hoole, P., & Geumann, A. (2007). Jaw and order. *Language and Speech*, *50*(2), 145–176. https://doi.org/10.1177/00238309070500020101

Morrison, G. S. (2007). *Matlab implementation of Aitken & Lucy's (2004) forensic likelihood-ratio software using multivariate-kernel-density estimation* [Computer software]. http://geoff-morrison.net/#MVKD

Morrison, G. S. (2009a). Forensic speaker recognition using likelihood ratios based on polynomial curves fitted to the formant trajectories of Australian English /aI/. *International Journal of Speech, Language and the Law*, *15*(2), 249–266. https://doi.org/10.1558/ijsll.v15i2.249

Morrison, G. S. (2009b). Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs. *The Journal of the Acoustical Society of America*, *125*(4), 2387–2397. https://doi.org/10.1121/1.3081384

Morrison, G. S. (2011). A comparison of procedures for the calculation of forensic likelihood ratios from acoustic–phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model–universal background model (GMM–UBM). *Speech Communication*, *53*(2), 242–256. https://doi.org/10.1016/j.specom.2010.09.005

Morrison, G. S. (2013). Tutorial on logistic-regression calibration and fusion: Converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences*, *45*(2), 173–197. https://doi.org/10.1080/00450618.2012.733025

Morrison, G. S. (2014). Distinguishing between forensic science and forensic pseudoscience: Testing of validity and reliability, and approaches to forensic voice comparison. *Science & Justice*, *54*(3), 245–256. https://doi.org/10.1016/j.scijus.2013.07.004

Morrison, G. S. (2018). The impact in forensic voice comparison of lack of calibration and of mismatched conditions between the known-speaker recording and the relevant-population sample recordings. *Forensic Science International*, *283*, e1–e7. https://doi.org/10.1016/j.forsciint.2017.12.024

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

Morrison, G. S., & Enzinger, E. (2019). Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (forensic_eval_01) – Conclusion. *Speech Communication*, *112*, 37–39. https://doi.org/10.1016/j.specom.2019.06.007

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

Morrison, G. S., Enzinger, E., Ramos, D., González-Rodríguez, J., & Lozano-Díez, A. (2020). Statistical models in forensic voice comparison. In D. L. Banks, K. Kafadar, D. H. Kaye, & M. Tackett (Eds.), *Handbook of forensic statistics* (pp. 451–497). Chapman; Hall/CRC. https://doi.org/10.1201/9780367527709-20

Morrison, G. S., Enzinger, E., & Zhang, C. (2016). Refining the relevant population in forensic voice comparison – A response to Hicks *et alii* (2015) The importance of distinguishing information from evidence/observations when formulating propositions. *Science & Justice*, *56*(6), 492–497. https://doi.org/10.1016/j.scijus.2016.07.002

Morrison, G. S., Enzinger, E., & Zhang, C. (2017). Reply to Hicks *et alii* (2017) Reply to Morrison *et alii* (2016) Refining the relevant population in forensic voice comparison - A response to Hicks *et alii* (2015) The importance of distinguishing info from evidence/observations when formulating propositions. *arXiv*. http://arxiv.org/abs/1704.07639

Morrison, G. S., Ochoa, F., & Thiruvaran, T. (2012). Database for forensic voice comparison. In H. Li, B. Ma, & K. A. Lee (Eds.), *Proceedings of Odyssey 2012: The Speaker and Language Recognition Workshop* (pp. 62–77). Chinese & Oriental Languages Information Processing Society, Speaker & Language Characterization SIG.

Morrison, G. S., & Poh, N. (2018). Avoiding overstating the strength of forensic evidence: Shrunk likelihood ratios/Bayes factors. *Science & Justice*, *58*(3), 200–218. https://doi.org/10.1016/j.scijus.2017.12.005

Morrison, G. S., Rose, P., & Zhang, C. (2012). Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice. *Australian Journal of Forensic Sciences*, *44*(2), 155–167. https://doi.org/10.1080/00450618.2011.630412

Morrison, G. S., Sahito, F. H., Jardine, G., Djokic, D., Clavet, S., Berghs, S., & Goemans Dorny, C. (2016). INTERPOL survey of the use of speaker identification by law enforcement agencies. *Forensic Science International*, *263*, 92–100. https://doi.org/10.1016/j.forsciint.2016.03.044

Mougeon, R., Rehner, K., & Nadasdi, T. (2004). The learning of spoken French variation by immersion students from Toronto, Canada. *Journal of Sociolinguistics*, *8*(3), 408–432. https://doi.org/10.1111/j.1467-9841.2004.00267.x

Muldner, K., Hoiting, L., Sanger, L., Blumenfeld, L., & Toivonen, I. (2019). The phonetics of code-switched vowels. *International Journal of Bilingualism*, *23*(1), 37–52. https://doi.org/10.1177/1367006917709093

Mullen, C., Spence, D., Moxey, L., & Jamieson, A. (2014). Perception problems of the verbal scale. *Science & Justice*, *54*(2), 154–158. https://doi.org/10.1016/j.scijus.2013.10.004

Myers, S. (2010). Regressive voicing assimilation: Production and perception studies. *Journal of the International Phonetic Association*, *40*(2), 163–179. https://doi.org/10.1017/S0025100309990284

Nagao, K. (2006). *Cross-language study of age perception* [Unpublished doctoral thesis]. Indiana University.

Nair, B., Alzqhoul, E., & Guillemin, B. J. (2014). Determination of likelihood ratios for forensic voice comparison using Principal Component Analysis. *International Journal of Speech, Language and the Law*, *21*(1), 83–112. https://doi.org/10.1558/ijsll.v21i1.83

Narayanan, S. S., Alwan, A. A., & Haker, K. (1995). An articulatory study of fricative consonants using magnetic resonance imaging. *The Journal of the Acoustical Society of America*, *98*(3), 1325–1347. https://doi.org/10.1121/1.413469

Nash, J. (2019). *The effect of acoustic variability on automatic speaker recognition systems* [Doctoral thesis, University of York]. White Rose eTheses Online. https://etheses.whiterose.ac.uk/27337/

Ng, M. L., Chen, Y., & Chan, E. Y. (2012). Differences in vocal characteristics between Cantonese and English produced by proficient Cantonese-English bilingual speakers—A long-term average spectral analysis. *Journal of Voice*, *26*(4), e171–e176. https://doi.org/10.1016/j.jvoice.2011.07.013

Niebuhr, O., Clayards, M., Meunier, C., & Lancia, L. (2011). On place assimilation in sibilant sequences—Comparing French and English. *Journal of Phonetics*, *39*(3), 429–451. https://doi.org/10.1016/j.wocn.2011.04.003

Nolan, F. (1983). *The phonetic bases of speaker recognition*. Cambridge University Press.

Nolan, F. (1997). Speaker recognition and forensic phonetics. In W. J. Hardcastle & J. Laver (Eds.), *The handbook of phonetic sciences* (1st ed., pp. 744–767). Blackwell.

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

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

O'Connor, K., Elliott, S., Sutton, M., & Dyrenfurth, M. (2015). Stability of individuals in a fingerprint system across force levels: An introduction to the Stability Score Index. *Information Technology in Industry*, *3*(2), 46–53.

Office for National Statistics. (2016). *Census 2011 Household Questionnaire - England.* https://www.ons.gov.uk/file?uri=/census/2011census/howourcensusworks/howwetookthe2011census/howwecollectedtheinformation/questionnairesdelivery-completionandreturn/h1-2011-watermark_tcm77-190660.pdf

Office for National Statistics. (2020). *Census 2021 Individual Questionnaire - England.* https://www.ons.gov.uk/file?uri=/census/censustransformationprogramme/questiondevelopment/census2021paperquestionnaires/englishindividual.pdf

Olson, D. J. (2013). Bilingual language switching and selection at the phonetic level: Asymmetrical transfer in VOT production. *Journal of Phonetics*, *41*(6), 407–420. https://doi.org/10.1016/j.wocn.2013.07.005

Olson, D. J. (2016a). The impact of code-switching, language context, and language dominance on suprasegmental phonetics: Evidence for the role of predictability. *International Journal of Bilingualism*, *20*(4), 453–472. https://doi.org/10.1177/1367006914566204

Olson, D. J. (2016b). The role of code-switching and language context in bilingual phonetic transfer. *Journal of the International Phonetic Association*, *46*(3), 263–285. https://doi.org/10.1017/S0025100315000468

Olson, D. J. (2019). Phonological processes across word and language boundaries: Evidence from code-switching. *Journal of Phonetics*, *77*, 100937. https://doi.org/10.1016/j.wocn.2019.100937

Olynyk, M., D'Anglejan, A., & Sankoff, D. (1987). A quantitative and qualitative analysis of speech markers in the native and second language speech of bilinguals. *Applied Psycholinguistics*, *8*(2), 121–136. https://doi.org/10.1017/S0142716400000163

Pang, J., & Rose, P. (2012). Likelihood ratio-based forensic voice comparison with the Cantonese diphthong /ei/ F-pattern. In F. Cox, K. Demuth, S. Lin, K. Miles, S. Palethorpe, J. Shaw, & I. Yuen (Eds.), *Proceedings of the 14th Australasian International Conference on Speech Science and Technology* (pp. 205–208). Australasian Speech Science & Technology Association Inc.

Park, H. (2013). Detecting foreign accent in monosyllables: The role of L1 phonotactics. *Journal of Phonetics*, *41*(2), 78–87. https://doi.org/10.1016/j.wocn.2012.11.001

Patil, H. A., & Basu, T. K. (2008). Development of speech corpora for speaker recognition research and evaluation in Indian languages. *International Journal of Speech Technology*, *11*(1), 17–32. https://doi.org/10.1007/s10772-009-9029-5

Pellegrino, E. (2013). The perception of foreign accented speech. Segmental and suprasegmental features affecting the degree of foreign accent in L2 Italian. In H. Mello, M. Pettorino, & T. Raso (Eds.), *Proceedings of the VIIth GSCP International Confer-*

ence: Speech and Corpora (pp. 261–267). Firenze University Press. https://doi.org/10.36253/978-88-6655-351-9

Pharao, N., Maegaard, M., Møller, J. S., & Kristiansen, T. (2014). Indexical meanings of [s+] among Copenhagen youth: Social perception of a phonetic variant in different prosodic contexts. *Language in Society*, *43*(1), 1–31. https://doi.org/10.1017/S0047404513000857

Picard, M. (2002). The differential substitution of English /θ ð/ in French: The case against underspecification in L2 phonology. *Lingvisticae Investigationes*, *25*(1), 87–96. https://doi.org/10.1075/li.25.1.07pic

Piccinini, P., & Arvaniti, A. (2015). Voice onset time in Spanish–English spontaneous code-switching. *Journal of Phonetics*, *52*, 121–137. https://doi.org/10.1016/j.wocn.2015.07.004

Pingjai, S., Ishihara, S., & Sidwell, P. J. (2013). A Likelihood Ratio–based forensic voice comparison using formant trajectories of Thai diphthongs. *Proceedings of Meetings on Acoustics*, *19*, 060043. https://doi.org/10.1121/1.4799433

Piske, T., MacKay, I. R., & Flege, J. E. (2001). Factors affecting degree of foreign accent in an L2: A review. *Journal of Phonetics*, *29*(2), 191–215. https://doi.org/10.1006/jpho.2001.0134

Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, *35*(4), 655–687. https://doi.org/10.1017/S0272263113000399

Poplack, S., & Meechan, M. (1998). Introduction: How languages fit together in codemixing. *International Journal of Bilingualism*, *2*(2), 127–138. https://doi.org/10.1177/136700699800200201

Poplack, S., Robillard, S., Dion, N., & Paolillo, J. C. (2020). Revisiting phonetic integration in bilingual borrowing. *Language*, *96*(1), 126–159. https://doi.org/10.1353/lan.2020.0004

Potowski, K. (2016). Bilingual youth: Spanish-speakers at the beginning of the 21st century. *Language and Linguistics Compass*, *10*(6), 272–283. https://doi.org/10.1111/lnc3.12192

Przybocki, M. A., Martin, A. F., & Le, A. N. (2007). NIST Speaker Recognition Evaluations utilizing the Mixer Corpora—2004, 2005, 2006. *IEEE Transactions on Audio, Speech and Language Processing*, *15*(7), 1951–1959. https://doi.org/10.1109/TASL.2007.902489

Queen, R. M. (2001). Bilingual intonation patterns: Evidence of language change from Turkish-German bilingual children. *Language in Society*, *30*(1), 55–80. https://doi.org/10.1017/S0047404501001038

Quené, H., Orr, R., & van Leeuwen, D. (2017). Phonetic similarity of /s/ in native and second language: Individual differences in learning curves. *The Journal of the Acoustical Society of America*, *142*(6), EL519–EL524. https://doi.org/10.1121/1.5013149

R Core Team. (2018). *R: A language and environment for statistical computing*. https://www.R-project.org/

Rasier, L., & Hiligsmann, P. (2007). Prosodic transfer from L1 to L2: Theoretical and methodological issues. *Nouveaux cahiers de linguistique française*, *28*, 41–66.

RCMP. (2016). *Voice ID Database* [Unpublished data set; Royal Canadian Mounted Police; Collected at University of Ottawa].

Recasens, D., & Espinosa, A. (2009). An articulatory investigation of lingual coarticulatory resistance and aggressiveness for consonants and vowels in Catalan. *The Journal of the Acoustical Society of America*, *125*(4), 2288–2298. https://doi.org/10.1121/1.3089222

Recasens, D., Pallarès, M. D., & Fontdevila, J. (1997). A model of lingual coarticulation based on articulatory constraints. *The Journal of the Acoustical Society of America*, *102*(1), 544–561. https://doi.org/10.1121/1.419727

Reidy, P. F. (2016). Spectral dynamics of sibilant fricatives are contrastive and language specific. *The Journal of the Acoustical Society of America*, *140*(4), 2518–2529. https://doi.org/10.1121/1.4964510

Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, *10*(1–3), 19–41. https://doi.org/10.1006/dspr.1999.0361

Rhodes, R. (2012). *Assessing the strength of non-contemporaneous forensic speech evidence* [Doctoral thesis, University of York]. White Rose eTheses Online. https://etheses.whiterose.ac.uk/3935/

Rindal, U. (2010). Constructing identity with L2: Pronunciation and attitudes among Norwegian learners of English. *Journal of Sociolinguistics*, *14*(2), 240–261. https://doi.org/10.1111/j.1467-9841.2010.00442.x

Robertson, B., & Vignaux, G. A. (1995). *Interpreting evidence: Evaluating forensic science in the courtroom*. J. Wiley.

Rogers, H. (1998). Foreign accent in voice discrimination: A case study. *International Journal of Speech, Language and the Law*, *5*(2), 203–208. https://doi.org/10.1558/sll.1998.5.2.203

Rose, P. (2002). *Forensic speaker identification*. Taylor & Francis.

Rose, P. (2003). The technical comparison of forensic voice samples. In I. Freckelton & H. Selby (Eds.), *Expert evidence* (Ch. 99). Thompson Lawbook Co.

Rose, P. (2007). *Going and getting it - Forensic speaker recognition from the perpsective of a traditional practitioner-researcher* [Paper presentation]. Australian Research Council Network in Human Communication Science Workshop: FSI not CSI – Perspectives in State-of-the-Art Forensic Speaker Recognition, Sydney, Australia.

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*(1), 139–163. https://doi.org/10.1558/ijsll.v16i1.139

Rose, R. L. (2017). A comparison of form and temporal characteristics of filled pauses in L1 Japanese and L2 English. *Journal of the Phonetic Society of Japan*, *21*(3), 33–40. Retrieved April 21, 2020, from https://doi.org/10.24467/onseikenkyu.21.3_33

Rudy, K., & Yunusova, Y. (2013). The effect of anatomic factors on tongue position variability during consonants. *Journal of Speech, Language, and Hearing Research*, *56*(1), 137–149. https://doi.org/10.1044/1092-4388(2012/11-0218)

Rutter, B. (2011). Acoustic analysis of a sound change in progress: The consonant cluster /stɹ/ in English. *Journal of the International Phonetic Association*, *41*(1), 27–40. https://doi.org/10.1017/S0025100310000307

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. https://doi.org/10.1017/S0025100318000130

San Segundo, E., & Gómez-Vilda, P. (2014). Evaluating the forensic importance of glottal source features through the voice analysis of twins and non-twin siblings. *Language and Law / Linguagem e Direito*, *1*(2), 22–41. https://ojs.letras.up.pt/index.php/LLLD/article/view/2430

San Segundo, E., Tsanas, A., & Gómez-Vilda, P. (2017). Euclidean Distances as measures of speaker similarity including identical twin pairs: A forensic investigation using source and filter voice characteristics. *Forensic Science International*, *270*, 25–38. https://doi.org/10.1016/j.forsciint.2016.11.020

Sankoff, D., Poplack, S., & Vanniarajan, S. (1990). The case of the nonce loan in Tamil. *Language Variation and Change*, *2*(1), 71–101. https://doi.org/10.1017/S0954394500000272

Santiago, F., & Delais-Roussarie, E. (2015). The acquisition of question intonation by Mexican Spanish learners of French. In E. Delais-Roussarie, M. Avanzi, & S. Herment (Eds.), *Prosody and language in contact: L2 acquisition, attrition and lan-*

guages in multilingual situations (pp. 243–270). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-45168-7_12

Schertz, J., Kang, Y., & Han, S. (2019). Cross-language correspondences in the face of change: Phonetic independence versus convergence in two Korean-Mandarin bilingual communities. *International Journal of Bilingualism*, *23*(1), 157–199. https://doi.org/10.1177/1367006917728389

Schleef, E., Meyerhoff, M., & Clark, L. (2011). Teenagers' acquisition of variation: A comparison of locally-born and migrant teens' realisation of English (ing) in Edinburgh and London. *English World-Wide*, *32*(2), 206–236. https://doi.org/10.1075/eww.32.2.04sch

Schneider, E. W. (2003). The dynamics of New Englishes: From identity construction to dialect birth. *Language*, *79*(2), 233–281. https://doi.org/10.1353/lan.2003.0136

Schneider, E. W. (2007). *Postcolonial English: Varieties around the world*. Cambridge University Press.

Schwab, S., & Goldman, J.-P. (2016). Do speakers show different F0 when they speak in different languages? The case of English, French and German. *Proceedings of Speech Prosody 2016*, 6–10. https://doi.org/10.21437/SpeechProsody.2016-2

Scrucca, L., Fop, M., Murphy, T., Brendan, & Raftery, A., E. (2016). Mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, *8*(1), 289–317. https://doi.org/10.32614/RJ-2016-021

Sewell, A., & Chan, J. (2010). Patterns of variation in the consonantal phonology of Hong Kong English. *English World-Wide*, *31*(2), 138–161. https://doi.org/10.1075/eww.31.2.02sew

Shadle, C. H., & Scully, C. (1995). An articulatory-acoustic-aerodynamic analysis of [s] in VCV sequences. *Journal of Phonetics*, *23*(1–2), 53–66. https://doi.org/10.1016/S0095-4470(95)80032-8

Shapiro, M. (1995). A case of distant assimilation: /str/ ⟶ /ʃtr/. *American Speech*, *70*(1), 101. https://doi.org/10.2307/455876

Shen, C., & Watt, D. (2015). Accent categorisation by lay listeners: Which type of "native ear" works better? *York Papers in Linguistics Series 2*, *14*, 106–131. https://www.york.ac.uk/language/ypl/ypl2/14.html

Skarnitzl, R., & Vaňková, J. (2017). Fundamental frequency statistics for male speakers of Common Czech. *AUC Philologica*, *2017*(3), 7–17. https://doi.org/10.14712/24646830.2017.29

Smorenburg, L., & Heeren, W. (2020). The distribution of speaker information in Dutch fricatives /s/ and /x/ from telephone dialogues. *The Journal of the Acoustical Society of America*, *147*(2), 949–960. https://doi.org/10.1121/10.0000674

Smorenburg, L., & Heeren, W. (2021). Acoustic and speaker variation in Dutch /n/ and /m/ as a function of phonetic context and syllabic position. *The Journal of the Acoustical Society of America*, *150*(2), 979–989. https://doi.org/10.1121/10.0005845

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-vectors: Robust DNN embeddings for speaker recognition. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5329–5333. https://doi.org/10.1109/ICASSP.2018.8461375

Soli, S. D. (1981). Second formants in fricatives: Acoustic consequences of fricative-vowel coarticulation. *The Journal of the Acoustical Society of America*, *70*(4), 976–984. https://doi.org/10.1121/1.387032

Sóskuthy, M. (2017). *Generalised additive mixed models for dynamic analysis in linguistics: A practical introduction. arXiv*. http://arxiv.org/abs/1703.05339

Sóskuthy, M. (2021). Evaluating generalised additive mixed modelling strategies for dynamic speech analysis. *Journal of Phonetics*, *84*, 101017. https://doi.org/10.1016/j.wocn.2020.101017

Sóskuthy, M., Foulkes, P., Hughes, V., & Haddican, B. (2018). Changing words and sounds: The roles of different cognitive units in sound change. *Topics in Cognitive Science*, *10*(4), 787–802. https://doi.org/10.1111/tops.12346

Spinu, L., Kochetov, A., & Lilley, J. (2018). Acoustic classification of Russian plain and palatalized sibilant fricatives: Spectral vs. cepstral measures. *Speech Communication*, *100*, 41–45. https://doi.org/10.1016/j.specom.2018.04.010

Spinu, L., & Lilley, J. (2016). A comparison of cepstral coefficients and spectral moments in the classification of Romanian fricatives. *Journal of Phonetics*, *57*, 40–58. https://doi.org/10.1016/j.wocn.2016.05.002

Statistics Canada. (1992). *1991 Census: Profile of census tracts – Part A*. https://www150.statcan.gc.ca/n1/en/catalogue/95F0171X

Statistics Canada. (2003). *Profile for Census Metropolitan Areas and Census Agglomerations, 2001 Census*. https://www12.statcan.gc.ca/datasets/Index-eng.cfm?Temporal=2001&Theme=-1&VNAMEE=Profile%20of%20Census%20Metropolitan%20Areas/Census%20Agglomerations%20(1709)&GA=-1&S=0&SR=1

Statistics Canada. (2017a). *Boundary Files, 2016 Census* (Catalogue no. 92-160-X). https://www150.statcan.gc.ca/n1/en/catalogue/92-160-X

Statistics Canada. (2017b). *Census in Brief: English–French bilingualism reaches new heights.* https://www12.statcan.gc.ca/census-recensement/2016/as-sa/98-200-x/2016009/98-200-x2016009-eng.cfm

Statistics Canada. (2017c). *Census profile, 2016 Census.* https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/index.cfm?Lang=E

Statistics Canada. (2017d). *Dictionary, Census of Population, 2016.* https://www12.statcan.gc.ca/census-recensement/2016/ref/dict/pop095-eng.cfm

Statistics Canada. (2017e). *Languages reference guide, Census of Population, 2016.* https://www12.statcan.gc.ca/census-recensement/2016/ref/guides/003/98-500-x2016003-eng.cfm

Stefanich, S., & Cabrelli Amaro, J. (2018). Phonological factors of Spanish/English word internal code-switching. In L. López (Ed.), *Issues in Hispanic and Lusophone linguistics* (pp. 195–222). John Benjamins. https://doi.org/10.1075/ihll.19.08ste

Stevens, K. N. (1998). *Acoustic phonetics.* MIT Press.

Stevens, M., & Harrington, J. (2016). The phonetic origins of /s/-retraction: Acoustic and perceptual evidence from Australian English. *Journal of Phonetics*, *58*, 118–134. https://doi.org/10.1016/j.wocn.2016.08.003

Strevens, P. (1960). Spectra of fricative noise in human speech. *Language and Speech*, *3*(1), 32–49. https://doi.org/10.1177/002383096000300105

Stuart-Smith, J. (2007). Empirical evidence for gendered speech production: /s/ in Glaswegian. In J. Cole & J. I. Hualde (Eds.), *Laboratory Phonology 9* (pp. 65–86). Mouton de Gruyter.

Stuart-Smith, J. (2020). Changing perspectives on /s/ and gender over time in Glasgow. *Linguistics Vanguard*, *6*(s1), 20180064. https://doi.org/10.1515/lingvan-2018-0064

Stuart-Smith, J., Sonderegger, M., Macdonald, R., Mielke, J., McAuliffe, M., & Thomas, E. (2019). Large-scale acoustic analysis of dialectal and social factors in English /s/-retraction. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019* (pp. 1273–1277). Australasian Speech Science & Technology Association Inc.

Stuart-Smith, J., Timmins, C., & Wrench, A. (2003). Sex and gender in /s/ in Glaswegian. In M.-J. Solé, D. Recasens, & J. Romero (Eds.), *Proceedings of the 15th International Congress of Phonetic Sciences: 15th ICPhS, Barcelona 3-9 August 2003* (pp. 1851–1854). Causal Productions.

Sullivan, K. P. H., & Schlichting, F. (2000). Speaker discrimination in a foreign language: First language environment, second language learners. *International Journal of Speech, Language and the Law*, *7*(1), 95–112. https://doi.org/10.1558/sll.2000.7.1.95

Sundara, M., Polka, L., & Baum, S. (2006). Production of coronal stops by simultaneous bilingual adults. *Bilingualism: Language and Cognition*, *9*(1), 97–114. https://doi.org/10.1017/S1366728905002403

Sung, C. C. M. (2015). Hong Kong English: Linguistic and sociolinguistic perspectives. *Language and Linguistics Compass*, *9*(6), 256–270. https://doi.org/10.1111/lnc3.12142

Tabain, M. (2001). Variability in fricative production and spectra: Implications for the Hyper- and Hypo-and Quantal Theories of speech production. *Language and Speech*, *44*(1), 57–93. https://doi.org/10.1177/00238309010440010301

te Grotenhuis, M., Pelzer, B., Eisinga, R., Nieuwenhuis, R., Schmidt-Catran, A., & Konig, R. (2017). When size matters: Advantages of weighted effect coding in observational studies. *International Journal of Public Health*, *62*(1), 163–167. https://doi.org/10.1007/s00038-016-0901-1

Thibeault, M., Ménard, L., Baum, S. R., Richard, G., & McFarland, D. H. (2011). Articulatory and acoustic adaptation to palatal perturbation. *The Journal of the Acoustical Society of America*, *129*(4), 2112–2120. https://doi.org/10.1121/1.3557030

Thompson, W. C., & Newman, E. J. (2015). Lay understanding of forensic statistics: Evaluation of random match probabilities, likelihood ratios, and verbal equivalents. *Law and Human Behavior*, *39*(4), 332–349. https://doi.org/10.1037/lhb0000134

Toda, M. (2009). *Etude articulatoire et acoustique des fricatives sibilantes* [Doctoral dissertation, Université de la Sorbonne Nouvelle - Paris III]. TEL. https://tel.archives-ouvertes.fr/tel-00448814

Tomić, K. (2017). Temporal Parameters of Spontaneous Speech in Forensic Speaker Identification in Case of Language Mismatch: Serbian as L1 and English as L2. *Comparative Legilinguistics*, *32*, 117–144. https://doi.org/10.14746/cl.2017.32.5

Tomić, K., & French, P. (2019). *Long-term formant frequencies as discriminants in cross-language FVC under LR framework* [Oral presentation]. 28th Annual Conference of the International Association for Forensic Phonetics and Acoustics, Istanbul, Turkey.

Tsui, R. K.-Y., Tong, X., & Chan, C. S. K. (2019). Impact of language dominance on phonetic transfer in Cantonese–English bilingual language switching. *Applied Psycholinguistics*, *40*(1), 29–58. https://doi.org/10.1017/S0142716418000449

Tucker, G. R. (1998). A global perspective on multilingualism and multilingual education. In J. Cenoz & F. Genesee (Eds.), *Beyond bilingualism: Multilingualism and multilingual education* (pp. 3–15). Multilingual Matters.

Turk, A., Nakai, S., & Sugahara, M. (2006). Acoustic segment durations in prosodic research: A practical guide. In S. Sudhoff, D. Lenertova, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, N. Richter, & J. Schließer (Eds.), *Methods in empirical prosody research* (pp. 1–28). De Gruyter. https://doi.org/10.1515/9783110914641.1

Ulbrich, C., & Mennen, I. (2016). When prosody kicks in: The intricate interplay between segments and prosody in perceptions of foreign accent. *International Journal of Bilingualism*, *20*(5), 522–549. https://doi.org/10.1177/1367006915572383

Vaillancourt, V., Laroche, C., Mayer, C., Basque, C., Nali, M., Eriks-Brophy, A., Soli, S. D., & Giguère, C. (2005). Adaptation of the HINT (hearing in noise test) for adult Canadian Francophone populations. *International Journal of Audiology*, *44*(6), 358–361. https://doi.org/10.1080/14992020500060875

van den Heuvel, H. (1996). *Speaker variability in acoustic properties of Dutch phoneme realisations* [Doctoral dissertation, Katholieke Universiteit Nijmegen]. Radboud Repository. http://hdl.handle.net/2066/76416

van Rij, J., Wieling, M., Baayen, R. H., & van Rijn, H. (2017). *itsadug: Interpreting time series and autocorrelated data using GAMMs* (Version 2.3) [Computer software]. https://CRAN.R-project.org/package=itsadug

Vaughn, C., Baese-Berk, M., & Idemaru, K. (2019). Re-examining phonetic variability in native and non-native speech. *Phonetica*, *76*(5), 327–358. https://doi.org/10.1159/000487269

Vergeer, P., van Es, A., de Jongh, A., Alberink, I., & Stoel, R. (2016). Numerical likelihood ratios outputted by LR systems are often based on extrapolation: When to stop extrapolating? *Science & Justice*, *56*(6), 482–491. https://doi.org/10.1016/j.scijus.2016.06.003

Vieru-Dimulescu, B., & Boula de Mareüil, P. (2006). Perceptual identification and phonetic analysis of 6 foreign accents in French. *Proceedings of Interspeech 2006*, 441–444. https://doi.org/10.21437/Interspeech.2006-142

Wade, T., Jongman, A., & Sereno, J. (2007). Effects of acoustic variability in the perceptual learning of non-native-accented speech sounds. *Phonetica*, *64*(2–3), 122–144. https://doi.org/10.1159/000107913

Walker, D. C. (1984). *The pronunciation of Canadian French*. University of Ottawa Press.

Wang, B. X., & Hughes, V. (2021). System performance as a function of calibration methods, sample size and sampling variability in likelihood ratio-based forensic voice comparison. *Proceedings of Interspeech 2021*, 381–385. https://doi.org/10.21437/Interspeech.2021-267

Wang, B. X., Hughes, V., & Foulkes, P. (2019a). Effect of score sampling on system stability in likelihood ratio based forensic voice comparison. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019* (pp. 3065–3069). Australasian Speech Science & Technology Association Inc.

Wang, B. X., Hughes, V., & Foulkes, P. (2019b). The effect of speaker sampling in likelihood ratio based forensic voice comparison. *International Journal of Speech, Language and the Law*, *26*(1), 97–120. https://doi.org/10.1558/ijsll.38046

Watermeyer, S. (1996). Afrikaans English. In V. de Klerk (Ed.), *Focus on South Africa* (pp. 99–124). John Benjamins. https://doi.org/10.1075/veaw.g15.08wat

Watson, C. I., & Harrington, J. (1999). Acoustic evidence for dynamic formant trajectories in Australian English vowels. *The Journal of the Acoustical Society of America*, *106*(1), 458–468. https://doi.org/10.1121/1.427069

Weinrich, U. (1953). *Languages in contact: Findings and problems*. Linguistic Circle.

Wester, F., Gilbers, D., & Lowie, W. (2007). Substitution of dental fricatives in English by Dutch L2 speakers. *Language Sciences*, *29*(2-3), 477–491. https://doi.org/10.1016/j.langsci.2006.12.029

Westfall, P. H. (2014). Kurtosis as peakedness, 1905–2014. *R.I.P. The American Statistician*, *68*(3), 191–195. https://doi.org/10.1080/00031305.2014.917055

Wiese, R. (1984). Language production in foreign and native languages: Same or different? In H. W. Dechert, D. Möhle, & M. Raupach (Eds.), *Second Language Production* (pp. 11–25). Gunter Narr Verlag.

Wilbanks, E. (2017). Social and structural constraints on a phonetically-motivated change in progress: (str) retraction in Raleigh, NC. *University of Pennsylvania Working Papers in Linguistics*, *23*(1), Article 33. https://repository.upenn.edu/pwpl/vol23/iss1/33

Wilson, I., & Gick, B. (2014). Bilinguals use language-specific articulatory settings. *Journal of Speech, Language, and Hearing Research*, *57*(2), 361–373. https://doi.org/10.1044/2013_JSLHR-S-12-0345

Wioland, F. (1985). *Les structures syllabiques du français: Fréquence et distribution des phonèmes consonantiques, contraintes idiomatiques dans les séquences consonantiques*. Champion.

Witteman, M. J., Weber, A., & McQueen, J. M. (2014). Tolerance for inconsistency in foreign-accented speech. *Psychonomic Bulletin & Review*, *21*(2), 512–519. https://doi.org/10.3758/s13423-013-0519-8

Wong, S. G.-J., & Papp, V. (2018). *Transferability of non-lexical hesitation markers across languages: Evidence from te reo Māori-English bilinguals* [Oral presentation]. 27th International Association for Forensic Phonetics and Acoustics Annual Conference, Huddersfield, UK.

Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed.). Chapman; Hall/CRC. https://doi.org/10.1201/9781315370279

Wrench, A. A. (1995). Analysis of fricatives using multiple centres of gravity. In K. Elenius & P. Branderud (Eds.), *Proceedings of the 13th International Congress of Phonetic Sciences* (pp. IV.460–IV.463). KTH & Stockholm University.

Xie, X., & Jaeger, T. F. (2020). Comparing non-native and native speech: Are L2 productions more variable? *The Journal of the Acoustical Society of America*, *147*(5), 3322–3347. https://doi.org/10.1121/10.0001141

Yim, A. C. S., & Rose, P. (2012). Are nasals better? Likelihood ratio-based forensic voice comparison with segmental cepstra from Cantonese and Japanese syllabic/mora nasals. In F. Cox, K. Demuth, S. Lin, K. Miles, S. Palethorpe, J. Shaw, & I. Yuen (Eds.), *Proceedings of the 14th Australasian International Conference on Speech Science and Technology* (pp. 5–8). Australasian Speech Science & Technology Association Inc.

Yu, A. C. L. (2016). Vowel-dependent variation in Cantonese /s/ from an individual-difference perspective. *The Journal of the Acoustical Society of America*, *139*(4), 1672–1690. https://doi.org/10.1121/1.4944992

Zampini, M. L. (2008). L2 speech production research: Findings, issues, and advances. In J. G. Hansen Edwards & M. L. Zampini (Eds.), *Phonology and second language acquisition* (pp. 219–249). John Benjamins. https://doi.org/10.1075/sibil.36.11zam

Zhang, C., & Enzinger, E. (2013). Fusion of multiple formant-trajectory- and fundamental-frequency-based forensic-voice-comparison systems: Chinese /ei1/, /ai2/, and /iau1/, 060044. https://doi.org/10.1121/1.4798793

Zhang, Y., He, L., & Dellwo, V. (2019). Speaker individuality in the durational characteristics of voiced intervals: The case of Chinese bi-dialectal speakers. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019* (pp. 3075–3079). Australasian Speech Science & Technology Association Inc.

Zsiga, E. C. (2000). Phonetic alignment constraints: Consonant overlap and palatalization in English and Russian. *Journal of Phonetics*, *28*(1), 69–102. https://doi.org/10.1006/jpho.2000.0109