UNIVERSITY OF NIŠ
FACULTY OF PHILOSOPHY

Kristina D. Tomić

VOICE QUALITY IN CROSS-LANGUAGE FORENSIC SPEAKER COMPARISON

DOCTORAL DISSERTATION

Niš, 2024

UNIVERSITY OF NIŠ
FACULTY OF PHILOSOPHY

Kristina D. Tomić

VOICE QUALITY IN CROSS-LANGUAGE FORENSIC SPEAKER COMPARISON

DOCTORAL DISSERTATION

Niš, 2024

УНИВЕРЗИТЕТ У НИШУ
ФИЛОЗОФСКИ ФАКУЛТЕТ

Кристина Д. Томић

КВАЛИТЕТ ГЛАСА У МЕЂУЈЕЗИЧНОЈ ФОРЕНЗИЧКОЈ КОМПАРАЦИЈИ ГОВОРНИКА

ДОКТОРСКА ДИСЕРТАЦИЈА

Ниш, 2024.

Data on Doctoral Dissertation

Doctoral Supervisor:	Professor Tatjana Paunović, PhD, full professor, University of Niš, Faculty of Philosophy
Title:	Voice Quality in Cross-language Forensic Speaker Comparison
Abstract:	The aim of the present research is to test the hypothesis that the anatomy of the vocal tract outweighs the linguistic factors in determining individual voice quality for the purpose of confirming that the acoustical measures of VQ could be used in cross-language forensic speaker comparison. The study consists of two perceptual experiments and evaluation of acoustic correlates of VQ under the likelihood ratio framework in same-language and cross-language sample pairs. The corpus for the research was created by recording fifty native speakers of Serbian while speaking Serbian and English over a mobile phone. In the first experiment, four expert listeners rated twenty speakers according to a truncated version of the Vocal Profile Analysis protocol (Laver et al., 1981). The results showed that within-speaker vocal profiles across Serbian and English exhibit lower variability than between-speaker profiles within each language, respectively. In addition, it was found that phonatory settings are more responsible for vocal profile similarity than articulatory settings. In the second experiment, sixty Serbian naïve listeners performed speaker discrimination and assessed voice similarity of same- and different-speaker pairs within and across languages. The study confirmed the existence of the "language familiarity effect" as listeners were able to discriminate speakers with higher accuracy in the same-language than in the cross-language context (92.68% vs 86.22%). Furthermore, while same-speaker samples were rated as slightly less similar in the cross-language context (8.45 vs 8.79), different-speaker samples have a notably higher similarity score in the language-mismatching than in the

language-matching condition (4.43 vs 3.61). Likelihood ratio assessment revealed that among the best performing parameters in cross-language comparison are CPP, HNR35, HNR25, H1*-A3*, LTF3 and H1*H2*, with EER ranging between 20%-24% and Cllr between 0.59 and 0.69. In combination, articulatory and phonatory parameters achieved an EER of 1.59% and Cllr of 0.06 in the cross-language context.

Scientific Field:	Philological Sciences
Scientific Discipline:	Linguistics, Phonetics, Forensic Phonetics
Key Words:	voice quality, VPA, phonatory settings, articulatory settings, long-term formants, creaky voice, breathy voice, speaker comparison, likelihood ratio, forensic phonetics
UDC:	811.111'342.1:811.163.41'342.1
CERIF Classification:	H 004 Philology, H 350 Linguistics, H 351 Phonetics, Phonology

Подаци о докторској дисертацији

Ментор:	Проф др Татјана Пауновић, редовни професор, Универзитет у Нишу, Филозофски факултет
Наслов:	Voice Quality in Cross-language Forensic Speaker Comparison
Резиме:	Циљ овог истраживања је да испита хипотезу да је анатомија говорног тракта одговорна за квалитет гласа појединца у већој мери него језик којим се говори, а са циљем да потврди да се акустички корелати квалитета гласа могу користити као параметри у међујезичној форензичкој компарацији говорника. Студија се састоји од два експеримента слушања и од евалуације акустичких корелата квалитета гласа под теоријским оквиром вероватноће у истојезичним и међујезичним паровима узорака. Корпус за истраживање настао је снимањем педесет изворних говорника српског језика док говоре на српском и енглеском преко мобилног телефона. У првом експерименту, четири експерта је оценило квалитет гласа двадесет говорника према протоколу за анализу гласовног профила (Laver et al., 1981). Резултати су показали да су интраспикерски гласовни профили сличнији у два језика него инерспикерски профили у било ком од језика појединачно. Такође, потврђено је да је фонација одговорна за сличност гласовног профила у већој мери него артикулација. У другом експерименту, шездесет појединца којима је матерњи језик српски вршило је лаичко препознавање говорника и оцењивало сличност гласова истих и различитих говорника у истојезичним и међујезичним паровима. Потврђено је постојање такозваног „ефекта познавања језика" с обзиром да је проценат препознавања био већи у истојезичним паровима (92,68% наспрам 86,22%). Такође, док су узорци истих говорника оцењени као мање слични у међујезичном контексту (8,45 наспрам 8,79),

узорци различитих говорника имају приметно већу оцену у међујезичном него у истојезичном контексту (4,43 наспрам 3,61). Евалуација односа вероватноћа показала је да су параметри са најбољим резултатима у међујезичној компарацији CPP, HNR_{35}, HNR_{25}, H1*-A3*, LTF3 и H1*H2*, за које је степен исте грешке у опсегу од 20% до 24% а C_{llr} између 0,59 и 0,69. Комбиновано, параметри артикулације и фонације достижу степен исте грешке од 1,59% и C_{llr} од 0,06 у међујезичном контексту.

Научна област: Хуманистичке науке, Филологија

Научна дисциплина: Лингвистика, Фонетика, Форензичка фонетика

Кључне речи: квалитет гласа, анализа гласовног профила, фонација, артикулација, дугорочни форманти, шкрипав глас, задихан глас, компарација говорника, вероватноћа, форензичка фонетика

УДК: 811.111'342.1:811.163.41'342.1

CERIF класификација: H 004 Филологија, H 350 Лингвистика, H 351 Фонетика, фонологија

"No man is an island",
especially when writing a PhD thesis.

"The jaw can be opened or dropped with much less strength that it can be closed.
We must work to keep our mouth shut. It opens quite easily."
Van Riper and lrwin (1958: p. 360)

Table of Contents

Acknowledgement

Now that this long journey has ended, I would like to express my sincere gratitude to all the people who have helped me realize this comprehensive project and who have not stopped encouraging me to go through with it; and there were times when I was on the verge of giving up.

First and foremost, I owe my deepest gratitude to two people who are not only marvellous professors and reputable experts in their scientific areas but also admirable, good-hearted souls, my academic parents, one of them the supervisor and the other advisor on this dissertation, Professor Tatjana Paunović at the University of Niš, and Professor Emeritus Peter French at the University of York and founder of JP French Associates, now JP French International, a forensic phonetic and acoustic laboratory in York, UK.

Eight years ago, when I approached Professor Peter French with the proposal to explore cross-language speaker comparison, he immediately recognized the potential of taking this direction in forensic research. His suggestions, research ideas and aid largely influenced the final form of the dissertation. In addition, he never stopped believing in my capability to complete the work, even at times when I stopped believing in myself. His encouragement, support and guidance gave me the strength and motivation to continue and conclude this project. And for all of that, I cannot thank him enough.

Indeed, I would have never even taken an interest in phonetics, and science in general, had it not been for my teacher and supervisor, Professor Tatjana Paunović, whom I met fifteen years ago on my first day as a freshman at the University. Not only has she taught me about the English language, phonetics, and academic writing, but also how to be an understanding and compassionate teacher and treat every student with respect. Her invaluable advice throughout my academic education and the writing of this dissertation has given me confidence and helped me become a more skilled researcher. I would also like to thank her for taking a direct part in the study as an expert in scoring the language competence of the participants.

A special thank you goes to my colleagues, linguists, Dr Katharina Klug and Dr Jessica Wormwald at the Department of Language and Linguistic Science, University of York, Prof Radek Skarnitzl at the Institute of Phonetics, Faculty of Arts, Charles University Prague and Professor Dominic Watt at the University of York, UK and JP French International, Zurich, Switzerland, for participating as expert listeners on scoring the voice quality settings. I would also like to thank my friends and colleagues, teachers Milena Videnović and Nevena Mitić for

participating as ESL experts in the dissertation and Dr Katarina Milenković, a lecturer of English language at the Faculty of Sciences and Mathematics, University of Niš, and my loyal friend, for participating as an expert on scoring the language competence of the speakers and helping me spread the word about the research and recruit more participants. Tackling a large corpus as mine in semi-automatic analysis is not a piece of cake; therefore. I am very grateful to my sister Veronika Tomić and my assisstant Zorana Bogdanović, a student at the Faculty of Philosophy, University of Niš, for helping with manual correction of vowel boundaries and formant paths. Without you the project would have taken many more daunting months.

As participant recruitment turned out to be one of the greatest challenges in completing the thesis, I would like to thank everyone who has contributed to my quest: Katarina Mitić Ivković, a teaching assistant at the Faculty of Medicine, University of Niš, Gordana Ignjatović, a lecturer of English language at the Faculty of Law, University of Niš, Marija Nešić, a lecturer of English language at the Singidunum University, Niš, Professor Jasmina Đorđević, Dr Petra Mitić, Ivana Šorgić and Nikola Tatar at Centre for Foreign Languages, Faculty of Philosophy, University of Niš, Neda Necić from the Department of Journalism and Communication, and other colleagues from the Department of English Language and Literature and Department of Serbian Language and Literature, University of Niš, as well as all of the students, more than 100 study participants – without you this project would not have been possible.

I would be remiss not to mention the cooperation with the School with Student Dormitory "Bubanj" in Niš, and their kind teacher Marija Pavlović, who let me use their audiometry booth for the recording of the part of the corpus, which, due to unfortunate circumstances of the world pandemics, did not end as part of this dissertation. Moreover, I would like to thank Professor Francis Nolan at the University of Cambridge for being so kind as to let me have his and Professor Laver's recordings of reproduction of distinct voice qualities exemplifying the categories from Laver's (1980) framework – the recordings have helped me learn and differentiate the voice quality settings, which otherwise I would have had to understand solely on the basis of the available reading material. I am also very grateful to Dr Justin Lo and Dr Elliot Holmes at the University of York for taking their time to discuss their projects with me and share the script for likelihood ratio calculations. Finally, I want to mention two people whose encouragement in my early research projects made a difference and propelled me into the vast area of science. You helped me realise that not only is forensic phonetics what I would like to do for a living but also that my research is relevant and important for the field – thank you, Professor Zorka Kašić and Professor Boban Arsenijević.

Among the last but not least significant people who have contributed to this thesis, I would like to mention my husband and colleague, Nikola Janković. Not only has he offered unmeasurable emotional support throughout these years, read my drafts and discussed every aspect of the research with me, prepared numerous cups of coffee and squeezed orange juice, followed me to numerous linguistic conferences and helped debug Praat scripts, but he has also directly contributed to the dissertation by developing online platforms for the conduction of the three listening experiments. This dissertation and my career would not have been possible without you – thank you.

Finally, everything I have achieved so far, professionally and personally, is due to my loving family, my parents and my sister, and my supportive friends. Thank you for believing in me and not sparing words to show it. I know this feels like the end of an epoch in my life, but the truth is that it is only the beginning, and I will do my best to make you all proud.

Ethical Declaration

The research and experiments in this doctoral dissertation were approved by the Ethical Committee of the Faculty of Philosophy, University of Niš by *Decision* issued on November 2nd 2021, in Niš. All participants in the present research are volunteers who gave informed consent for the results obtained from their data to be used in this dissertation, shared at linguistic conferences, and published in relevant scientific publications. The participants' personal information is confidential, and the data is presented anonymously so that their identity cannot be inferred or otherwise revealed. The experiments did not pose any health risks to the participants.

List of Abbreviations

CI	correct identification
CL	cross-language
DS	different-speaker
EER	equal error rate
ER	error rate
FA	false alarm / false acceptance
FR	false rejection
FSC	Forensic Speaker Comparison
FVC	Forensic Voice Comparison
LLR	log-likelihood ratio
LR	likelihood ratio
LTF	long-term formant
MH	missed hit
SL	same-language
SS	same-speaker
VPA	Vocal Profile Analysis
VQ	voice quality

List of Tables

List of Figures

1. Introduction

1.1. Defining the Problem

Forensic speaker comparison involves comparing a speech sample of an unknown speaker with a sample of a known one in order to assist the courts in determining whether the same speaker uttered them (French, 2017; Morrison, 2009; Nolan, 2007). There is a decades-long practice of comparing voices with outstanding reliability when both samples are in the same language; however, comparing speech samples in different languages still poses many challenges. Cross-language forensic speaker comparison was strongly discouraged at the beginning of the century since "not enough [was][1] known yet about bilingual speakers to say whether any voice quality remains the same across two samples of the same speaker speaking in two different languages or dialects" (Rose, 2002, p. 342). In addition, Article 3.10 in the Code of Practice of the International Association of Forensic Phonetics and Acoustics states that "members should exercise particular caution with cross-language comparisons" (IAFPA, 2020). However, globalisation and increased mobility of the world's population have contributed to the rising number of multilingual speakers. As a result, forensic speech scientists have begun to encounter cases requiring them to compare samples in different languages (Künzel, 2013; Milne et al., 2019). Therefore, if such work is to be undertaken, there is an obvious need for structured research to identify which parameters perform well in these circumstances.

Bearing in mind that traditional forensic phonetic parameters (such as fundamental frequency, pitch range, and segmental features) could be incomparable in different languages (Rose, 2002), Köster et al. (2007: p. 1845) suggested that experts should rely on paralinguistic and extralinguistic components of speech. Namely, similarities in speech are more likely to be retained once we remove language-conditioned variation from the equation. Thus, the scientists focused on testing parameters related to speaker habits that are not necessarily linguistically conditioned, such as temporal parameters (Tomić, 2017), or aspects of speech conditioned by the anatomy of the human vocal tract, such as long-term formant frequencies (Asiaee et al., 2019; Heeren et al., 2014; Krebs & Braun, 2015; Meuwly et al., 2015; Tomić & French, 2019; Lo, 2021).

[1] Changes marked by square brackets were made by the author.

One of the features that, at least partially, depend on the human vocal tract geometry and dimensions is voice quality. Voice quality in the broad sense is described as the cumulative effect of laryngeal and supralaryngeal characteristics of human speech, which make a voice recognisable and specific to every individual (Laver, 1980); thus, it is not surprising that this feature is described as "most useful for discriminating speakers" by a large number of forensic practitioners (Gold & French, 2011, p. 302). Two comprehensive surveys on forensic speech science practices revealed that the perceptual analysis of voice quality had been widely employed not only in forensic speaker comparison (Gold & French, 2011) but also in speaker profiling, voice line-ups, and other forensic tasks (San Segundo, 2021). Laver (1994) defines voice quality as an extralinguistic feature, an index of someone's speaking habit and the nature of their vocal apparatus, rather than a bearer of communicative information (p. 22-23). As such, it presents a potentially useful parameter in cross-language forensic speaker comparison. A simple thought experiment provides support in favour of this hypothesis. Imagine a professor talking to a group of familiar students in English without being able to see them. During the conversation, everyone switches to a foreign language that all in the class understand. Would the professor still be able to recognise the students' voices? Most likely, the answer to this hypothetical question is "yes" because even though their pitch, intonation and pronunciation of segments may change, something in their voice will remain "recognisable and specific".

Let us now take another thought experiment into consideration. A colleague of yours is an English-French simultaneous bilingual but you never knew of their French origin. The colleague decides to prank you and calls you on the phone from an unknown number, speaking French with the competence of a native speaker. How likely would you recognise their voice instantly, if at all? A plethora of research has explored sociophonetic aspects of voice, and it has already been shown that different languages, or even different dialects or social groups, are characterised by specific voice qualities (Esling, & Wong, 1983). For instance, as early as 1964, Honikman noted that French speakers exhibit considerable mobility of the lips and jaw as opposed to English speakers, as well as that they have a lowered tongue and prominent lip rounding. Slavic languages are characterised as palatalised (Honikman, 1964), while Spanish has prominent tongue-fronting (Cruttenden, 2014: p 302). Speakers of Danish and Dutch may be described as having a breathy voice (Cruttenden, 2014), whereas many accents of English are often characterised by creakiness (Hewlett & Beck, 2006: p. 275, 277; Stuart-Smith, 1999). The results of such research contradict the thesis that voice quality can be used in cross-language forensic speaker comparison.

In the present study, we set out to understand whether the biological, that is, anatomical factors outweigh the acquired, language-specific, factors in defining one's voice quality. This research question is yet another fragment of the nature-nurture dichotomy and it will help us understand how reliably voice quality can be employed in forensic speaker comparison in the conditions of a language mismatch. We will explore the hypothesis by identifying which features of voice remain the same despite the language switch and whether they depend on foreign language proficiency. The research is performed on the samples of native Serbian and English as a foreign language through a range of perceptual and acoustic experiments. Prior to outlining the research and experiment design, we will briefly discuss the issues related to the relevant terminology.

1.2. Terminology Disambiguation

Forensic speaker comparison (FSC) has existed as a discipline within forensic sciences for almost half a century; however, the terminology used to describe this process and its definitions have changed throughout the years. Towards the end of the 20th and at the beginning of the 21st century, the term *forensic speaker identification* (FSI) was used to describe the process of comparing the identity of speakers of a known and unknown speech sample (French, 1994; Hollien, 1990; 2002; Jovičić, 2001; Kašić & Đorđević, 2009; Nolan, 1999; Rose, 2002). However, some experts underline that the usage of this term is inappropriate, bearing in mind that forensic speech scientists the not perform the determination of speakers' identities; instead, they merely compare the speakers' voices and provide the information to the court or relevant authorities to help in a complex process of speaker identification (Nolan, 2007; Morrison, 2009). In recent literature, the term *forensic voice comparison* (FVC) can also be encountered for the same procedure (*e.g.* Jessen, 2018; Morrison & Enzinger, 2019; Morrison et al., 2021; Rose, 2010; 2013; 2013b). Whereas the semantics of these two terms may imply some differences, whereby *forensic speaker comparison* may have a broader meaning than *forensic voice comparison* (see Rose & Morrison, 2009), in the present study primarily dealing with voice quality, the two terms will be used interchangeably.

C*ross-language* forensic speaker comparison is one of the terms used to refer to the forensic comparison of speech samples that are in different languages (*e.g.* Tomić & French, 2019; IAFPA, 2020; Künzel, 2013); other terms that could be encountered in the literature include *cross-linguistic* (e.g. de Boer & Heeren, 2020; Zhong, 2019) or *cross-lingual* (e.g. Askar et al., 2015) forensic speaker comparison, forensic speaker comparison *in language mismatch* (e.g. Drygajlo et al., 2015; ENFSI, 2001; Tomić, 2017) or *of bilingual speakers* (e.g.

Armbrecht, 2015; Asiaee et al., 2019; Cho & Munro, 2017; Dorreen, 2017; Heeren et al., 2014; Krebs & Braun, 2015; Lo, 2021). While language mismatch is more of an expression than a term describing, appropriately, conditions of voice comparison, we argue that *cross-linguistic* FSC should not be used to denote FSC in which the known sample and questioned sample are in different languages. Namely, the term has a different meaning: it implies comparative/contrastive research or outcomes of the analyses obtained across different linguistic systems, not necessarily involving samples in different languages (for instance, practices/results of FSC in the English language as opposed to FSC in Serbian). On the other hand, as the term *bilingual* traditionally implies a person who uses two or more languages on a regular basis (Grosjean, 1982: p. 1), while not incorrect, it is not appropriate for the present study in which the participants are foreign language learners who do not necessarily communicate in their second language daily. Therefore, this paper will maintain a difference between bilinguals and foreign language learners and use the term *cross-language forensic speaker comparison* to refer to FSC in language mismatched conditions.

1.3. Research Goals

The current research has both theoretical and applied significance. In the theoretical sense, the goal of the study is to explore the language effect on the voice quality of individual speakers across languages. In the applied sense, the aim is to investigate to what extent the acoustical measures of voice quality can be used in cross-language forensic speaker comparison.

The study commences with the hypothesis that anatomy of the vocal tract outweighs the sociolinguistic factors in determining individual voice quality. The hypothesis is tested within the forensic-phonetic theoretical framework by answering the following primary questions:

- How similar are the voices of same/different speakers when speaking Serbian (L1) and English (L2)?
- What is the effect of language mismatch on forensic speaker comparison using the acoustic correlates of voice quality with Serbian (L1) and English (L2) samples?
- How does foreign language proficiency/fluency affect voice perception and cross-language forensic speaker comparison?

The perceptual aspect of the study can be analysed through two perspectives, expert listening by trained phoneticians, who would be able to score the prominence of specific voice

quality features across languages, and the naïve listeners, who would assign holistic, impressionistic values to the pairs of voices. The advantages of the former approach are (1) that it will assist in identifying specific, measurable voice quality settings that are noted to differ or remain the same across languages and could, thus, be used for the selection of appropriate acoustic parameters for forensic speaker comparison and (2) that it can be replicated in forensic speaker comparison casework. The latter approach provides the subjective perspective on the issue, appealing to our "common sense" hypothesis that if an untrained, lay listener can recognise a person's voice in a foreign language, then there indeed are speaker-specific features of voice that outweigh the language effect in speaker recognition.

To estimate the effect of language in cross-language forensic speaker comparison, we will first compare the same-language samples using the same parameters under the selected methodological framework. The framework chosen for this study is the Bayesian Likelihood Ratio, which is in accordance with contemporary forensic tendencies of probabilistic expression of the outcome. The scores obtained through comparisons in two contexts (single-language and cross-language) will subsequently be contrasted and analysed further to estimate the effect of language mismatch on the performance of the entire system and individual speakers, respectively. Given the lack of forensic speaker comparison research in Serbian, the significance of the present study is also reflected in the assessment of voice quality parameters in single-language comparisons, which would not only act as a control analysis for cross-language comparison but would also contribute to understanding whether there is any universality in these features as discriminants across different linguistic systems.

The effect of foreign language proficiency will be assessed both in the perceptual experiments and in the analysis under the Likelihood Ratio framework. Previous research on cross-language forensic speaker comparison has often detected that the FSC system performance may vary from speaker to speaker (e.g. de Boer & Heeren, 2020; Lo, 2021); however, it has not been explored whether this difference might be a result of the speakers' fluency in the second language. Understanding the effect of foreign language proficiency on forensic speaker comparison would have implications for both voice quality theory and FSC practice.

Finally, since this is exploratory research, as the relevant literature is reviewed, the issues that arise along the way will be discussed, and the study will include additional secondary questions that will assist in the decision-making process concerning the construction of reference population and other relevant issues. The research questions will be revisited in Chapter 5.1.

1.4. Thesis Outline

Chapters 2, 3 and 4 of this dissertation contain the theoretical background necessary to understand the motivation and methodology for the present study. Chapter 2 introduces the field of Forensic Speech Science, providing a brief historical development of the discipline in the West and in Serbia, and presents the theoretical and methodological base for Forensic Speaker Comparison. The chapter demonstrates the Bayesian Likelihood Ratio framework and discusses the previous research in cross-language forensic speaker comparison. In this chapter, we also explore speaker recognition by naïve listeners, with particular interest in the cross-language recognition.

Chapter 3 primarily concerns the voice quality theory, the perceptual framework of voice quality analysis and its acoustic correlates, as well as its communicative and informative functions. The chapter further discusses previous research on the application of voice quality in forensic speech science and the voice quality in a telephone-transmitted signal. Finally, it reviews the studies that observe how voice quality features vary in bilingual speakers.

Chapter 4, provides a concise overview of the vowel systems of Serbian and English, with the focus on the acquisition of English as a foreign language by native speakers of Serbian. This chapter shall provide the ground for understanding and analysing the effect of language proficiency in cross-language forensic speaker comparison.

In Chapter 5, we revisit the research goals and raise more specific research questions for the present study. Next, we provide details regarding the corpus compilation and participants and discuss the methodology of language proficiency and voice quality scoring. Methodologies of individual experiments will be demonstrated prior to presenting the results in Chapters 6 and 7.

Chapter 6 is divided into two parts, the former comprising the listening experiment with the expert (6.1) and the latter with naïve listeners (6.2). For each of the experiments in this chapter, we elaborate on the utilised methodology, present the results and discuss the findings. An interim discussion concerning both experiments is provided at the end of the chapter.

In Chapter 7, we perform the acoustic analysis and present the results through descriptive and inferential statistics (7.1) as well as through same-language and cross-language Likelihood Ratio comparisons (7.2). The cross-language comparisons are performed with various reference population combinations, and finally, the effect of foreign language fluency is explored.

Chapter 8 summarises the findings from the perceptual, acoustic and Likelihood Ratio experiments and provides a general discussion of the results in the context of Forensic Speaker Comparison. In this chapter, we once again reflect on the research goals and the initial hypothesis, explaining how the present study contributed to the issues in question. Finally, the thesis concludes with prospects for future research in the domain of cross-language Forensic Speaker Comparison.

2. Forensic Speaker Comparison

In the present chapter, we will consider the historical and methodological developments in forensic speech science, addressing the terminological debate and presenting an overview of the practices in Serbia. Further on, we will provide the theoretical background of forensic speaker comparison by exploring between-speaker and within-speaker variability, as well as the fundamentals of the Likelihood Ratio framework and ongoing "paradigm shift" in the field. Finally, we will contemplate the issues related to cross-language forensic speaker comparison, review previous research on cross-language corpora and consider the implications for the present study.

2.1. Forensic Speech Science - Evolution and Practices

2.1.1. Discipline definition and scope

Throughout literature, speech scientists, forensic experts and linguists have defined and redefined the interdisciplinary area between linguistics, speech acoustics and forensic sciences multiple times. As a result, today, there are a plethora of overlapping sub-disciplines, and their scope is described as varied, depending on authors, institutions and scholarly influence under which they have developed and grown.

In the 1980s, Harry Hollien wrote about *Forensic Communication* as an emerging area within Forensic Sciences that encompasses the elements primarily drawn from Phonetics but also Psychoacoustics, Electrical Engineering, Psychology, and Computer Sciences, interfacing with areas of Linguistics, Mechanical Engineering, and Medicine, Otolaryngology and Speech Pathology in particular (Hollien, 1983). A few decades later, Hollien (2012) differentiates between three Forensic Communication sub-disciplines, *Forensic Linguistics*, *Forensic Psychoacoustics*, and *Forensic Phonetics*, whereby the first one analyses spoken or written language to determine authorship, individual intent, deception and speech/language decoding. The second sub-discipline is primarily concerned with perceptual aspects of human hearing and audition, while the third includes tasks such as forensic speaker identification, enhancing and decoding of speech, analysis of voice and emotion, or authentication of recordings. According to Hollien, the tasks of Forensic Linguistics and Forensic Phonetics overlap considerably (Hollien, 2012: p. 27).

It is now widely accepted that the term *Forensic Linguistics* was first introduced by Svartvik (1968) in a case study, where he disputed the authorship of statements in the Evans case by identifying certain stylistic discrepancies by employing quantitative and qualitative

analysis (Olsson, 2008), even though the term *forensic English* was mentioned by Philbrick (1949) 20 years before (Coulthard & Johnson A., 2007). Today, *Forensic Linguistics* is used as an umbrella term for a large area that represents an interface between language, crime and law (Olsson, 2008) and encompasses the study of the language of legal documents, the language of the police and law enforcement, police interviews, courtroom interaction, authorship attribution and plagiarism, trademarks and their protection (Coulthard & Johnson A., 2007).

Kašić and Đorđević (2009) argue whether Forensic Linguistics (referring to both Forensic Phonetics and Linguistics) should be regarded as a skill or an academic discipline. The argument in favour of observing it as a skill is that it draws from the already established knowledge base and applies the theories and findings to forensic purposes. However, throughout the past decades, there has been an abundance of research directed not only at improving forensic linguistic practices but also at uncovering new facts and providing novel insights into the already vast linguistic/phonetic knowledge base, which is why it is fair to say that Forensic Linguistics (Forensic Phonetics included) has become a full-fledged academic discipline (Kašić & Đorđević, 2009: p. 470-471). As a result, nowadays, several accredited postgraduate academic programmes are offering a degree in these disciplines, such as MA programmes in Forensic Linguistics in the UK and USA (Aston University, 2022.; California University, 2022; Cardiff University, 2022; Hofstra University, 2022) or MA in Forensic Phonetics (previously MSc in Forensic Speech Science) at the University of York (The University of York, 2022).

Even though Phonetics is regarded as narrower in scope than Linguistics, Forensic Phonetics branched out from Forensic Linguistics and is today considered a separate discipline with distinct research interests and methods. The division is also observed in the fact that, around the world, forensic laboratories specialise in either Forensic Linguistics or Forensic Phonetics and Acoustics, which is also reflected in the existence of two separate international associations, *The International Association of Forensic Linguistics* and *The International Association of Forensic Phonetics and Acoustics* (Kašić & Đorđević, 2009). Since the research interests of Forensic Linguistics are beyond the scope of this study, we will not dwell on its practices and methodology any further.[2]

Probably one of the most comprehensive definitions of Forensic Phonetics is given by Jessen (2008), who defines it as "the application of the knowledge, theories and methods of

[2] For anyone interested in reading more about Forensic Linguistics, the following literature constitutes a fair starting point: Coulthard and Johnson A. (2007); Gibbons and Turell (2008); McMenamin (2002); Olsson (2008).

general phonetics to practical tasks that arise out of a context of police work or the presentation of evidence in court, as well as the development of new, specifically forensic-phonetic, knowledge, theories and methods" (p. 671).

Hollien (1990; 2002) identifies five areas of Forensic Phonetics, and these include (1) speaker identification, (2) vocal behaviour/stress in voice, (3) speech enhancement, (4) speech decoding, and (5) tape authentication. Nolan (1999: p. 2) also mentions the task of determining a speaker's origin by inferring the facts about their regional accent (today known as speaker profiling) and questions whether Forensic Phonetics is an appropriate term to encompass all of these tasks, bearing in mind that most of the traditional phonetic research assumes a shared linguistic system, while all of the individual differences are observed as "noise", and there has been considerable effort to normalise and eliminate these between-speaker differences in research (Nolan, 1999: p. 2). Furthermore, in most forensic cases, the expertise of phoneticians alone is not enough; the forensic task often cannot be performed without specialists in acoustic signal analysis and speech technology. This is why one of the most prominent international associations in this domain that was established in 1991 (The International Association of Forensic Phonetics) was renamed the International Association of Forensic Phonetics and Acoustics in 2004 (Jessen, 2008: p. 672). Due to the reasons mentioned above, there are some alternative terms to refer to this discipline: *Forensic Speech and Audio Analysis* (Jessen, 2008) or, nowadays, widely accepted, *Forensic Speech Science* (French & Stevens, 2013; Nolan, 1999).

French (1994) groups the tasks that a forensic phonetician can face into five main areas: (1) speaker identification, (2) determination of unclear or contested utterances, (3) authenticity examinations of audio recordings, (4) evaluation of speaker recognition evidence given by lay witnesses, and (5) speaker profiling. A few years later, Foulkes and French (2001) write about four main applications of phonetics to legal context: "deciphering the content of 'difficult' recordings, speaker profiling, speaker identification, and constructing voice 'line-ups' in order to evaluate ear-witness testimony" (p. 329). Finally, using the updated terminology, French and Stevens (2013) mention only three main sub-areas of forensic speech science (FSS), including speech content analysis, speaker profiling and forensic speaker comparison. The gradual decreased representation of other FSS tasks in theoretical literature overtime is most likely the reflection of their scarcity in practical work. Namely, as French and Stevens (2013) point out, around 70% of the laboratory work of an expert is comprised of forensic speaker comparison assignments (p. 187).

In the following part of the section, we will briefly outline and explain the main tasks of forensic speech science. For more information about the methodology and case examples, the reader is advised to consult the relevant literature.

Speech content analysis or *deciphering the content of "difficult" recordings* corresponds to what Hollien (1990; 2002) defines as *speech enhancement* and *speech decoding* and presents the investigation of an audio recording to determine what was being said. It may involve general transcription of the content or the analysis of certain disputed utterances, the former usually being required when the recording is of poor quality or the speaker is unintelligible due to a foreign accent, while the latter implies a detailed analysis of a short utterance or even a single word (Foulkes & French, 2001; French & Stevens, 2013).

Speaker profiling is a procedure utilised in criminal cases with a voice recording but without a suspect. It is often employed by the police during the investigation phase, for instance, when a kidnapper leaves a message on the telephone, and it implies extracting as much information as possible about the speaker's regional, socioeconomic or ethnic background (Foulkes & French, 2001; French & Stevens, 2013; Jessen, 2007; Rose, 2002;). The regional markers are usually researched and confirmed by reference to published studies, and the experts sometimes record the representatives of the potential "target" population for comparison (French & Stevens, 2013). The strength of conclusions depends on the length and quality of the recorded material, potential voice disguise, and the extent of available descriptive dialectological and sociolinguistic information (Foulkes & French, 2001). Speaker profiling has also found its application in Language Analysis for the Determination of Origin (LADO), a procedure employed by immigration authorities to determine the origin of asylum seekers (French & Stevens, 2013). Cambier-Langeveld (2016), however, warns that the term LADO may not fully describe the task of the language analysts in these cases, as they tend to "investigate whether the language skills of the asylum seeker support the claimed origin", and since the origin is not being determined but examined, this procedure presents a verification rather than an identification task (p. 28). As a result, recently, the term LAAP (Language Analysis in the Asylum Procedure) has gained influence in the field (see Hoskin et al., 2020; Hoskin, 2022).

Detection of stress in voice is determining whether a speaker is suffering from stress states. Hollien (1990) gives examples of a control tower talking to a pilot in trouble, a civil help centre talking to a caller threatening to commit suicide, or law enforcement personnel in various situations. While there are a plethora of studies within forensic speech science focusing on emotive speech (see Đorđević et al., 2004; Hansen & Clements, 1987; Ivanović & Kašić, 2011a;

Ivanović & Kašić, 2011b; Kašić & Ivanović, 2011; Steeneken & Hansen, 1999; Williams, 1972) and speech characteristics under the influence of narcotics and alcohol (see Gawell, 1981; Hollien et al., 1998; Hollien et al., 2001a; Hollien et al., 2001b; Pisoni & Martin, 1989; Sobell L. & Sobell M., 1972; Tisljár-Szabó et al., 2014), in literature, this area of research is rarely recognised as a separate discipline within the FSS.

Authentication of recordings is performed when one of the parties in court challenges the validity of the recorded material claiming that the recording has been modified so that it does not correctly represent the events that took place at the time it was made (Hollien, 1990). In the past, when the material was recorded via analogue tape recorders, the procedure involved physical and acoustic tape examination (Hollien, 1990). Today, experts often rely on the encoding parameters, including bitrate, sampling rate, or timestamps, auditory analysis, sound spectrum analysis etc. (see Grigoras, 2005; In Park et al., 2022; Rappaport, 2000; Xu et al., 2022). Of particular importance for authenticity analysis may be Electric Network Frequency (ENF). Namely, a recording device may capture an alternating current power hum that varies smoothly and randomly around a nominal operative value, depending on the location (e.g. 50 Hz in continental Europe). The recording then contains a series of harmonic tones, the fundamental of which is the ENF and manipulation of the recording content may create discontinuities in the ENF signal (see ENFSI, 2009; 2022; Grigoras, 2005

The construction of voice line-ups or *voice parades* is a procedure performed by experts in earwitness testimonies. It involves setting up a listening experiment for the purposes of speaker recognition by naïve listeners. The suspect's voice is presented in a group of other voices – foils – and the witness is asked to identify which among the group belongs to the offender (Foulkes and French, 2001, Hollien, 1990; 2002; McDougall, 2013b; Nolan, 1999). Since earwitnesses are untrained individuals known as naïve or lay listeners, to obtain the best possible results, the voice line-up must be administered in a rigorous and structured manner with paying particular attention to the proper selection of foils (see Künzel, 1994; Broeders, 1996; Broeders & van Amelsvoort, 1999; de Jong-Lendle et al., 2015). In the UK, voice line-ups are administered according to the guidelines prepared by John McFarlane of the Metropolitan Police and Professor Francis Nolan in the publication "Advice on the use of Voice Identification Parades" (Home Office, 2003). In addition, recent research by McDougall and colleagues from the Cambridge University under the IVIP project has had a large impact on the understanding of the factors that affect naïve listener judgements for the purpose of creation of fair voice parades (see McDougall, 2021; McDougall et al., 2022; McDougall et al., 2023; Paver et al., 2021).

Forensic speaker identification is often seen as part of *forensic speaker recognition*, the other being *forensic speaker verification* (Nolan, 1983; 1990; 1999; Hollien, 2002; 2013). Nolan (1990) holds that *speaker recognition* is a general term that refers to any process of "attributing a speech sample to a person on the basis of its phonetic-acoustic content" (p. 457), and he differentiates between *naïve speaker recognition* (also *"casual" speaker recognition*) and *technical speaker recognition* (Nolan, 1983: p. 7; 1990: p. 457; 1999: p. 1; 2001: p. 3). The former occurs in everyday life when people employ their natural ability as language users to recognise the voices of their parents, children, acquaintances, or a famous person on the radio/television without relying on any technical methods of analysis, whereas the latter entails that trained experts perform the task using either auditory analysis or a machine-based technique. *Speaker verification* (SV) usually implies that the speaker wants their identity to be confirmed or that such validation is necessary in order for the person to be granted certain access. In this case, there may be an established repository of voices according to which the system compares the disputed voice for identity verification purposes (Broeders, 2001: p. 2; Hollien, 2002: p. 12; 2013: p. 2; Nolan, 1990: p. 458; 1999: p. 1-2). On the other hand, speaker identification usually involves an uncooperative speaker coming from a population of unknown size and composition, who needs to be identified by their speech and voice analysis. As such, it presents a much more challenging task (Hollien, 2013: p. 2; Nolan, 1990: p. 458; 1999: p. 2). Even though the earlier literature specifies that recognition is broader in scope than identification, the two terms are often used interchangeably (e.g. Nolan, 2001; Jessen, 2008). Jessen (2008: p. 673) uses the term *speaker identification* in a broader sense and defines it through three different tasks: voice or speaker comparison (the term also used by Braun and Künzel, 1998), voice profiling or speaker classification, and speaker identification by victims and witnesses. Jessen's (2008) definition of voice comparison largely corresponds to what Nolan (1983; 1990; 1999), Foulkes and French (2001), Rose (2002) and Hollien (2013) refer to as speaker identification.

2.1.2. Brief historical overview

The admissibility of aural-perceptual testimony in the UK courts is traced back to 1660 when William Hewlett was accused of regicide (Hollien, 1990). On the other hand, in the USA, the acceptance of earwitness identification testimony dates back to 1907, when a cross-racial suspect in Florida was identified as a rapist by the victim on the basis of two spoken sentences (Hollien, 1990: p. 192). Broeders (2001), however, notes that one of the most remarkable applications of earwitness identification in courts concerns the Lindbergh baby

kidnapping case in the 1930s, when the father of the abducted child, a famous aviator, claimed to have recognised the voice of the perpetrator by his German-accented English almost three years after the crime was committed. The controversial validity of this testimony stimulated the rise of research in the area of speaker identification by lay listeners (Broeders, 2001).

Forensic speaker comparison by experts emerged in the first half of the 20th century, after the invention of the tape recorder and sound spectrograph when it became possible to "capture, replay and visually represent" the acoustic signal of human speech (Broeders, 2001: p. 4). One of the earliest UK court cases in which forensic phonetic evidence by experts was used was at Winchester Magistrates Court in 1967 (Ellis, 1990, as cited in French, 1994: p. 169). While "the work of forensic phoneticians consisted almost exclusively of the identification of speakers in criminal recordings" in the UK until 1980s (French, 1994: p. 169), in the USA, there are reports of experts also working on cases that required their expertise in validating the authenticity of tape recordings (see Hollien, 1990: p. 3). Nowadays, it is estimated that experts in the UK in this field are consulted in around 500 to 600 criminal investigations and legal cases per year (French, 2017: p. 1; French & Stevens, 2013: p. 196), 70% of which involve the task of forensic speaker comparison (French & Stevens, 2013: p. 187).

In the earliest stages of the development of Forensic Linguistics, forensic tasks were scarce, and there were not any attempts to establish the analytical framework or methodology. Instead, the linguists engaged in such endeavours for the sake of intellectual challenge and creativity (Kašić & Đorđević, 2009: p. 474). There were no professional bodies or organisations to provide regulations regarding the practice, and the depth and detail of the analysis were inadequate by modern standards (French, 2017: p. 2). The first conference on forensic applications of phonetics was organised in the United Kingdom in 1989 (Rose, 2002: p. 18) and was soon followed by the forming of the International Association for Forensic Phonetics (IAFP), renamed the International Association for Forensic Phonetics and Acoustics (IAFPA) in 2004 (French, 2017: p. 2). The Association aimed to provide a forum for discussion among active forensic speech science practitioners or academics interested in the subject. It has established a professional Code of Practice, as well as Guidelines for Keeping a Record of Analysis (Braun & Künzel, 1998: p. 13).

The earliest approaches to forensic speaker comparison in the 1960s in the USA and the UK were diametrically opposed, the experts in the USA relying on the spectrograms, the notion known as 'voiceprint', whereas the experts in the UK relied exclusively on the auditory analysis of the recordings. By the end of the 20th century, both approaches were heavily criticised and known to have limitations (French, 1994: p. 170).

The voiceprint method was developed during the Second World War, the underlying theory being that just like our fingers have a unique print, so does our voice (Kersta, 1962; Smrkovski, 1975). Therefore, it was assumed that spectrograms of different realisations of linguistically identical utterances produced by the same speaker are bound to have similarities in patterns, while the realisations by different speakers would produce different spectrographic images (Broeders, 2001: p. 4-5; French, 1994: p. 170). However, it is now known for a fact that, unlike fingerprints, the speech of an individual is susceptible to prominent within-speaker variability and is not invariant over time, the variations being reflected in the appearance of spectrograms (Broeders, 2001: p. 5; French, 1994: p. 171; Nolan, 1999: 759-765). Furthermore, French (1994) criticises that the proponents of this method failed to explain "what should be taken to constitute a forensically significant or diagnostic *similarity* or *difference* between spectrograms" (p. 172), while Braun and Künzel (1998) question the expertise of the individuals and organisations performing the training and the procedure (p. 11). Similarly, Hollien (1990) describes the voiceprint method as a mere 'pattern-matching' procedure (p. 212) and examines various controversies regarding the methodological framework and research performed to justify its application in legal systems. The official position of the IAFPA since 2007 is that the method is without scientific foundation, as described in Tosi (1979), and should not be employed in forensic casework (IAFPA, 2022).

In the seventh and the eighth decade of the 20th century, some phoneticians believed that the auditory analysis in FSC was sufficient on its own (see Baldwin, 1979). In the auditory-phonetic approach, trained phoneticians would undertake narrow phonetic transcriptions of both the questioned and the known speech sample to capture the details of vowel and consonant production. The phoneticians would also address the intonation, rhythmical and fluency features. However, apart from the comparability of samples, not much could be concluded from the auditory analysis of pitch unless it was combined with the acoustic analysis. In this approach, auditory impressions of voice quality were expressed holistically, without a systematised scoring protocol and with no regard to its constituent phonatory and vocal tract settings (French, 2017: p. 2).

A method that was pioneered by the German Bundeskriminalamt in 1980 (Künzel, 1995, as cited in Rose, 2002: p. 18) and still comprises most of the modern-day FSC is the combined auditory-acoustic approach (French, 1994: p. 173–4; Nolan 1999: p. 14), also known as the phonetic-acoustic approach (Rose, 2002: p. 49). In this framework, the speech signal is viewed through components including segmental (consonant and vowel features and connected speech processes), suprasegmental (voice quality, intonation, general pitch, speech rate,

rhythmical features) as well as high-level and non-linguistic features (morphology, lexico-syntax or discourse organization, speech pathologies, disfluencies) (French & Stevens, 2013). At the beginning of the 21st century, forensic speaker comparison by using a combination of auditory and acoustic methods was performed in several government forensic laboratories, including Germany, Austria, Sweden, the Netherlands and Spain, and in private practice in countries like the United Kingdom and Germany (Broeders, 2001: p. 5).

A significant development in forensic speaker comparison occurred with the advancement of computer technology and the emergence of automatic speaker recognition (ASR) software. Such software implements powerful signal-processing pattern recognition algorithms, reducing the recordings to a statistical model based on mel frequency cepstrum coefficients (MFCCs), and then, using intensive probabilistic-statistical processing, produces a measure of similarity and difference between two samples, also comparing them to a reference population of other speakers that exist within the system (French & Stevens, 2013: p. 188; Rose, 2002: p. 56). Common terms encountered in literature for this technology also include FASR – forensic automatic speaker recognition, and FSASR – forensic semiautomatic speaker recognition (see Drygajlo et al., 2015). The difference between the two lies at the feature extraction level: it operates automatically in FASR but manually in FSASR (p. 14). The ASR systems have proven rather successful and have been employed both in government and private laboratories around the world since the beginning of the century (Broeders, 2001: p. 6; Rose, 2002: p. 56), their main advantages being replicability of results, efficiency and numerical output of the likelihood ratios (French & Stevens, 2013: p. 188-189). However, French and Stevens (2013) claim that despite all their advantages, the ASR systems cannot be regarded as infallible (p. 189), especially since previous research has established that there is a substantial degree of convergence across same-sex speakers within ethnic groups as well as across speakers of particular language varieties in terms of the vocal tract settings they tend to adopt (p. 189-191). In addition, these systems have proven to perform with higher error rates when the recordings are of poor quality (French & Stevens, 2013: p. 189), and they are known for their sensitivity to transmission channels, including the effects of different handsets, telephone lines and GSM-coding (Broeders, 2001: p. 6). As French (2017) reported, at the time of writing the paper, the ASR system evidence was unlikely to be accepted in courts in England and Wales (p. 6).

2.1.3. Forensic speaker comparison in Serbia

In Serbia, the application of FSS in courts is performed sporadically (Đorđević et al., 2011); the requirements for FSC depend on the extent to which the participants in the investigation and judicial process are generally informed about its existence and possibilities (Kašić & Đorđević, 2009).

The records indicate that the first FSC case in Serbia can be tracked to 1977, the so-called "Mystery of Matejevac" (Zarić, 2004), on which the information is now only available in the media. Namely, years after the disappearance of V. N., whose body was never found, his wife D. N. and her lover M. K. were arrested and accused of murder. The widow was interviewed 14 times during the investigation, changing the details of her statement until she finally admitted that her lover had murdered her husband and she had helped him burn the body. Her statement was recorded on a reel-to-reel tape recorder; however, the prosecutor was not satisfied with D. N.'s confession and required another hearing. Subsequently, she changed her statement several times, altering the details of who and how exactly killed V. N.

After D. N. committed suicide in prison, the accused, M. K., challenged the authenticity of one of the recordings of D. N.'s confessions, claiming that the person speaking was not D. N. As a result, the assigned judge of the Regional Court in Niš (now Higher Court) decided to send the recordings to the Police Department for Forensic Science in Belgrade for analysis. However, the expert analysis was performed in the Belgrade Radio and Television studio, where the sound experts expressed an opinion that the voice on both recordings originated from the same person. The difference was perceptible, as they claimed, due to the recording speed, greater exploitation of one of the reels, the difference in recording devices, and the surroundings in which the recording was performed (Zarić, 2004).

Even though there are no legal constraints regarding FSC methods and expressing the outcome, the review of limited court practice available at Sudska praksa (https://sudskapraksa.sud.rs) reveals that the methods used by expert witnesses in this process are relatively consistent. In the process of FSC, experts mostly rely on the analytical approach. Audio-perceptual listening by trained listeners allows the experts to extract the markers for the phonetic-acoustic analysis. The holistic approach is sometimes used, but it is not optimal for presenting the results in court as the judges want to hear the specificities of what makes the two speech samples similar or different (Đorđević et al., 2011).

In Serbia, to express their opinion in courts, FSC experts sometimes rely on a verbal scale of ranked probability scores (Jovičić, 2001) that was used by German federal and state

forensic speech and audio laboratories and private experts at the beginning of the 2000s (Gfroerer, 2003). The scores are: cannot be assessed ("no-decision" vote) / is probable / is highly probable / is very highly probable / can be assumed with near certainty (Gfroerer, 2003; Jovičić, 2001).

Reliance on automatic speaker recognition software is not explicitly forbidden in courts in Serbia; however, there is a consensus in the Serbian scientific community that ASR software should never replace expert analysis, even though it may be used as a tool to assist in the process. This is mainly because such software is unable to detect dialectological information that is often crucial in FSC in Serbian (Đorđević et al., 2011).

2.2. Speaker Specificity and Sources of Variation

The task of forensic speaker comparison (FSC) involves comparing one or more speech samples of an unknown speaker to one or more samples of a known speaker in order to assist the courts or relevant authorities in determining whether the samples were uttered by the same person (French, 2017; Jessen, 2018; Morrison, 2009; Nolan, 2007). In literature, the voice recording of the unknown speaker may be referred to as a *disputed*, *offender*, *criminal*, *perpetrator* or *questioned* sample, whereas the known voice can also be termed *defendant* or *suspect* sample (see French, 2017; Hollien, 1990; Jessen, 2018; Nolan, 1999; 2001; Rose, 2002).

At the core of forensic speaker comparison is the assumption that individuals differ in how they speak and how their voices sound (Rose, 2002), even though, from what we know so far, it is believed that "there is no unique pattern that distinguishes one speaker from everyone else without any overlap" (Jessen, 2018: p. 219). The differences between speakers are termed *between-speaker* or *inter-speaker variations*, while the voice variation of a single person due to different circumstances is called *within-speaker* or *intra-speaker variation* (Rose, 2002: p. 25).

2.2.1. Forensic parameters

One particular feature by which voices are compared is labelled *dimension* or *parameter*. Voices can be compared in terms of different dimensions; however, the most powerful are the ones that exhibit greater between-speaker than within-speaker variation. (Nolan, 1983: p. 11; Rose, 2002: p. 33). When selecting a forensic phonetic parameter, it has been recommended to inspect the ratio of between-speaker to within-speaker variation (Kinoshita, 2001; Nolan, 1983; Pruzansky & Mathews, 1964; Wolf, 1972). The ratio is called

F-ratio and is obtained by performing the statistical procedure known as Analysis of Variance or ANOVA (Rose, 2002: p. 33).

According to Nolan (1983), an ideal forensic parameter should not only assume high between-speaker and low within-speaker variability, but it should also be resistant to attempted disguise or mimicry; it should occur frequently in relevant materials, be largely independent of transmission channels and relatively easy to extract and measure (p. 11). Rose (2002) adds that "each parameter should be maximally independent of other parameters" (p. 66). However, an ideal parameter as described above does not exist; thus, the list of criteria should be regarded rather as a guide. In addition, parameters that could easily discriminate specific speakers may not be useable for others, especially if the pair of speakers are close to the average value of the reference population (Rose, 2002: p. 318). In other words, the rarer the speaker features in the population according to a specific dimension, the more likely the dimension will be an effective discriminant in that particular case of forensic speaker comparison. According to the survey by Gold and French (2011), practitioners agreed that "despite some individual parameters holding significant weight, it is the overall combination of features that they consider crucial in discriminating between speakers" (p. 754).

Laver (1968) introduces the term evidential information to denote the attributive markers that listeners use as the basis on which to characterise speakers. These can be grouped into three categories: physical markers, relating to physical characteristics such as age, sex or physical state of health (such as voice quality), social markers, indicating social characteristics such as regional, social or educational background, occupation or social role (accent and lexicon), and psychological markers that reveal psychological characteristics of personality and affective state of mood (tone of voice). The evidential markers, in this sense, have a semiotic status; they are "indexical" of a speaker, i.e. reveal even the information that speakers do not deliberately intend to convey (Laver, 1994: p. 15).

Linguists differentiate between three types of speech behaviour: linguistic, paralinguistic and extralinguistic, whereby all three types are informative, but only linguistic and paralinguistic behaviour is coded and communicative (Laver, 1994: p. 21). Accordingly, forensic parameters can be either linguistic or non-linguistic (Rose, 2002: p. 57). A linguistic parameter is any feature "that has the potential to signal a contrast, either in the structure of a given language or across languages or dialects" (Rose, 2002: p. 58). Paralinguistic parameters, on the other hand, relate to the speech behaviour that non-verbally communicates the speaker's current affective, attitudinal or emotional state, such as anger, sadness, excitement, disappointment, or happiness (Laver, 1994: p. 21), while the extralinguistic parameters refer to

those non-coded aspects of speech which signal the information about the identity of the speaker, particularly concerning habitual factors such as the speaker's voice quality, and overall pitch and loudness range (Laver, 1994: p. 23). The attribution of social, psychological and physical characteristics from speech cannot, however, be correlated directly with linguistic, paralinguistic and extralinguistic information, respectively (p. 23).

It is important to understand, however, that what could belong to the paralinguistic domain in one language could bear linguistic meaning in another. For instance, the choice of phonation type in English may be an index of social background (see Henton & Bladon, 1985; 1988; Laver, 1968; Wright et al., 2019; Yuasa, 2010), or the speaker's physical or emotional state (see Gobl & Ni Chasiade, 2003; Laver, 1968; Laver & Trudgill, 1979; Ni Chasiade & Gobl, 2005), while at the same time, it may be a linguistic feature in another language (see Esposito & Khan, 2020; Gordon & Ladefoged, 2001; Keating et al., 2010). This is why the choice of forensic parameters in forensic speaker comparison must be informed by knowledge of the nature of the language in question (Rose, 2002: p. 62).

Another classification of forensic phonetic parameters is into auditory and acoustic, the latter being further divided into traditional or automatic (Rose, 2002; Jessen, 2010). Forensic parameters can also be quantitative, provided they can be expressed with numeric values, and qualitative (nominal or ordinal), if they can only be expressed descriptively (Aitken and Taroni, 2004). Quantitative parameters are either discrete, if they can assume only a fixed number of values, or continuous, wherein the samples can be quantified more precisely (Rose, 2002).

In forensic speaker comparison, there is no predetermined set of parameters that would usually be tested; the choice depends on the actual circumstances of the case, the perceived similarities and differences of the audios, as well as the language in question (Rose, 2002: p. 47). These could be features from the phonetic domain, including segmental or suprasegmental features (different aspects of vowel formants and consonants, fundamental frequency, voice quality, intonation, tempo, rhythm), higher-order linguistic features (discourse markers, conversational behaviour, lexico-grammatical usage) or non-linguistic features (filled pauses, tongue-clicking, audible breathing, throat clearing, and laughter) (Gold & French, 2011; French et al., 2010; French, 2017).

Finally, there is not a single parameter which is an absolute discriminant that can unmistakably be used for forensic speaker comparison in every forensic case. Due to the plasticity of the human vocal system, every individual can produce a range of acoustic characteristics for each forensic parameter (Nolan, 1983: p. 59). Ever since 1970s, there has been extensive research on the factors that condition the variability of speech production (see

Hollien, 2012; Nolan, 1983; Stevens, 1971). In the following sections, we will examine the prevailing views on the sources of within-speaker and between-speaker variability, respectively, even though, as will be shown, the boundary between the two sets of features is not clear-cut.

2.2.2. Sources of between-speaker variability

One of the proposed views is that between-speaker differences can be categorized as "organic" and "learned" (Garvin & Ladefoged, 1963: p. 194; Kašić & Đorđević, 2009a; Tosi, 1979: p. 55; Wolf, 1972: 2045). Namely, our vocal apparatus varies in size and shape, much like our external appearance does, and since the dimensions of the vocal tract and larynx condition phonetic properties such as resonant frequencies and vocal cord vibration rate, how we sound does depend on our physique (Mackenzie Beck, 2010; Gobl & Ní Chasaide, 2010). On the other hand, ever since our first exposure to language, we acquire more than just the linguistic system; we acquire a socially and regionally marked variety of pronunciation, constructing a linguistic-phonetic system that defines us as belonging to a specific sub-group of the population, and this is what constitutes the "learned" factors (Garvin & Ladefoged, 1963). To distinguish between vocal features that are under the speaker's control and those that are not, Laver (1991a) uses the terms "extrinsic" and "intrinsic" features (p. 163).

However, Nolan (1983; 1999) argues that this dichotomy is simplified, as there is no precise boundary between the biologically determined features of voice and those representing learned social behaviour. He illustrates that the "intrinsic" aspect of idiosyncrasy does not reflect absolute values; rather, these should be regarded as "limitations on the variation that the speaker could induce on his vocal apparatus" (Nolan, 1983: p. 72). Nolan (1983) develops a model for revealing the basis of speaker-specific information in the speech wave and the sources of variability, which he claims are in a symbiotic relationship (p. 72). At the top of his model is the communicative intent, mapped onto two sets of phonetic resources, segmental and suprasegmental, the integration of which yields the phonetic representation that contains all the details of an utterance that are of potential linguistic relevance. The specifications of the phonetic representation are then acted upon by implementation rules, which result in neuromuscular commands, that is, the movement of vocal organs and the production of the acoustic signal (Nolan, 1983). According to his model, the phonetic resources also incorporate two second-order long-term strands, corresponding to the two primary strands (segmental and suprasegmental), each contributing their long-term target specifications to the phonetic representation (Nolan, 1983: p. 34). The long-term segmental strand relates to the

resonance characteristics of the vocal tract, such as voice quality and phonation type, while the long-term suprasegmental strand is reflected in default values for mean pitch, pitch range, loudness or speech rate (p. 51). Nolan (1983) notes that the "learned" aspect of speaker-specificity that occurs at the lower level of the model is performed by trial and error rather than through direct imitation of implementation strategies, whereas at higher levels, the speaker learns through exposure to language use and on the basis of innate understanding that language is a complex mechanism of expression. The mechanism serves for the mapping of different aspects of communicative intent and many parts of the mechanism (segmental, suprasegmental, short and long term, primes and realisation rules) could be affected by a single aspect of the communicative intent, for instance, use of nasalisation to communicate irony in a specific social context (p. 72-73).

Relying on Nolan's (1983) model for revealing the basis of speaker-specific information in the speech wave and the sources of variability, Rose (2002) observes voice from a semiotic perspective by presenting a voice model according to which a speaker's voice results from two inputs processed through two mechanisms (p. 293-294). The primary input of the system is the communicative intent, understood as in Nolan (1983), whereby the speaker chooses to convey specific meaning; it incorporates cognitive, affective, social, regulatory, and self-presentation intent (Rose, 2002: p. 300-305). The other input, intrinsic indexical factors, denotes the "intersection of indexical and intrinsic information" in the speech wave (p. 305). Rose's (2002) intrinsic indexical factors are Laver's (1968; 1991a) concepts that correspond to Nolan's (1983) second-order long-term strands. The term indexical refers to the information revealing characteristics of a speaker, which could be extrinsic and intrinsic. Intrinsic information that is revealed non-volitionally is primarily biological, and, in Rose's (2002) model, it consists of the age, sex, physique, and physical and psychological health of the speaker (p. 306). The communicative intent and intrinsic indexical factors are then processed through the linguistic and vocal mechanism, the former being the result of language (phonetics, phonology, morphology, syntax, lexicon) and the tone of voice, while the latter is the anatomical structure of the vocal tract organs used in the speech and other vocalic production (p. 285-300). Rose's (2002) model demonstrates that the variation in a speaker's output is a function of their communicative intent and the dimensions and condition of their individual vocal tract. According to both Nolan (1983) and Rose (2002), only if we understand what underlies within-speaker variation can we correctly evaluate differences between speakers in forensic speaker comparison (p. 311).

More recently, an extended classification of between-speaker differences was suggested by Jessen (2010), who categorized them as "organic", "idiolectal", and "habitual" (p. 387). According to him, organic features are partially determined by our biological construct, such as the length of the vocal tract or vocal folds, and these include fundamental frequency, formant frequencies and voice quality (p. 391). For instance, despite the low correlation, it is confirmed that a speaker with very low formant frequencies is likely not to be a short person, and someone with very high formants is likely not to be tall (Greisbach, 1999; Jessen et al., 2005). Furthermore, speakers with long-term third formant in the high range above 2,500 Hz have a small or medium body size, whereas speakers with formants in the low range below 2200 Hz have an above-average body size (Jessen, 2010: p. 382). Vocal tract length, which is reflected in vowel formants and fundamental frequency, may also be indicative of the biological sex of the speaker (Jessen, 2010: p. 283). Mature males are estimated to have a vocal tract that is, on average, 20% longer than females', resulting in lower fundamental and formant frequencies (Rose, 2002: p. 307). According to some earlier studies, the maximum range of fundamental frequency is 50-250 Hz for men, and 120-480 Hz for women (Fant, 1956), with means for men between about 80 and 170 Hz (Jessen et al., 2005; Künzel, 1989) and for women, between 165 to 260 Hz (Künzel, 1989; Simpson & Ericsdotter, 2007).

Medical conditions within the domain of language, speech and voice pathology also belong to the sphere of organic differences. What is primarily meant here are all those permanent conditions that have an invariable long-term effect on someone's speech, unlike laryngitis or nasal cavity congestion that may appear as the effects of a current health state. For example, Jessen (2010) lists stuttering and sigmatism as potential distinctive features between speakers (p. 382-383). In addition, medical conditions outside speech-language pathology that have been applied forensically include obstruction in the breathing pathways, which may be an index of obesity (p. 383).

Age correlates are also primarily based on biological factors (Jessen, 2010: p. 383). Research has shown that, both for male and female speakers, the fundamental frequency may decrease through early and mid-adulthood and then increase again later in life (Baken & Orlikoff, 2000; Hollien & Shipp, 1972; Russell A. et al., 1995). Similarly, it was found that speech rate is a good age correlate because it decreases gradually over time for speakers of both sexes due to physiological conditions of the vocal tract (Bóna, 2014; Jacewicz et al., 2009; Ramig, 1983).

In Jessen's (2010) classification of between-speaker differences, the term "idiolectal" concerns our speech with regard to social context, regional or dialectal

characteristics that will remain even if we change our location and setting in the course of our lifetime, the degree of dialect or foreign accent influence, some speaker-specific segmental phenomena or prosody, as well as idiosyncratic aspects of syntax and the lexicon (p. 391-392). Namely, just as people living in different geographical regions speak different linguistic varieties – dialects, different social groups have different sociolect, which may be reflected in the lexicon, grammar, phonetic and phonological features of speech (Trudgill, 2000). As Nolan (1983) explains, the speakers typically choose to signal their membership of a specific social, ethnic or regional group by manipulating aspects of linguistic structure, and this constitutes their social intent, a sub-category of the communicative intent (p. 63). Even though attitudes to language may play a role in preserving or removing social and dialectal markers in speech (Trudgill, 2000), remnants of the automated articulation base acquired as the mother tongue in the native environment can seldom be erased (Kašić & Đorđević, 2009a), which is why social and regional markers pose as important speaker discriminants in forensic speaker comparison.

Finally, Jessen's (2010) "habitual" features refer to those characteristics of speech that "do not have any obvious organic foundation nor are they related to the linguistic conventions that are required or expected by the language system or the social community" (p. 392). In this group, he lists features such as articulation rate, fundamental frequency variability, and speech disfluencies (p. 392). Similarly, Köster O. and Köster J. P. (2004) refer to Gfroerer and Baldauf (2000), who group all the features in three rough categories: (1) voice in the narrowest sense, (2) speech or articulation, and (3) manner of speaking or suprasegmental phenomena. As they explain, "voice phenomena" can either refer solely to the direct production of voice at the level of vocal folds or also include the filtered signal of the vocal tract, that is, resonance (p. 10).

To conclude, despite the attempted classifications of between-speaker differences, the consensus in the literature is that these categories are not always clearly distinguished as certain features may at the same time be determined by human biology and social context (see Jessen, 2010; Köster O. & Köster J. P., 2004; Nolan, 1999). For instance, according to Köster O. and Köster J. P. (2004), intonation may be observed as both a voice feature and a manner of speaking (p. 10). Similarly, a fundamental frequency may be an indication of someone's biological age or used to convey certain meaning (see Jessen, 2010: p. 383). In addition, while phonation types in English are observed as habitual phenomena, in some languages, they are used contrastively on different phonemes, such as breathy voice in Gujarati or creaky voice in Jalapa Mazatec (Gordon & Ladefoged, 2001: p. 163). Therefore, what can be observed in English as a feature of between-speaker variability, or an idiosyncratic possibility, in another

language may be part of the linguistic repertoire or voluntary variation of the glottis actions (Gordon & Ladefoged, 2001: p. 163). As Nolan (1999) underlines, an individual's anatomy does indeed impose certain limits on speech; however, each person has a "very wide scope for controlled variation" within these limits (p. 3).

2.2.3. Sources of within-speaker variability

According to Nolan (1983) sources of within-speaker and between-speaker variation are not two unrelated issues to discuss. The way someone speaks on a particular occasion is "the result of a complex interaction between his communicative intent, the language mechanism he controls and the context in which he is speaking" (Nolan, 1983: p. 73). In order to evaluate whether two voices come from the same person or not, a forensic expert must understand how voices differ with regard to these factors. Some of the variations may reflect our voluntary choice to exploit the linguistic system's and vocal tract's plasticity, while others may be an involuntary side-effect of our physical mechanism of speech undergoing certain changes (Nolan, 1983: p. 27-28; Nolan, 1999). Some of the common factors mentioned in the relevant literature include speaker's emotional and physical state, various health conditions, intake of psychoactive substances such as drugs or alcohol, social context and familiarity with the interlocutor, deliberate voice disguise and the effects of recording devices and transmission channels (see Hollien, 2012; Jessen, 2010; Köster O. & Köster J. P., 2004; Nolan, 1999; Rose, 2002).

People can choose to signal an emotional state, their short-term attitudes and feelings when speaking, which Nolan (1983) refers to as the affective intent (p. 62). To signal emotions linguistically, we can use different words or syntax, but also different intonation pitch or phonation type. For example, in some varieties of British English, using breathy voice can convey sympathy, whereas using creaky voice at the end of an utterance can signal boredom (Rose, 2002: p. 301). Factors such as fatigue, stress, and the diurnal cycle may affect fundamental frequency and phonation type (Hollien, 2012; Nolan, 1999). Stress is reflected in the increase in pitch, fundamental frequency or frequency variability and intensity in speech. Speech disfluencies also increase with stress increments. On the other hand, speech rate tends to be reduced when the speaker is stressed out (Hollien, 2012: p. 42-43). Emotion such as anger may also condition changes in loudness, mean pitch, pitch range, and phonation type (Nolan, 1999).

As Rose (2002) explains: "any changes in health that affect the size or shape or organic state of the vocal tract, or its motor control, will alter its acoustic output, thus

contributing to within-speaker variation" (p. 308). The health-related changes can be temporary (common cold), periodic, chronic (vocal fold polyp), or permanent (effects of surgery or hormonal therapy) and can contribute to perceiving the same speaker as sounding more different in certain parameters such as mean fundamental frequency, or fundamental frequency range (Nolan, 1999; Rose, 2002). Recent studies have confirmed that fundamental frequency and the perceived gender of voice can change due to hormonal therapy, which transgender people often undertake. In the study by Nygren et al. (2016), patients with female-to-male transition were able to reduce their mean and mode fundamental frequency over the course of 12 months to match male reference data (mean 125 Hz), though the change was not equally prominent for all the participants. Similarly, Marquez (2018) tracked the change in vocal features of male-to-female gender-transitioning people, confirming that the endocrine therapy in combination with voice modification therapy can result in the fundamental frequency increase, even though most participants remained in the lower spectrum of the reference population numbers (p. 11). The arrangement of teeth in the mouth as well as various dentures may affect the production of sound segments, in particular, sibilant fricatives and the resonance patterns of vowels (Rose, 2002: p. 308).

As far as the alcohol consumption is concerned, in some earlier research, it was found that, under the influence of alcohol, intensity and fundamental frequency were lowered while fundamental frequency variability, the number and length of speech pauses often increased (Chin & Pisoni, 1997; Pisoni & Martin, 1989). However, Hollien et al. (2001a; 2001b; 2009) tested the effects of alcohol intoxication on speech and found that, as intoxication increases, speaking fundamental frequency (heard pitch) is raised, and speech is slowed. In accordance with previous studies, they detected a strong correlation between disfluencies and intoxication level, which was also confirmed by Schiel and Heinrich (2015). Bearing in mind that psychoactive substances affect the motor functions of the vocal tract, various inconsistencies in pronunciation of segments are possible. For instance, the tongue might not achieve closure for dental plosive [d], which would result in a [z]-like fricative (Rose, 2002: p. 309). Similarly, the consumption of cannabis was found to have the acute effects on speech, such as the increase in mean pause and mean vowel duration and decrease in the phrase length (Zeidenberg et al., 1973), while the long-term effects may include changes in the spectral tilt (vocal effort and intensity) Vogel et al., (2021). A recent study on the effect of language under the effects of lysergic acid diethylamide (LSD) revealed that affected speakers exhibit increased verbosity and a reduced lexicon (Sanz et al., 2021). Furthermore, in the comprehensive research on the effects of smoking on voice it was reported that smoking increases the weight of the

vocal cords, which in turn decreases the fundamental frequency (Awan, 2011; Murphy & Doyle, 1987; Pinar et al., 2015) and is likely to result in changed voice quality, affecting the acoustical parameters such as jitter, shimmer, harmonics-to-noise ratio or smoothed pitch perturbation quotient (Gonzalez & Carpi, 2004; Vuković et al., 2022). Similar changes may occur under the exposure to electronic cigarettes, though to a smaller degree (Tuhanioğlu et al., 2019).

Inasmuch as sociolect and dialect may help distinguish between speakers, one must be careful when performing comparison on the basis of these markers due to a phenomenon known as code-switching – the ability to switch between one language or dialect and another depending on the social context (Soares & Grosjean, 1984; Trudgill, 2000: p. 201). Nolan (1999) warns that "differences of pronunciation between speech samples should only be interpreted in the light […] of sociolinguistic and stylistic variation within a speech community" (p. 7). One of these aspects is the perceived formality of the situation, or a stylistic variation; another is the perceived status relationship of the speaker to the interlocutor, or others present (Rose, 2002: p. 303). For example, in the forensic speaker comparison context, it is common that the questioned recording of a speaker is an informal conversation between two familiar associates, while the known recording is that of a police interview, which is a formal context. In formal contexts, speakers often adhere to the standard variety of the language, whereas communication between friends or peers can be marked by various non-standard segmental and prosodic features, choice of vocabulary and grammar (Nolan, 1999). For instance, in the rehearsed, formal speech, syntactic and informational structure are often more fully marked with intonation patterns than in informal, spontaneous speech. The linguistic choices may also be affected by whether the speaker wishes to appear friendly or rude to the interlocutor (Nolan, 1999).

A common occurrence in forensic casework is purposeful change of voice for the sake of concealing one's identity, also known as voice disguise (Künzel, 2000; Leemann & Kolly, 2015). According to customary classification (see Künzel, 2000; Perrot et al., 2007), it may include the modification of one or more of the following features in speech: a change in voice source characteristics, such as fundamental frequency or phonatory changes (see Růžičková & Skarnitzl, 2017), a change in resonance features, such as placing an object in the vocal tract, hypo- or hypernasality, face covering (see Fecher & Watt, 2011; Llamas et al., 2009), a change in language, accent, and dialect (see Eriksson A., 2010; Neuhauser, 2008; Sjöström et al., 2006), or a change in the manner of speaking, including reduction or exaggeration of pitch variation, speaking tempo and stress pattern. Research has shown that

voice disguise can be used to trick both the listeners and automatic speaker recognition systems into mistaking a person's identity (Farrús et al. 2006; González-Hautamäki et al., 2017; González-Hautamäki et al., 2018; Tan, 2010); therefore, practitioners must be aware of the possible disguise strategies and how they contribute to within-speaker variability when undertaking a forensic speaker comparison task.

Other phenomena that can affect the speech include difference in recording equipment or transmission channels, noise or somebody else's voice in the background (Broeders, 2001; Rose, 2002). Bearing in mind that the corpus for the present study is recorded over GSM network, a more extensive literature review concerning the effects of telephone transmission will be discussed in more detail in section 3.4.2.

To sum up, for forensic speaker comparison, it is crucial to understand that the speech of an individual is not a constant, either in terms of those properties which result primarily from the physical mechanism of speech, or those which are a function of the linguistic system. In "speaker space", one speaker does not take a single static point, but an area considering all the variations (Rose, 2002: p. 29). For example, it would be a mistake to claim that two samples differing in pitch height were uttered by different people if there are other indications that the voice in one sample is angry and the voice in the other is not (Rose, 2002: p. 302). On the other hand, concluding that such samples were uttered by the same person would also be incorrect if there were no indications that such difference signified some paralinguistic or extralinguistic information in the given samples. In forensic practice, it is almost impossible to encounter the case where the known and questioned sample are recorded in identical conditions and the expert need not take into account the abovementioned "real-world" effects on the voice dimension in question (Rose, 2002: 35). This is exactly why Nolan (1999) underlines that we must acknowledge the limitations on any act of speaker recognition. Similarly, Coulthard and Johnson A. (2007) also remind that the members of the International Association of Forensic Phonetics and Acoustics always attach a "warning that their evidence should only be used corroboratively in criminal cases, because it is their collective opinion that it is not possible to establish the identity of a speaker with absolute certainty" (p. 202).

2.2.4. Speaker recognition by naïve listeners

In contrast to technical recognition (forensic speaker comparison by experts), naïve speaker recognition implies the application of our natural abilities as language users to recognise (identify) a speaker (Nolan, 2005). Nolan (2005) remarks that "given the sophistication of these abilities, the term naïve is perhaps inappropriate" (p. 386), yet, it is used

since it emphasises the lack of specific training of the person who is performing the recognition task.

Aural-perceptual approach to voice recognition has mainly been studied in light of earwitness evidence, the construction of voice line-ups or voice parades (Hollien, 1990; 2002; Nolan, 1999). Unconsciously, we perform voice recognition based on a composite analysis, processing the cues in speech relative to the speaker's sex, maturity, psychological or physical state, intelligence, social, economic, geographic, educational and other factors (Hollien, 1990: p. 191). Hollien (1990: p. 197-198) provides an analytical framework listing the elements of speech that cumulatively contribute to speaker recognition, including fundamental frequency, articulation (individual phoneme production), general voice quality, prosody (timing and melody), vocal intensity, and other speech characteristics (dialect, specific use of stress, idiosyncrasies in language use and pronunciation, speech impediments); however, he underlines that it is difficult to isolate and quantify the exact contribution of each aspect.

Depending on the nature of the task, speaker recognition may imply speaker identification or speaker discrimination, although, in the literature, the terminological difference is not necessarily maintained (O'Brien et al., 2021). Speaker identification implies that the listener is presented with the target voice prior to the recognition task and is subsequently asked to identify the speaker in a series of other voices (foils), the typical example being voice parades in earwitness testimonies (see de Jong-Lendle et al., 2015). In speaker discrimination, however, the listener is asked to assess the (dis)similarity of two speech samples (see Fleming et al., 2014). Another common approach to speaker recognition is the paired comparison technique, also known as the ABX (see Hollien & Schwartz R., 2000), in which the listener is presented with Speaker A, Speaker B and Voice X and asked to determine to which of the two speakers the unknown voice belongs.

The success and accuracy of speaker recognition depend on numerous factors related to both the recordings/speakers and the listeners. For instance, speech sample duration may affect voice recognition as longer samples tend to offer a more expansive phonemic repertoire (Hollien, 1990: p. 196; Yarmey, 1995). Using the paired comparison technique, Hollien and Schwartz R. (2000) demonstrated that using non-contemporaneous samples reduces the accuracy of the recognition experiment, with a very sharp drop in performance for longer delays (6 to 20 years). In addition, voice disguise, dialects/accents, speaking style, linguistic context and poor recording quality were also found to reduce recognition accuracy (see Atkinson, 2015; Das et al., 2020; Hollien & Schwartz R., 2000; Lavan et al., 2019; Nolan et al., 2008; Reich & Duke, 1979; Růžičková & Skarnitzl, 2017; Smith H. M. J. et al., 2018;

Stevenage et al., 2012; Tomić, 2020; Yu, 2019). Similarly, the negative impact can be observed when a larger number of speakers and speakers with similar voices are used for the procedure (Hollien, 1990: p. 197).

With regard to the listeners, it is generally held that gender is not related to recognition ability (Atkinson, 2015; Hollien & Schwartz R., 2000; Yarmey & Matthys, 1992). On the other hand, experiments have shown that voice memory gradually fades as time passes; the longer the period after the first encounter with the voice, the lower the accuracy (Clifford et al., 1981; Papcun et al., 1989). The decline in retention seems to depend on "an individual's ability to store information relative to both short-term and long-term memory" (Hollien, 1990: p. 195). Furthermore, the presence of familiar voices can significantly improve the listeners' performance (Papcun et al., 1989; Wenndta, 2016; Yarmey et al., 2001), which is also true for psychological stress/arousal associated with the circumstances under which the target voice was heard for the very first time (Hollien 2002: p. 34). Research has shown that non-native speakers of the target language perform worse than native speakers in recognition tasks (Cháfer, 2019; González Hautamäki et al., 2017; Köster & Schiller, 1997). In addition, numerous experiments have confirmed that professional training and experience will result in superior performance (Bartle & Dellwo, 2015; Hollien & Schwartz R., 2000; Schiller & Köster, 1998). In her PhD thesis, de Jong (1998: p. 116) provided evidence that even the musical aptitude of the listener, in particular, rhythm, tonal memory and timbre, may affect the recognition accuracy. Finally, despite the undisputable effect of all the previously mentioned factors, voice recognition ability is highly person-specific – some people score consistently high in recognition tasks across various circumstances (Aglieri et al., 2017; Bull & Clifford, 1984, as cited in Atkinson, 2015; Hollien, 1990; 2002; Künzel, 1994; Lavan et al., 2019). In literature, such listeners are often referred to as "super-recognisers" – the term initially coined by Russell R. et al. (2009) for people with an excellent ability to recognise faces.

2.3. Likelihood Ratio Approach

When forensic speech scientists provide the results of their analysis to the court, they are usually required to state their opinion on whether the disputed recording contains the suspect's voice. However, providing a categorical, binary answer to this question is not considered the best practice. The prevailing opinion is that forensic sciences, in general, must be probabilistic in the estimation of evidence (Aitken & Taroni, 2004; Aitken et al., 2021; Champod & Meuwly, 2000; Evett, 1991; Good, 1991; Nolan, 2001; Robertson & Vignaux. 1995; Rose, 2004; Taroni et al., 2006). As per this view, one should not claim that the evidence

shows that the two voices come from the same speaker; instead, one can only state "how much more probable the observed difference between questioned and suspect samples is, assuming that they have come from the same speaker, and assuming they have come from different speakers" (Rose, 2002: p. 46). Similarly, it may be tempting when assessing evidence to try to determine the guilt of the suspect; however, the odds of the suspect's guilt are solely the concern of the court – the job of a forensic scientist does not imply expressing opinions on the probability of the suspect's guilt, only on the probability of evidence (Aitken, 1995: p. 4 as cited in Rose, 2002: p. 69; Lindley, 1991: p. 42).

Considering the growing tendency towards expressing the results of forensic speaker comparison in a probabilistic manner, Bayesian Likelihood Ratio (LR) approach has become one of the leading methods in assessing forensic evidence. The framework is based on the Bayes' theorem (Bayes, 1763, as cited in Aitken & Taroni, 2004) and is considered objective as it guides the scientist to consider the evidential value under two competitive hypotheses, the *prosecution* and *defence hypotheses* (Champod & Meuwly, 2000). In forensic speech science, if the evidence supports the *prosecution hypothesis* (that the voices come from the same person), then it is said that the evidence is N times more likely to be observed were the samples from the same speaker, and if the evidence supports the *defence hypothesis* or *alternative hypothesis* (that the samples come from different people) then it is said that the evidence is *n* times more likely to be observed if the samples were from different speakers (Rose, 2002: p. 312).

2.3.1. Calculation and strength of evidence

Bayes' theorem observes the probability (*p*) of prosecution and defence hypotheses (H_p, H_d) given the evidence (*E*). Such a formulation contains the presupposition about the posterior odds in favour of a hypothesis, which are the product of the prior odds in favour of the hypothesis and the Likelihood Ratio (LR) calculation (Champod & Meuwly, 2000; Rose, 2004; Rose & Morrison, 2009):

$$\frac{pH_p|E}{pH_d|E} = \frac{pH_p}{pH_d} * \frac{pE|H_p}{pE|H_d}$$

The product of the equation above represents the latter probability, that is, posterior odds of the outcome, whereas the multiplier contains the information regarding the prior background knowledge regarding the case, i.e. prior odds. The multiplicand on the right of the formula is the Likelihood Ratio; it represents the ratio of probability (*p*) of evidence (*E*) given the prosecution and defence hypothesis (H_p, H_d) and is of vital interest to a forensic expert.

Namely, considering that forensic experts do not have information regarding the prior odds at their disposal, they can only estimate the probability of evidence, not the probability of the hypothesis – estimating the probability of the hypothesis should be the task of the court (Champod & Meuwly, 2000; Evett, 1991; Lindley, 1991; Rose, 2004; Stoney, 1991).

In Likelihood Ratio, the same-speaker hypothesis is determined by comparing the similarity between questioned and suspect samples, taking into account the intra-variability, while the different-speaker hypothesis is tested by estimating a relative frequency of the concordant features, i.e. their typicality, in a reference sample (Champod & Meuwly, 2000; Nolan, 2001; Rose, 2002).[3] The reference sample largely depends on the case in question (Champod & Meuwly, 2000) and should be comprised of the recordings similar to the one of the offender in terms of speaker sex, age, dialectological background, recording conditions and other relevant parameters (Rose, 2002). The reference sample should ideally be comprised of recordings that are not employed in the test itself, however, due to limited resources, scientist often rely on a leave-one-out cross-validation technique where the reference sample is comprised of all the speakers' values except of the ones that are being compared (e.g. Kinoshita, 2001; Li & Rose, 2012; Tomić & French, 2019). If the number obtained in the end is larger than unity (1), we have evidence supporting the prosecution hypothesis. On the other hand, if the number is below unity, the formula implies that the evidence is more probable given the defence hypothesis (Evett, 1991). Furthermore, by multiplying the LRs obtained from different parameters, we may reach the *overall likelihood ratio* (OLR), which is particularly useful since individual likelihood ratio values are often too small to reach meaningful conclusions (Robertson & Vignaux, 1995). The process of combining LR values is not entirely straightforward, however, as numerous experiments have confirmed that prior to the combination of parameters, it is essential to account for existing correlations between them (Gold, 2014; Nair et al., 2014; Rose, 2006; 2013b). The issue of parameter correlation is properly addressed by Multivariate Kernel Density Likelihood Ratio formula (MVKD) by Aitken and Lucy (2004) that is commonly applied in forensic sciences. The formula is also recommended because "it models two levels of variation, […] allows for non-normal between-group distribution and the results are not extreme" (Aitken & Lucy, 2004: p. 18) The methodology has become a staple in forensic speaker comparison research in the past decade as it has been applied in countless studies (e.g. Frost & Ishihara, 2015; Gold, 2012; 2014; Gold

[3] The approach was first applied to measuring glass refractive index and later to DNA analysis (see Evett, 1991).

et al., 2013; Ishihara, 2017; Ishihara & Kinoshita, 2008; Kinoshita, 2014; Lo, 2021; 2021b; Morrison, 2009b; Rose, 2013; 2015; 2017; Rose & Wang, 2016; Tomić & French, 2019), as well as in the casework (e.g. Rose, 2013b; 2022).

The output of MVKD likelihood ratio is a score that supports either the same-speaker or different-speaker hypothesis – the raw scores, however, may need to be calibrated before they are interpretable. Logistic regression calibration (Brümmer & du Preez, 2006) has become a customary method of converting LR scores to interpretable logLRs by performing a linear shift (in the logarithmic scale) on the scores relative to a decision boundary (Frost & Ishihara, 2015; Morrison, 2013). The weights involved in the shift are calculated by using a training set of data, that is, running sets of known-origin pairs through the system to obtain scores, resulting in a development of a training model. Ideally, the training data would not be compiled of the same recordings as the testing data; however, similar as with reference population, scientists often resort to a cross-validated procedure where each derived score is referenced against every other score in the database to produce the weights (see Frost & Ishihara, 2015; Morrison et al., 2012).

System performance under the Likelihood Ratio framework is often evaluated through measures of Equal Error Rate (EER) and log-likelihood ratio cost (C_{llr}). Irrespective of the chosen metric, the validity is estimated by running a large number of same-speaker and different-speaker samples through the system, each time noting whether the output was in accordance with the expectations or not, that is, whether the system correctly identified the same-speaker and different-speaker samples (Morrison, 2011). The likelihood ratio value of 1 (or 0 for the logarithm of the likelihood ratio) is chosen as a threshold for the proposition under the same-speaker and different-speaker hypotheses (Aitken et al., 2021: p. 959). A value describing the average performance over all of the test pairs is taken as an overall system validity (Morrison, 2011). An error when the system mistakenly identifies two different-speaker samples as originating from the same speaker is called false positive (false acceptance or false alarm); conversely, an error when the system fails to detect two same-speaker samples is termed false negative (false rejection or missed hit) (Aitken et al., 2021: p. 959-960; Brümmer and du Preez, 2006: p. 231; Cubic & Buscaglia, 1991: p. 216; Good, 1991: p. 91; Morrison, 2011: p. 93; Rose, 2002: p. 97). From a judiciary perspective, false positives (FP) are considered to have more severe consequences than false negatives (FN) (Rose, 2002: p. 28).

Equal Error Rate (EER) represents an operating point where probability of a false positive is equal to that of a false negative (Brümmer and du Preez, 2006; Bhattacharjee & Sarmah, 2012). Another possible measure is the half total error rate (HTER), which is the

average of the two (Frédéric et al, 2004, as cited in Bhattacharjee & Sarmah, 2012: p. 378), but apart from the convenience of calculation, the number itself does not reveal much regarding the validity of the system. Equal Error Rate for the particular set of measures is obtained through manipulating the threshold of acceptance (τ) for same-speaker and different-speaker hypotheses ratio until the percentage of false positives is equal to that of false negatives (Brümmer and du Preez, 2006). Equal Error Rate is often illustrated on a *detection error trade-off* (DET) plot, which illustrates how the rates of false positives and false negatives are inversely proportional (Aitken et al., 2021: p. 960). However, since EER is based on a categorical threshold (whether the system has correctly identified two samples as originating from the same/different speakers or not), and because of the ongoing tendency of expressing the outcomes of a forensic speaker comparison in a probabilistic manner, EER is often described as a useful metric for the overall discriminability of a system but is seen as inappropriate for the evaluation of the system and strength of evidence (Frost & Ishihara, 2015; Morrison, 2011).

Another measure, that is more in accordance with the contemporary probabilistic tendencies in assessing the strength of evidence is log-likelihood ratio cost (C_{llr}) (Aitken et al., 2021; Brümmer & du Preez, 2006; Morrison, 2011). It is a continuous measure that evaluates the strength of likelihood ratio output by considering the magnitude of consistent-with-fact (and contrary-to-fact) LR values, and assigns them appropriate penalties (or *cost*); the larger the value of misleading evidence, the higher the penalty/cost (Aitken et al., 2021: p. 961; Frost & Ishihara, 2015: p. 44; van Leeuwen and Brümmer, 2007: p. 344) For instance, an erroneous $\log_{10}LR = -5$ for a same-speaker comparison would have a higher C_{llr} than $\log_{10}LR = -0.5$ for the same pair. In contrast, in different-speaker comparisons, the lower the number, the stronger C_{llr} score. It is generally considered that C_{llr} value less than 1 implies that the LR output is reliable, and the system validity increases as C_{llr} approaches 0 (Frost & Ishihara, 2015: p. 44). In C_{llr} calculation, it is assumed that the prior probabilities of the two propositions of same source (H_p) or different source (H_d) are taken to be equal: $p(H_p) = p(H_d) = 0.5$ (Aitken et al., 2021: p. 961).

Among others, some of the common means to assess the likelihood ratio system performance involve probability distribution (histograms) and Tippet plots. With probability distribution, the discriminating power of a method at a particular value of log(LR) is the amount of overlap of the distributions for data from H_p and H_d at that value. If there is no overlap, then there is 100% discrimination; which is rare considering that log(LR) values are the continuous data. Conversely, if there is no separation, then one distribution is entirely encompassed within the range of the other, which indicates absence of discrimination (Aitken et al., 2021: p. 956).

Tippett plots, on the other hand, are generalisations of rates of contrary-to-fact evidence in comparisons, the x-axis representing the log(LR) plotting and y-axis the proportion of comparisons greater than a particular value (in percentage). For instance, in same-speaker comparisons, it is to be hoped that all log(LR) values are greater than 0, thus for $x < 0$, the optimal scenario is that the corresponding value on the y-axis will be 100%. The distance from the intersection of the same-speaker plot with the line log(LR) = 0 and the line y = 100% is the percentage of false negatives. The inverse is true for false positives (Aitken et al., 2021: p. 957-958).

2.3.2. Verbal expression of likelihood ratio

Champod and Meuwly (2000) warn that "the calculation of the LR, however, is not a `Bayesian analysis', as this term usually implies the assignment of prior probabilities" (p. 200). They note that scientists do not usually have access to the background information available to a jury or a judge and, therefore, cannot assess the prior odds correctly. Consequently, forensic scientists cannot provide an opinion on the posterior odds implied by the Bayesian analysis (p. 201). The observation is particularly relevant to the verbal expression of the forensic results to the court.

Namely, one of the frequent ways of expressing the outcome of forensic speaker comparison to the court includes a variety of ranked probability scales (e.g. Baldwin and French 1991: p. 10 as cited in Broeders, 1999: p. 229; French, 2017: p. 7; Gfroerer, 2003: p. 708), also known as classical probability scales (Gold & French, 2019). However, as Champod and Evett (2000) explain, "scales that use terms such as 'probable', 'highly probable' or 'with a probability bordering to certainty', invoke statements of the posterior odds on the issue" (p. 238), combining "the effect of the scientific observation and an assessment of the prior probability that encapsulates all the other evidence available to the court" (p. 238). It is why they propose the reporting convention based on the likelihood ratio calculation: "the […] evidence supports the proposition that […]", thus avoiding taking a position on the posterior probability (p. 239).[4]

Champod and Evett's (2000) interpretation of the numerical likelihood ratio is hitherto one of the most cited verbal scales for expressing the outcome of forensic analysis to

[4] For a more detailed account of logical fallacies, formation of hypotheses, implementation of Bayesian principles to forensic speaker comparison and debate on the expression of the outcome, the reader is advised to consult Champod and Evett (2000), Champod and Meuwly (2000), French and Harrison, (2007); French et al., (2010); Morrison (2016), and Rose and Morrison (2009).

the court and is actively used by multiple forensic speech science practitioners (see French, 2017; Gold, 2014; Gold & French, 2019; Rose, 2002;). The recreation of their proposition can be observed in Table 2-1:

Table 2-1
Verbal expression equivalents of likelihood ratio values

Likelihood Ratio	***Strength of Evidence Expression***	***Hypothesis***
>10 000	Very strong evidence to support...	Prosecution hypothesis
1000 to 10 000	Strong evidence to support...	
100 to 1000	Moderately strong evidence to support...	
10 to 100	Moderate evidence to support...	
1 to 10	Limited evidence to support...	
1 to 0.1	Limited evidence to support...	Defence hypothesis
0.1 to 0.01	Moderate evidence to support...	
0.01 to 0.001	Moderately strong evidence to support...	
0.001 to 0.0001	Strong evidence to support...	
<0.0001	Very strong evidence to support...	

Note: The table is adapted according to the verbal expressions in Champod and Evett (2000)

Champod and Evett (2000) do acknowledge the weaknesses of the scale, such as having to explain its meaning to the jury or court, a categorical classification of the values that are, in essence, continuous, its inadequacy to distinguish between very high or very small likelihood ratio values and a difficulty to combine other evidence the strength of which is expressed verbally (p. 241). As an alternative, the authors suggest employing the logarithmic form of the likelihood ratio, also used for expressing the power of earthquake or sound (Good, 1950 as cited in Champod & Evet, 2000: p. 241). A logarithm of the likelihood ratio with a value greater than 1 is a positive number, while a logarithm of the likelihood ratio between 0 and 1 is a negative number. Such a scale is considered symmetrical and thus more intuitive to be applied in the legal process (Aitken et al., 2021).

The impact of likelihood ratio usage for forensic speaker comparison purposes can be witnessed in the significant growth of the research on this methodology. According to Gold's (2014) literature review, at the time, likelihood ratio was primarily applied in forensic research but there were not many reports of its application in forensic practice. In research, the framework was used either to test potentially useable forensic speaker comparison parameters or for methodology improvement and revision (Gold, 2014: p. 60). Gold (2014) notes that, across studies, the likelihood ratio framework is mainly employed with vowels and formant-based numerical values and only occasionally with fundamental frequency or voice onset time.

She cites Rose's (2012; 2013b) report on a fraud case in Australia as a single publication on incorporating this methodology in forensic speaker comparison casework.

The trend has continued in the past decade as well, yielding numerous studies on the likelihood ratio methodology testing and improvement (e.g. Enzinger, 2016; Hughes, 2017; Hughes & Foulkes, 2014; Kinoshita & Ishihara, 2014; Meuwly et al., 2017; Morrison et al., 2014; Morrison & Enzinger, 2018; Nair et al., 2014; Xiao Wang et al., 2019; in press) as well as in forensic parameter testing. In addition, the latest survey on forensic practices by Gold and French (2019) revealed a slight rise in the application of numerical LR (13.2% of experts as opposed to 8.6% in Gold & French, 2011) and a significant rise in the application of verbal LR (26.3% of experts as opposed to 11.4% in Gold & French, 2011) as conclusion framework in casework. Apart from the parameters related to vowel and formant values (e.g. Diesner & Ishihara, 2016; He et al., 2019; Heeren, 2020; Irfan et al., 2021; Jessen, 2021; Rose, 2015; Tomić & French, 2019), the methodology has been used to explore temporal parameters of speech (e.g. Gold, 2014; Hughes et al., 2016; Tomić, 2017), various aspects of consonants, including plosives (e.g. Earnshaw, 2016) and fricatives (e.g. Rose, 2022), higher-level features, such as tone (e.g. Rose, 2017; Rose & Wang, 2016), click rate (e.g. Gold, 2014) or even to evaluate authorship attribution (e.g. Ishihara, 2014; 2017). The research on the Likelihood Ratio framework in forensic speaker comparison has flourished with the development of automatic speaker recognition (ASR) systems. Numerous researchers have worked on testing and improving the methodology as well as examining various effects on the strength of evidence in FSC with the help of ASR systems (see Drygajlo et al., 2003; Franco-Perdoso & González-Rodríguez, 2016; González-Rodríguez et al., 2002; 2003; 2006; Kelly & Harte, 2015; Kockmann et al., 2021; Morrison et al., 2020; 2022; Rhodes, 2017; Sztahó et al., 2021; Wang & Zhang, 2015; Watt et al., 2020; Xiao Wang & Hughes, 2022). The increased presence of likelihood ratio methodology in both research and practice indicates the increased awareness of the benefits of such methodology over the binary decision or classical probability scales.

Even though the arguments against the Bayesian Likelihood Ratio approach are hardly sustainable nowadays, whether this methodology can completely substitute other conclusion frameworks in the forensic speaker comparison practice is negotiable. Nolan (2001) noted that some experts at the time did not find it necessary to compare the questioned and suspect samples against the reference population, but only against each other. Also, someone may argue that Bayesian LR is too complex for expressing the results in court or too complex for calculation (see Evett 1991 for his elaboration on communication between the expert witnesses and the court of law). Broeders (2001) believes that applying the LR approach could

be unrealistic because experts find it difficult to adequately express their findings for the court to understand their actual meaning, especially if the jury is involved. He also points out the necessity to have data on the statistical distribution of relevant parameters in the relevant reference population, which is quite challenging to obtain (Broeders, 2001). This opinion is seconded by French (2017), who agrees that most of the features are subject to regional, social and ethnic variation, as well as change over time and that even if we had unlimited research resources at our disposal, it would not be possible to establish distributional information for every analysable feature for every variety at every period. A complete transfer to the likelihood ratio methodology is also aggravated by the fact that forensic speaker comparison still depends on qualitative analysis, and some qualitative features are challenging to quantify.

Nolan (2001) notes that, despite the abovementioned challenges, the LR approach is the right direction for developing forensic speaker comparison practice. Namely, experts should always evaluate the evidence in probabilistic terms, stating how likely it is to observe the evidence given the prosecution and alternative hypothesis and not succumb to the pressure of giving categorical statements (Nolan, 2001). French (2017) agrees that even the limited availability of reference population values increases the objectivity of the assessment of distinctiveness of the analysed features and, therefore, the objectivity of the interpretation of findings, "even if our conclusions have to remain as opinions - in verbal rather than numerical form" (p. 13). Finally, despite the challenges this methodology may pose when presenting results to courts, the numerical likelihood ratio framework remains one of the most objective methodologies for testing and evaluating the effectiveness of acoustic features in forensic speaker comparison experiments, which is why it will be employed as an instrument in the present study.

2.4. Cross-Language Forensic Speaker Comparison

2.4.1. Current practices and reasoning

Even though up-to-date literature describes just a few example cases concerning cross-language forensic speaker comparison, practitioners are fairly familiar with this type of requirement in their forensic laboratories. For instance, Künzel (2013) wrote that, in his forensic practice, "the majority of speaker-recognition cases involve[d] speech material from more than one language" (p. 22). Similarly, in her paper on the examples of FSC casework, Wagner (2019) confirms that, at the Bundeskriminalamt, they do encounter cases with language mismatch. However, forensic speaker comparison surveys published so far do not reveal much information regarding the actual amount of such casework.

Künzel (2013) presented a case in which the police intercepted phone calls in Igbo that revealed incriminating evidence regarding illicit drug deals. However, after the suspect was apprehended, he claimed that his phone had been stolen and that the only languages he could speak were (Nigerian) English and a little German. In addition, he invoked his right not to deliver a speech sample for analysis, which left the police no choice but to compare the incriminating recordings in Igbo and several calls in German the suspect had made to the social welfare department (p. 22). Twenty years before, in 1992, in a case that was only recently brought to light by Lo (2021), a merchant from Toronto was threatened anonymously over a phone call by a man speaking English with a Cantonese accent. The recipient identified a potential caller, and the police apprehended the suspect. Subsequently, the defence presented the linguistic analysis of segmental features of the suspect's speech and expressed an opinion that the offender's accent was "not as strong" as the suspect's Cantonese accent, resulting in the suspect's acquittal (Rogers, 1998, as cited in Lo, 2021: p. 24). Two other cases of cross-language forensic speaker comparison that occurred in Australia in 2002 and 2007 are presented in Edmond et al. (2011). Both cases have piqued the public and scientific interest due to the notorious circumstance of how the court approached the speaker comparison task. In the former case, a Vietnamese appellant was found guilty of heroin importation, among other evidence, based on incriminating phone calls. What renders the case preposterous is that the speaker comparison of the incriminating recordings in Vietnamese and known samples, some of which were in English, was performed by the court interpreter and the jury in a rather layman's fashion. The latter case concerned an appellant similarly convicted of cocaine importation. Namely, several incriminating telephone recordings in Igbo and a known recording in English were played to the jury to decide whether it was the same person or not. A more detailed description of these cases and the quotes from the trial transcript explaining the reasoning behind these procedures are available in Edmond et al. (2011).

In a talk at the International Association of Forensic Phonetics and Acoustics conference in 2019, Milne et al. (2019) reviewed the voice-related case requests received during 2017 and 2018 in three forensic laboratories, the NTF's Speech and Audio Group (Netherlands), the RCMP's Audio and Video Analysis Unit (Canada), and the University of Zurich's Centre for Forensic Phonetics and Acoustics (Switzerland). One of the aspects they examined is the language mismatch between the suspect and questioned samples. Unfortunately, the authors did not publish the survey results after the conference; therefore, we cannot specify the percentage that language-mismatch cases take in the overall caseload.

Furthermore, two major surveys on forensic practices, Gold and French (2011; 2019), published at the beginning and the end of the previous decade, respectively, primarily focus on the methodologies and analytical framework in FSC while omitting to report on the type of casework that is performed across the forensic laboratories. Gold and French (2011) mention that 56% of the surveyed practitioners disclosed that they have worked with samples in foreign languages; however, it is not specified whether, in these cases, both samples were in the same language or it was cross-language analysis (p. 300).

At the beginning of the century, in his comprehensive book on forensic speaker identification theory, practice, and methodology, Rose (2002) wrote:

> "Sometimes, a forensic comparison is requested between samples in different languages. [...] Unfortunately, not enough is known yet about bilingual speakers to say whether any voice quality remains the same across two samples of the same speaker speaking in two different languages or dialects. Most likely it will depend on how good a command the speaker has of both varieties. Until we have a much better knowledge of this area, cross-linguistic forensic comparison is clearly counter-indicated."
>
> (Rose, 2002: p. 342)

Two decades later, there seem to be no clear guidelines or principles regarding the practices when performing cross-language forensic speaker comparison. Namely, in the *Best Practice Manual for the Methodology of Forensic Speaker Comparison* by the European Network of Forensic Science Institutes (ENFSI, 2021), there is only a brief article mentioning cross-language FSC in the context of the comparability of speech material:

> "[…] mismatch of spoken languages between the questioned and the reference material could play a substantial role. It limits the number of phonetic-linguistic features that could, in principle, be analysed. Caution should therefore be exercised in analysing cases involving language mismatch."
>
> (ENFSI, 2001: p. 14)

A similar warning is provided in Article 3.10 of the IAFPA Code of Practice (IAFPA, 2020), stating that "members should exercise particular caution with cross-language comparisons." However, both documents fail to disclose what "caution" actually implies in this context. A more detailed elaboration on how to approach FSC in language mismatch conditions is given by Drygajlo et al. (2015) in the context of automatic speaker recognition:

> "Language mismatch does not generally preclude the application of FASR and FSASR methods because vocal tract characteristics and prosodic phenomena can remain fairly stable across first and second language. However, the language structure itself can impose its influence on the features, for example the system of vowel phonemes and their phonetic implementation in a language has an influence on formant frequencies and MFCCs. The extent to which the factors mentioned above influence FASR and FSASR has to be determined empirically through method validation, either without mismatch compensation or, if possible, with application of mismatch compensation methods."
>
> (Drygajlo et al., 2015: p. 35)

Mismatch compensation methods that Drygajlo et al. (2015) mention include various statistical procedures based on feature extraction, feature modelling and similarity

scoring (p. 15). However, as above, in this case, language mismatch seems to be treated the same way as the problem of recording quality or device mismatch, which could be described as an oversimplification. Namely, with regard to cross-language FSC, the features that need to be analysed may depend on multiple factors, including but not limited to the language in question and foreign language proficiency. Unfortunately, the amount of available research is barely enough to scratch the surface of the problem, let alone provide some universal principles when conducting cross-language FSC, which is why the contemporary manuals/codes on the topic remain vague.

Despite the lack of official positions and guidelines on cross-language FSC, there is obviously no shortage of practice. However, as seen from the reviews above, the practitioners appear reluctant to disclose many details concerning such cases. Understandably, one of the reasons must be the sensitivity of the data, yet, another may be the fear of critique by the scientific community, as a significant part of such analysis must be based on a subjective decision-making process driven by the experts' previous experience.

In order to push the cross-language forensic speaker comparison from the margins of forensic speech science and demystify the "fog of caution" encircling it, we need structured, scientifically driven research and practice principles – especially considering that cross-language FSC has already been undertaken for years. In the present chapter, the challenges of FSC under language mismatch are approached analytically. First, we will review a selection of the cross-language forensic speaker comparison research undertaken so far, and then we will raise some issues and elaborate on the implications it has for the present study.

2.4.2. Previous research

While forensic speaker comparison research dates back to the first half of the 20th century, studies concerning FSC with language mismatch only emerge much later. What sparked a deeper interest in cross-language forensic speaker comparison is the development of software for automatic speaker recognition (ASR). Traditional phonetic parameters such as fundamental frequency (f_0), local and long-term formant frequencies (LTFs), formant dynamics, temporal aspects of speech (articulation rate, pauses, hesitation), and consonants have mostly come into focus in the past decade. For clarity and convenience of future reference, the presented research will be grouped according to the mentioned topics and described in the sub-sections below. The previous research examination, however, begins with a brief review of cross-language speaker recognition studies by naïve listeners.

Cross-language speaker recognition by naïve listeners

A wide range of research has confirmed that listeners recognise voices better in their mother tongue than in a second or a foreign language, the phenomenon known as the "language-familiarity effect" (Perrachione, 2019: p. 516). The effect was first described by Hollien et al. (1982, as cited in Perrachione, 2019: p. 519), and research has shown that it can be observed regardless of the number of included voices, the languages spoken, the nature of the recognition task, previous exposure to the target voices, delay in exposure and test administration or content (Perrachione, 2019: p. 518). One of the possible explanations for this effect is that our "memory for voices is encoded via 'schemata' that consist of norms for all aspects of a language, including its syntax, lexicon, and phonology […] learned through exposure to voices in a local area"' (Goggin et al., 1991, as cited in Perrachione, 2019: p. 520).

In discrimination tasks, where listeners were presented with pairs of voices to decide if it was the same person, native English listeners performed better with English-speaking samples than German-speaking samples, deteriorating even further with cross-language comparisons (Winters et al., 2008). The authors conclude that the listeners rely on both language-dependent and language-independent information in the speech signal to perform discrimination. Wester (2012) obtained similar results for German-, Finnish- and Mandarin-speaking voices and Mok et al. (2015) for Cantonese-English. The lower discrimination performance seems to stem from the fact that subjectively perceived similarity between different voices tends to be higher for a foreign or unfamiliar language, not due to comprehension but rather because of the familiarity with the phonology of one's native language, analogous to the "other-race" effect in face recognition (Fleming et al., 2014). There is some indication that the holistic perception of voice quality may outweigh the language-familiarity effect, provided the listeners are presented with short stimuli (3-second recordings). Namely, in the experiments with Spanish-speaking monozygotic twins, San Segundo et al. (2016; 2017) found that twin pairs were consistently rated as more similar than non-twin pairs by both the Spanish and English or German listeners. When asked to judge cross-language voice pairs, however, listeners tend to mark them as sounding more distinct than either foreign or native within-language pairs (Fleming et al., 2014).

Numerous studies have explored the effect of listeners' foreign language proficiency, age of first exposure and immersion in the foreign language community on the speaker recognition ability (Köster & Schiller, 1997; Sullivan & Kügler, 2001; Sullivan & Schlichting, 2000). The results generally suggest that earlier acquisition or greater exposure to a second language can improve people's ability to recognise voices speaking in that language,

the exposure sometimes playing a more significant role than the actual ability to speak the language (Goggin et al., 1991; Orena et al., 2015).

As most of the cross-language perceptual research is focused on the listeners, few studies deal with the effect that speakers' language proficiency and fluency can have on the listeners in voice recognition or discrimination tasks. In addition, very few studies exploring cross-language speaker recognition have considered the inherent features of voice, such as voice quality, which may significantly influence voice perception. For instance, Das et al. (2020), working with native and accented English, found that the effect of voice quality is five times as large as that of the non-native accent in speaker recognition, but, surprisingly, the effect disappears when speakers share the same (native) accent.

Cross-language Automatic Speaker Recognition

Automatic speaker recognition systems are often described as text-independent as they do not rely on language-specific "high-level" features such as dialect, sociolect, intonation patterns, phonetic and linguistic parameters of hesitations (Künzel, 2013); instead, they extract "low-level" spectral envelope features, such as MFCCs – Mel Frequency Cepstral Coefficients, PLPCCs – Perceptual Linear Prediction Cepstral Coefficients and LPCCs – Linear Prediction Cepstral Coefficients (Drygajlo et al., 2015) that are supposed to be characteristic of the general resonance behaviour of the vocal tract of a speaker. Depending on the type of material, feature extraction method and statistical procedures involved, researchers have presented different outcomes concerning cross-language forensic speaker comparison with the aid of ASR systems.

Some of the earliest studies to examine the performance of automatic speaker recognition software in multilingual circumstances were performed on short sequences of read-out speech, with the error rates consistently deteriorating when different training and testing languages were used. For instance, Durou (1999), who recorded 82 native speakers of Dutch in four languages (Dutch, English, French and German), relying on LPC feature extraction, reached perfect system performance (EER 0%) with same-language pairs, as well as with Dutch and English as a pair; however, the results obtained for Dutch/German and Dutch/French were slightly weaker (around 2% and 5%, respectively). Similar performance was perceived by Faundez-Zanuy and Satué-Villar (2006), who compared 49 bilingual speakers of Catalan and Spanish by extracting LPCC features and relying on two different speaker modelling methods, vector quantisation and covariance matrices. The former speaker modelling technique yielded better results, up to 100% of correct identifications for same-language and 99.6% for different-language pairs. Notably, inferior results (the overall system performance of 85.74%) were

obtained by Kumar et al. (2009), who tested 25 people speaking four Indian languages and English by extracting multiple LPC and Reflection Coefficient (RC) features and analysing the results using Neural Network Model (Kumar et al., 2009). Similarly, Luengo et al. (2008), in their study with 22 speakers of Spanish and Basque, combine the prosodic features (intonation and absolute energy extracted every ten milliseconds alongside their first and second derivatives) with MFCC features to account for language-specific variation. The identification accuracy drops dramatically when the training and testing language are not the same (63.55% and 67.34% v 98.34% and 97.29%); however, if double-language training is performed, the results are very close to those obtained in the same-language condition (96.77% and 95.58). Finally, somewhat improved results were obtained when 200 speakers with some of the Arunachali languages of North-East India as their mother tongue were recorded in English, Hindi and their native language. Relying on MFCC feature extraction, the system reached the performance of EER of 4.55% for same-language pairs and up to 11.36% for different-language pairs (Bhattacharjee & Sarmah, 2012). Namely, the system performed poorer when English was compared to any language; however, it performed equally well when the same languages were compared as when Hindi was compared to some of the local languages. The authors believe that the reason for this is the fact that Hindi and the local languages are spoken in a similar manner (Bhattacharjee & Sarmah, 2012). Nagaraja and Jayanna (2013) performed single-language and cross-language speaker comparisons of 30 Kannada speakers who could also speak Hindi and English. Surprisingly, the best results were obtained for single-language comparisons with English samples, even though it was not the participants' native language. The researchers suspect that the reason for this is the presence of unvoiced consonant clusters in the Kannada corpus that resulted in fewer frames/features for the analysis (p. 19).

The results obtained by these early studies seem attractive, with a rather strong system performance and minimal equal error rates. However, one must understand the nature of corpora used to train and test the systems in this pioneering research. Namely, the speech material was comprised of highly structured, phonetically balanced read-out text, not varying across participants. In forensic reality, experts are frequently engaged to examine a small portion of spontaneous speech that, apart from language mismatch, may have numerous inconsistencies across the questioned and known sample. Therefore, the results presented above need to be considered in light of limitations imposed by the corpora they employed.

More recent research concerning automatic speaker recognition in a realistic forensic setting has brought new answers and, at the same time, raised more issues on the topic. In their review paper, Nagaraja and Jayanna (2016) examine available feature extraction and

modelling techniques used by contemporary software in multilingual ASR. Using one such system, Batvox 3.1 (Agnitio, 2009), Künzel (2013) examined the spontaneous speech of 75 bilingual speakers (German-Russian, German-Polish, German-English, German-Spanish, German-Chinese and Spanish-Catalan) recorded over a microphone, a landline phone and a GSM phone. The EERs of his comparisons are predominantly below 1%. Slightly higher error rates are noted for German-Spanish and Spanish-Spanish pairs recorded over the GSM phone (5.9% and 5%). In addition, in 9 out of 20 scenarios, the cross-language condition involves slightly lower EERs than the corresponding same-language condition, whereas the opposite relation occurs in two cases, with English-German and Chinese-German recorded over the GSM phone. The overall system's performance appears to be reasonably reliable in direct recordings; however, it deteriorates when landline and GSM phones are used (Künzel, 2013). Alamri (2015) used the same system to compare speech samples in various Arabic dialects and English; however, the author does not report error rates; instead, he focuses on the potential problems that could arise due to speech sample quality. Jovičić and Grozdić (2014) examined Speech Interactive System (Speech Technologies Center, n.d.) with three native speakers of Serbian who were also able to speak English and Hungarian, the system being able to confirm the identity only of the person with strong Serbian-accented English. As ASR methodology improvement, Askar et al. (2015) propose a linear transform approach that projects speech signals from one language to another so that the language mismatch between samples is normalised. They evaluate its efficiency with 113 female speakers of Standard Chinese and Uyghur, concluding that the proposed approach can achieve up to 10% improvement in the EER scores. Van der Vloed et al. (2017) used the recordings of native Dutch and native Turkish speakers to examine the test and reference data language mismatch in two ASR systems. The researchers observe the rise in log LR scores, which they refer to as the "right shift" pattern and conclude that the automatic forensic speaker comparison with mismatched reference population may be used with caution. Kahil et al. (2018) tested ALIZE/LIA_RA, an open-source toolkit, with 52 native speakers of Arabic also fluent in English. According to their findings, the error rates for same-language training and testing were close to 7% and 8%; however, these numbers increased in the mismatched conditions (around 12%). Most recently, Saleem et al. (2020) propose implementing a tool for extracting accent and language information (Accent Classification – AC and Language Identification - LI) from short utterances. Their results show that with the x-vector feature extraction method, the ASR system reached an accuracy of 80.4%, while AC achieved 85.4% and LI - 90.2%. The combination of the AC and LI methods yielded an accuracy of 95.1%, which, the researchers conclude, is a promising result

(Saleem et al., 2020). In his doctoral dissertation, among other parameters, Lo (2021) investigates the performance of a contemporary ASR system (Phonexia Voice Inspector v4.0) on a single-language and cross-language corpus of 60 Canadian French and English bilinguals. Under very well-controlled conditions, the researcher obtained perfect identification (EER 0%) and stellar C_{llr} scores for same-language comparisons (0.0047 in English; 0.012 in French) and very reliable results for cross-language comparisons (EER up to 4% for English-French and up to 0.5% for French-English, C_{llr} scores up to 0.15). The effect of language mismatch, however, varies between speakers (p. 265). The researcher does not provide potential reasons for deteriorated speaker recognition of individual speakers, however, as the study observes bilingualism in a broad view, the proficiency, fluency and accent influence were not considered as a variable in the study, and this may be precisely why for specific speakers, the ASR system yielded higher C_{llr} scores.

Dependency of ASR system performance on individual voices and universality of methods and results across datasets and languages is currently being investigated by Dr Vincent Hughes and colleagues at the University of York as part of the project Person-Specific Automatic Speaker Recognition (Hughes et al., 2022a; 2022b).

Vowel-based parameters in cross-language FSC

Some of the earlier work on formant-based cross-language FSC includes Heeren et al. (2014), who analysed LTF2 and LTF3 for 12 speakers of Dutch and Turkish, concluding that within-speaker variability between languages is lower than between-speaker variability within a language. Next, Meuwly et al. (2015) compared LTF2 and LTF3 of a single bilingual speaker of Dutch and Turkish under the Likelihood Ratio framework within each language and across languages, respectively. While the system was able to confirm the speaker's identity both times with samples in the same language, the identification in the cross-language comparison was unsuccessful (Meuwly et al., 2015).

Krebs and Braun (2015) analysed the LTFs of 16 bilingual speakers of German and French and noted small but significant differences in the values between the two languages. They identify the correlation between the two sets of values and single LTF3 out as the steadiest parameter with the greatest between-speaker and lowest within-speaker variability.

Frost and Ishihara (2015) built an FSC system based on formant frequency values measured from the trajectories of the vowels and surrounding segments to compare the speech of 15 Hong Kong Cantonese speakers when speaking Cantonese and English. The comparisons were made on the bases of selected vowels in six predetermined target words. The reported

error rates are quite low (below 2%), and the obtained log likelihood-ratio cost (C_{llr} between 0.158 and 0.527) is comparable to the performance of similar systems designed for monolingual comparisons (p. 46).

Zuo and Mok (2015) analysed formant dynamics of the first four formants in Shanghainese-Mandarin bilingual identical twins, concluding that the differences between the twins were significant enough to discriminate them using Discriminant Analysis. In addition, the differences between the twins became more prominent in their non-dominant language (p. 1).

Cho and Munro (2017) explored f_0, LTFs and long-term average speech spectra (LTAS) in 10 L1 Korean and L2 English speakers. According to their results, LTFs seem to be most speaker-specific, while f_0 may vary across styles. Finally, LTAS appeared to be most similar across languages for a speaker with low language competence (p. 5).

Some aspects of fundamental frequency in cross-language FSC were also explored by Dorreen (2017), who split creak and modal phonation into separate distributions to obtain more precise results. According to this researcher, the antimode of f_0 is a most promising parameter as it exhibits greater between-speaker than within-speaker variability. The corpora used in this study included Maori speakers of New Zealand English and a variety of European and East Asian languages (p. 24).

More recently, Tomić and French (2019) compared the performance of LTF1-LTF4 under the Bayesian Likelihood Ratio framework, analysing the speech of 35 native speakers of Serbian fluent in English. The researchers obtained higher EER and C_{llr} scores when comparing samples in Serbian and English than when the suspect and offender samples were in the same language. The combination of all four LTFs yielded the lowest EER (around 11%), while the lowest C_{llr} score of 1.2494 was noted for LTF2. The researchers conclude that there is a bias towards different-speaker identification errors ('missed hits') in cross-language comparisons (p. 30).

A slightly different approach was taken to vowel formants by Zhong (2019), who compared the F1 and F2 values of individual vowels in Chinese-English bilinguals under the LR framework. The lowest C_{llr} was obtained for the second formant of the vowel /i/, followed by the first formant in vowels /ə/ and /u/. The researcher also discusses the choice of the reference population, which, according to his results, largely depends on the parameters in question (p. 61).

Next, applying the analysis of variance, Tomić (2020) compared the long-term formant values of 10 native speakers of Serbian (L1) and foreign English (L2). The results

revealed significant differences between speakers in the four tested formants, with LTF3 and LTF4 exhibiting significantly higher between-speaker than within-speaker variation across the two languages.

Most recently, using a corpus of Canadian English–French bilinguals, Lo (2021b) examined the impact of language mismatch on the performance of long-term formant distributions (LTFD) in FVC under the LR framework. Despite the noted impact of language mismatch on the system performance, the discriminatory potential of LTFDs should not be underestimated as C_{llr} scores always remained below 1: 0.46-0.94 for cross-language comparisons as opposed to 0.29-0.74 for single-language comparisons (p. 419).

Finally, Asadi et al. (2022) explored the within- and between-speaker variability of long-term f_0 and long-term formant frequencies (F1-F4) in two speaking styles (read and spontaneous speech) of Persian and English bilinguals. Their results suggest that language is more important in speaker classification compared to style, and that f_0, F1, and F3 were better at distinguishing Persian-English bilinguals from each other than F2 and F4 for both genders.

Temporal parameters in cross-language FSC

For the purposes of FSC, Amino and Osanai (2015) compared the articulation rate of native Chinese, Korean and Thai speakers when speaking Japanese as a foreign language. The authors did not perform FSC through likelihood ratio in their research, but they revealed that the cross-language difference of AR in L1 was transferred and retained in L2.

Next, Armbrecht (2015) investigated hesitation phenomena in native Spanish and foreign English, concluding that the distribution of silent pauses remains the same across languages, while the use of filled pauses in the foreign language is more frequent for certain speakers, most probably due to lower language proficiency. The research does not focus on speaker-specificity of hesitation phenomena as the title suggests; however, the author provides the potential significance of the results for forensic speaker comparison across languages (p. 39-41).

Furthermore, Tomić (2017) explored temporal parameters of spontaneous speech (articulation rate, speaking rate, degree of hesitancy, percentage of pauses, and average pause duration) in cross-language FSC under the LR framework. The participants were ten native speakers of Serbian speaking English as a second language. The results showed that the most successful discriminant was the degree of hesitancy with error rates of 42.5%/28% (EER: 33%), followed by average pause duration (35%/45.56%, EER: 40%). As the researcher did not perform the comparison of same-language samples, it is impossible to observe how the obtained

error rates in cross-language comparison compare to the same-language counterparts. The author, however, indicates that the obtained results are in accordance with previous studies dealing with same-language FSC (p. 139).

More recently, de Boer and Heeren (2020) investigated the acoustics of filled pauses (*uh*, *um*) in 58 female speakers of L1 Dutch and L2 English. Mixed-effects models showed that, whereas duration and fundamental frequency remained similar across languages, vowel realization was language-dependent, and speakers used *um* more often in English than in Dutch. Results, furthermore, showed that filled-pause acoustics in the L1 and L2 depend on the position in the utterance, and cross-linguistic forensic speaker comparison using filled pauses may be restricted.

Consonant-based parameters in cross-language FSC

Studies exploring consonants as a parameter in cross-language forensic speaker comparison are the latest addition to the field. Cheung and Wee (2008) researched voice onset time (VOT) in 5 native Cantonese and Hong Kong English bilinguals across languages and emotional states. Their results indicate that certain speakers do retain the values across languages, but for some, the values change (p. 10). The study examines a relatively small number of speakers; therefore, it would be incorrect to draw any general conclusions.

More recently, de Boer and Heeren (2020; 2022) explored language dependency of the bilabial nasal /m/ and fricative /s/ in the spontaneous speech of about 50 L1 Dutch and L2 English speakers. The results showed that cross-linguistic differences in /m/ acoustics within the same speakers were minor, with N2 being the feature with the largest cross-linguistic difference (de Boer & Heeren, 2020). As for /s/, the results indicate that the language effect is speaker dependent; however, the spectral Centre of Gravity is, on average, higher in English than in Dutch (de Boer & Heeren, 2022). By reviewing the results of both studies, it can be concluded that retention of consonant quality across languages is more of a speaker-dependant than a general phenomenon.

Lo (2021) also measured the spectral features of /s/ (including Centre of Gravity, standard deviation, skewness, and kurtosis) with 60 French-English bilinguals and compared the values within and across languages under the likelihood ratio framework. In same-language comparisons, C_{llr} ranged between 0.41 and 0.84 and EER between 11.2% and 32.9%. In cross-language comparisons, the average C_{llr} for each of the measured parameters varied between 0.72 and 0.92 (with individual replications nearing 2), while average EER ranged between 25.8% and 34% (reaching up to 50% in individual replications) (Lo, 2021: p. 180). Lo (2021)

concludes that spectral moments of /s/ yield significantly weaker evidence under cross-language comparisons.

2.4.3. Implications for the present study

Even at a glance at the literature review in the previous section, it becomes evident that the most fruitful domain of cross-language forensic speaker comparison research includes the employment of ASR systems, yielding more robust results with every technological improvement. Bearing in mind that these systems do not rely on "higher-level" lexical features and are supposed to be characteristic of the general resonance behaviour of a speaker's vocal tract, it is reasonable that most researchers interested in this area of FSC have opted for such technology. Nonetheless, the studies have repeatedly detected the existence of the language effect, even with state-of-the-art systems that deliver relatively stable results in cross-language comparisons.

By analysing the previous research, we can infer that two significant factors interact to contribute to the so-called "language effect". Namely, several studies have noted that the more distinct phonemic systems of the compared languages, the stronger the effect. In contrast, when the languages are spoken "in a similar way" or "with a strong native accent", the effect is lower (see Bhattacharjee & Sarmah, 2012; Cho & Munro, 2017; Jovičić & Grozdić, 2014; Nagaraja & Jayanna, 2013). Furthermore, studies that reflect on the obtained results at a speaker level have detected that the language effect is not equal for all the speakers and that the system performance in language-mismatched conditions is speaker-dependent (e.g. Cheung & Wee, 2008; de Boer & Heeren; Lo, 2021). However, what the mentioned studies have in common is that they do not estimate language proficiency, fluency or the strength of the native accent of individual speakers; instead, the participants are roughly taken to be of the same level of proficiency, and fluency is not taken into consideration at all. If we knew how "far" each speaker goes when speaking the second language, that is, how much the pronunciation and phonetic realisation of phonemes deviate from the native language, we might be able to understand the scale of language effect on cross-language forensic speaker comparison. The present study aims to fill the gap in the existing literature by taking into consideration the speakers' fluency and pronunciation.

Given the vast array of features explored in single-language forensic speaker comparison, it can be said that the research in cross-language comparison has not even scratched the surface. So far, the feature extraction method based on MFCC or LPC/LPCC has generated the best results in cross-language forensic speaker comparison under well-controlled

conditions (see Durou, 1999; Faundez-Zanuy & Satué-Villar, 2006; Künzel, 2013; Lo, 2021). Parameters providing almost equally robust results are long-term formant frequencies (see Frost & Ishihara, 2015; Lo, 2021b), which is not surprising, bearing in mind that these parameters are already proven as reliable discriminants in single-language comparisons (see Asadi & Dellwo, 2019; Gold et al., 2013; Hughes et al., 2018; Lo, 2021b; Moos. 2010; Nolan & Grigoras, 2005; Tomić & French, 2019). This leads us to the question of what the next steps in cross-language FSC should be and what direction further research should take. By analogy, if we are to obtain better results in cross-language comparison, we ought to select the parameters that are considered reliable in single-language comparison. Features of voice quality chosen for the analysis in the present study have already been explored in single-language comparisons with relative success. More on the previous research concerning voice quality in FSC can be read in Section 3.3. in the following chapter.

Another vital issue to address when engaging in cross-language forensic speaker comparison under the likelihood ratio framework is the choice of the reference population. To solve this dilemma, we need to consider two perspectives, a perspective of a forensic practitioner working on real-world cases in real-world conditions and of a scientist, a statistician if we may, looking to obtain the neatest possible numbers. When dealing with mismatched conditions in the known and questioned recording, it has been suggested that the reference population should match the conditions of the known sample (Alexander & Drygajlo, 2004; González-Rodríguez et al., 2006; Morrison et al., 2012). However, when compiling the training set used for system calibration, to achieve the best results, it is recommended that the recordings be "representative of the relevant population and have the same channel and speaking-style conditions as the suspect and offender recordings, including any mismatches" (Morrison et al., 2012: p. 63). In cross-language comparisons, it implies that the training data set should be built on both L1 and L2 recordings for best results.

Considering the two factors contributing to the language effect mentioned above, such an outcome seems logical and reasonable. The previously reviewed studies, however, have reached inconclusive results concerning the combination of languages in the reference population and training data, presumably due to differences in the analysed parameters and chosen methodology. While on the one hand, it is in our best interest to calibrate the system in such a way as to accomplish the best possible performance, on the other hand, we need to consider some practical implications. Namely, bearing in mind that, in forensic reality, it is rather challenging to obtain the reference population matching the case material even in single-language comparisons – many times, the experts need to manipulate the recordings in a certain

way to match the original case files (Gold & French, 2019) – the idea of having access to the recordings of one particular group of speakers speaking the foreign language in question becomes but a myth.

Therefore, while science should strive to provide ideal results, the research needs to be mindful of real-world conditions and estimate the outcome considering the absence of such training data. Accordingly, in the present study, we will explore the influence of the reference population on the performance of the selected parameters by performing cross-language forensic speaker comparison in three conditions (1) reference population in both Serbian and English (L1 and L2), (2) reference population in Serbian (L1), and (3) reference population in English (L2),

With this, we conclude the chapter on Forensic Speaker Comparison. In this chapter, we have touched upon the development of Forensic Speech Science, examined the concept of speaker-specificity and speaker recognition by naïve listeners, explored the sources of speaker variability and introduced the Likelihood Ratio framework. The final section, concerned with the narrow field of interest for the present study, examined the previous research on cross-language forensic speaker comparison and speaker recognition by naïve listeners and discussed its implications for the present study. In the following chapter, we will survey the theoretical background and previous research relating to another significant concept for this dissertation – voice quality. We will explore the application of voice quality in forensic speaker comparison and elaborate on the selection of parameters for the acoustic analysis.

3. Voice Quality

When we hear the host of our favourite television show on the TV in another room, we will instantly recognise who is speaking even without looking at the screen. Similarly, if a comedian or a voice actor impersonates a celebrity we know, we would not need much time to grasp who they are supposed to be – this is because each person's voice has a specific "colouring", or "timbre" and we tend to associate people with their voice colour. In phonetics, the set of speaker-specific features of voice that determine its "colouring" and make it recognisable is termed voice quality. One of the earlier definitions that set the basis for the most influential voice quality theory to this day was given by Abercrombie (1967), who wrote that voice quality does not only mean "sound resulting from phonation, i.e. vibration of the vocal cords" – it refers to "those characteristics which are present more or less all the time that a person is talking: it is a quasi-permanent quality running through all the sound that issues from his mouth" (p. 91).

In the present chapter, in section 3.1. we will reflect on the voice quality theory and provide a basic anatomical overview of the vocal tract necessary to understand the laryngeal and supralaryngeal voice quality settings. Perceptual and acoustic measures will be discussed in section 3.2., with a brief overview of the instruments and technology for measuring physiological properties of voice quality. Section 3.3. focuses on the voice quality functions, covering its linguistic, paralinguistic and extra-linguistic, that is, speaker-specific aspects. Finally, in section 3.4., we will discuss some of the previous research relevant to the voice quality in forensic speaker comparison and voice quality of bilingual speakers.

3.1. Voice Quality Theory

3.1.1. Voice quality models

In the narrow sense, voice quality may refer to the vibratory patterns of the laryngeal vocal tract, coinciding with phonatory quality (see Esposito & Khan, 2020; Keating & Esposito, 2007). Laver (1980), however, defines voice quality in a broad sense as the cumulative effect of laryngeal and supralaryngeal characteristics of speech, which are "manifested as short-term articulations used by the speaker for linguistic and paralinguistic communication" but in combination create a long-term effect on perception, giving "the characteristic auditory colouring [to] an individual speaker's voice" (p. 1). The interest in voice quality research originated with linguistic motivation to characterise the segmental and suprasegmental phonetic phenomena in languages of the world. Traditionally, the vocal tract is

observed through the source-filter theory of speech production (Fant, 1960), whereby the larynx, which is perceived as the source of the sound, interacts with the cavities of the vocal tract, which act as an acoustic filter that modifies that energy, to produce different speech sounds. Ladefoged (1971) considers speech to be the product of four separate processes: the airstream process, the phonation process, the oro-nasal process, and the articulatory process.

Initiation, or airstream mechanism, denotes the source of energy for generating speech sounds, whereas phonation refers "specifically to the production of voice at the glottal opening through the larynx" (Esling, 2013: p. 110). As the airstream passes through the larynx, it is modified by the movement of the large number of muscles within and around the larynx[5], resulting in various phonation types, which can be identified by the turbulence (noise) or vibratory patterns (periodic waves) that can be heard (p. 110). The most common type of initiation is pulmonic egressive, with the energy originating in the lungs, others being glottalic and lingual (egressive and ingressive) phonation (Esling, 2013: p. 112). As initiation is not the subject of the present study, it will not be explored further[6]; it is worth noting, however, that both Serbian and English have pulmonic egressive phonation in the production of speech sounds. Laver's (1980) description of the supralaryngeal vocal tract encompasses Ladefoged's (1971) articulatory and oro-nasal (in Laver: velopharyngeal) processes.

In an attempt to build a model for phonatory contrasts in languages, Ladefoged (1971) presented the continuum of phonation types which are arbitrarily aligned along the degrees of glottal constriction, ranging between the complete closure of glottis (glottal stop) and the state of the open glottis (voicelessness). While the original continuum consisted of nine states (glottal stop, creak, creaky voice, tense voice, voice, lax voice, murmur, breathy voice, voiceless) (Ladefoged, 1971: p. 17), in recent literature, it is often presented as a range between three categories: creaky voice on one end, breathy on the other, and modal voice in between these two (see Gordon & Ladefoged, 2001), which seems appropriate considering that no languages make phonation contrasts in more than three categories (e.g. Burmese, Chong, Jalapa Mazatec), whereas most languages that have contrastive phonation have only two-way contrasts along one of the ends of the glottal stricture continuum (Ladefoged, 1971; Gordon & Ladefoged, 2001). Moreover, the phonation types according to this model should not be observed as absolute, and their realisation could vary not only between different languages but

[5] Some recent literature that describes the anatomy of the larynx includes Hewlett and Beck (2006: p. 258-264), Esling et al. (2019: p. 5-9), Hirose (1999), and Wrench and Beck (2022: p. 17-20)

[6] More information on airstream mechanism and initiation can be found in Abercrombie (1967: p. 24-33), Esling (2013: p. 110-112), Ladefoged (1971: p. 23-31) and Laver (1994: p. 161-183).

also between different sociolinguistic communities within one language as noted by Ladefoged (1971) and proven in recent research (see Keating et al., 2010). As Ladefoged (1971) himself remarks, the model is tentative – for instance, it does not account for whispery voice – however, it is influential considering that a fair portion of contemporary phonation research has been grounded in it.

Laver (1991b; 1991c) distinguishes between phonetic quality, that is, "the qualitative aspect of all learned, controllable vocal activity on the part of the speaker" (Laver, 1991c: p. 382), and voice quality – the source of the voice. Phonetic quality refers to both short- and long-term extrinsic vocal activity, including but not limited to those aspects of the sound of a voice that signal linguistic – in particular phonological – information (p. 382). Voice quality, on the other hand, has an organic (intrinsic) component, which refers to aspects of the sound that are determined by the anatomy and physiology of a speaker's vocal tract that they have no control over, such as vocal tract length or the volume of nasal or pharyngeal cavity (Laver, 1991b: p. 187). A common ground between the phonetic and voice quality is a setting component; it refers to the muscular settings that an individual adopts when speaking, which could be in service of phonetic quality to convey specific linguistic meaning or habitual, such as speaking with rounded lips, nasalisation, or a creaky voice – in both cases controllable and learnable (p. 187). Therefore, voice quality results from the organic and habitual adjustments of the vocal organs, which characterise speakers' voices on a long-term basis, beyond segmental level – a term corresponding to Nolan's (1983: p. 121) "long-term quality". Long-term tendencies in positioning the articulators in the supralaryngeal vocal tract (larynx, lips, tongue, faucal arches, pharynx, jaw and velum) are referred to as supralaryngeal or articulatory settings, whereas those referring to the laryngeal activity and the vocal cords are called phonatory settings (Laver, 1980; 1994). Regarding phonation, as opposed to Ladefoged (1971), Laver (1980) distinguishes between different types of glottal constriction and includes the dimension of overall muscular tension. He explains the whispery phonation as the airflow through the posterior glottis that can combine with any other phonation type (Laver, 1980: p. 136).

Whereas Laver (1980) does acknowledge the overlap in voice quality settings as induced by the laryngeal and supralaryngeal vocal tract, a more precise description of the interplay of the two parts of the vocal tract has been provided after technological advances and a surge of research using contemporary imaging techniques such as laryngoscopy, parallel cineradiography, ultrasound, or real time Magnetic Resonance Imaging (rtMRI). The innovations in technology and research have led to the proposal of a new model of speech production – Laryngeal Articulator Model (Esling, 2005; Esling et al., 2019), according to

which the vocal tract has two major parts: a laryngeal and an oral vocal tract. At the next level, the vocal tract contains five major articulators: larynx, velopharyngeal port, tongue, jaw, and lips, which may serve to identify the key settings in voice quality description (Esling et al., 2019: p. xvi). The theory diverges from previous linguocentric models, the primary differences being, as the name itself suggests, that the larynx is not only a source of voicing but also an articulator with "multiple sites of potential vibration" (p. xv), and the tongue is not the primary active articulator of the oral (supralaryngeal) vocal tract – its movement is instead seen as "an accompanying action to a dominantly laryngeal manoeuvre" (Esling, 2017: p. 14). For instance, the larynx incorporates the pharynx and the retraction of the tongue is considered an articulatory gesture of the larynx because, physiologically, this is where the initiation of this action occurs (Esling, 2005; Esling, 2017; Esling et al., 2019). The acoustic resonance of the vocal tract and the auditory quality are not shaped solely by the "filter" of the supralaryngeal vocal tract – instead, the articulations of the lower vocal tract interact with the vertical aspect of the laryngeal mechanism to affect both the quality of voice and individual speech sounds (Esling, 2017; Esling & Moisik, 2022: p. 252).

Voice quality – including the medium-term modifications used to express mood and emotion and the long-term variations that signal speaker identity – is usually categorised as a prosodic (Cruttenden, 2014) suprasegmental feature (Hewlett & Beck, 2006). Laver's (1980) voice quality model is strongly linked to Fant's (1960) source-filter theory of speech production; however, as recent studies have shown, the larynx is not merely a phonatory-source modulator but also a complex articulator which interacts with the supralaryngeal vocal organs (Esling, 2005; Esling et al., 2019). Furthermore, Laver's (1980) descriptions of articulatory settings through key susceptible segments are, to a great extent, Anglocentric and, therefore, difficult to adapt to different languages. Nonetheless, despite its limitations, it is undeniable that Laver's (1980) model has shaped current theoretic trends and research in the field of voice quality, both with regard to its linguistic and habitual aspects. His nomenclature of voice quality settings is in accordance with the classification of stricture points defined as places of articulation in the IPA system, sharing phonetic reference points with other sounds of the same articulatory origin. Therefore, the auditory description of voice quality is intuitive as it corresponds to the identification of other sounds (Esling & Moisik, 2022: p. 237, 241). Furthermore, the Vocal Profile Analysis protocol (Laver et al., 1981, reproduced in Laver, 1991), which is grounded in Laver's (1980) theory, with modifications depending on the field of research, remains, up to this day, one of the most nuanced tools for description of individual voice quality (Wrench & Beck, 2022: p. 242). It is for these reasons that, in the present study,

voice quality settings will be observed through Laver's (1980) framework while taking into account recent technological developments.

3.1.2. Definition of a setting

Each segment that we pronounce is characterised by a specific position of articulators in our vocal tract; for instance, /k/ in English (and Serbian) is pronounced with open, non-vibrating vocal folds, raised velum, the tip and blade of the tongue in the rest position and with the back of the tongue raised to central velum. Such an analysis of a segment is termed parametric analysis (Hewlett & Beck, 2006: p. 101; Laver, 1994: p. 115). However, if a particular position of articulators in the vocal tract is persistent throughout the speech of an individual, these long-term tendencies are abstracted from the segmental analysis and described as habitual voice quality settings (Laver, 1980: p. 2). According to this theory, a setting is not a static position but rather "a long-term-average adjustment of some part of the vocal tract, which then acts as a background for segmental articulations" (Hewlett & Beck, 2006: p. 102). This means that if a speaker has a habit of lowering their velum when speaking (resulting in more air escape through the nasal cavity, that is, nasal articulation), they will still be able to differentiate between nasal and non-nasal segments. Their long-term average position will be reflected in a tendency to make nasal segments more nasal than usual and slightly lower the velum for segments that would typically be pronounced without nasalisation.

The term "articulatory setting" was first introduced by Honikman (1964), who described vocal tract settings in different languages as learned behaviour. She considered the articulatory settings to mean "the disposition of the parts of the speech mechanism and their composite action" without including laryngeal settings. Laver (1994) defines a *setting* as "a featural property of a stretch of speech which can be as long as a whole utterance; but it can also be shorter, characterising only part of an utterance, down to a minimum stretch of anything greater than a single segment" (p. 115). According to him, the critical difference between a segment and a setting is that of span, whereby a setting is "by definition multisegmental" (p. 116). Laver (1994) remarks that a setting should be seen as "continual rather than continuous" in the sense that it could not possibly affect all of the segments, giving an example of a whispery voice that could not be observed on voiceless consonants, bearing in mind that these segments are produced without vocal fold vibration (p. 115). In addition, not all segments are necessarily equally influenced by each setting; instead, segmental susceptibility to settings should be regarded on a scale ranging from maximally susceptible to non-susceptible (Laver, 1980: p. 20-21). The segments in which a particular setting is most audible are termed key segments (Laver,

1994: p. 402). The susceptibility is primarily conditioned by the physiological relationship between the muscles and organs involved in the production of the segment and the setting, but sometimes phonological requirements override the potential susceptibility of given segments for the sake of maintaining linguistic intelligibility (Laver, 1980: p. 20-21; 1994: p. 401-402). Intermittency of settings, however, is not only a consequence of segment susceptibility to different settings; it also occurs due to the mutually exclusive nature of some settings, dynamics of speech in individual speakers or speakers' paralinguistic communicative intentions (Laver, 1980: p. 21-22).

Rose (2002) warns that it may be challenging to differentiate between phonetic features and voice quality features; for instance, retracted tongue body may be reflected in the pronunciation of specific vowels as more to the back, which could at the same time be legitimate allophones of specific phonemes. A key to understanding whether a particular feature is segmental or that of voice quality is to observe its span, that is, whether the nearby segments are affected as well (p 289). Laver (1980) proposes that a relationship between phonetic and voice quality is that of a reciprocal figure-ground, whereby what counts as one cannot be defined independently of understanding what counts as the other (p. 4-5). For instance, the pitch accent in Serbian is not grounded in absolute frequency values of rising and falling tones; instead, it can be interpreted only against the background of the overall pitch range of the speaker[7]. Correspondingly, the pitch variation due to the pitch accent should not be mistaken for individual voice dynamics. The significance of the distinction between phonetic and voice quality for forensic speaker comparison is reflected in the fact that the questioned and known speech samples may differ in four ways: (1) samples can have the same/similar voice and phonetic quality, (2) sample can have different voice and phonetic qualities, (3) same voice quality but a different phonetic quality or (4) different voice quality but same phonetic quality (Rose, 2002: p. 290), Rose (2002) also notes that in naïve speaker recognition, voice quality has more weight than phonetic quality for the listeners assessing whether two speech samples originated from the same person (p. 290).

The controllable aspect of voice quality is observed through components, yet it is crucial to understand that one speaker may exhibit several identifiable settings at the same time; thus, a voice might be described as raised larynx and nasalised, or "whispery with a backed and lowered tongue body and a rounded and protruded lip setting" (Hewlett & Beck, 2006: p. 102). The constraints on the co-occurrence of settings are imposed only by the physiology of the

[7] The example adapted from Rose (2002: p. 289).

human vocal tract (Laver, 1994: p. 153-154). Laver (1994) stresses that the notion of a setting could be applied to every level of phonetic description, including articulation, phonation, overall muscular tension factors and prosodic activities in speech (p. 153). In the present study, however, we are primarily concerned with articulation and phonation. A more detailed description of these settings will be presented later in the chapter. The conceptual framework and details for such a description of voice quality were set out by Laver (1980) and were used as the basis for the development of one of the most influential voice perception frameworks – Vocal Profile Analysis Protocol (Laver et al., 1981).

A *neutral setting* (Laver, 1980), a *neutral reference setting* (Laver, 1994) or a *neutral baseline setting* (Mackenzie Beck, 1988) is a term used to denote a baseline against which we can measure the deviation of each setting. Laver (1994) describes an ideal neutral reference setting as follows:

> "the vocal tract is as nearly as anatomy allows in a posture giving equal cross-section to the vocal tract along its full length; the tongue is in a regularly curved convex shape; the velum is in a position of closure with the back wall of the pharynx, except for phonemically nasal segments; the lower jaw is held slightly open; the lips are held slightly open, without rounding or spreading."
>
> (Laver, 1994: p. 402-403)

This description corresponds to the pronunciation of the English central vowel [ə] (Mackenzie Beck, 1988: p. 137). Laver (1994) provides additional properties to the neutral reference setting. Namely, the voice must have modal phonation, the vocal apparatus should exhibit moderate muscular tension throughout, and the pitch and loudness must be moderate in terms of mean, range and variability (p. 403).

In particular, for articulatory settings, the neutral position implies that the length of the vocal tract is not muscularly distorted – that is, the lips are not protruded, and the larynx is neither raised nor lowered. In addition, the cross-section of the vocal tract should not be distorted by the lips, jaw, tongue or pharynx and should be kept as equal as possible along its entire length. With regard to phonation, the neutral, that is, modal phonation, is achieved only with the regularly periodic (efficient) vibration of true vocal folds (not the ventricular folds), without audible roughness or friction and with the moderate muscular tension of the phonatory systems (Laver, 1980: p. 14-15; Laver, 1991b: p. 187-188; Laver, 1994: p. 404).

However, even Laver (1994) admits that "virtually nobody speaks" with a completely neutral voice considering all setting categories (p. 404). First and foremost, speakers from different accents and languages start from different phonological defaults (including the vowel space and frequency of occurrence); therefore, it would be unrealistic to expect that the centre of gravity for each language and accent would result in the position for [ə] (p. 404-405). Secondly, the nature of a speaker's articulation and phonation is often determined by

physiological differences and constraints; for instance, the asymmetry of the vocal folds would result in aperiodic vibration that would deviate from modal phonation. The neutral reference setting, as Laver (1980; 1991b; 1994) describes it, should, therefore, be regarded more as a Chomskyan (1968) competence concept, "an idealised capacity" – what we know to be the neutral position of the vocal tract, rather than something that we consistently produce in actual communication. The following sections will present the articulatory and phonatory settings, including their anatomical basis. The acoustic correlates of various settings will be reviewed in section 3.2.3. Considering that the overall muscular tension and prosodic settings are not central to the present study, they will only be briefly mentioned. More detailed information regarding these two groups of settings can be found in Laver (1980: p. 141-156; 1994: p. 416-420, 506-508).

3.1.3. Articulatory settings

The vocal tract consists of three cavities, pharyngeal, oral and nasal[8]. While the term supralaryngeal vocal tract is nowadays used to refer to the mobile speech organs in the oral cavity, above the larynx – lips, tongue, velum and the lower jaw (Esling et al., 2019; Hewlett & Beck, 2006: p. 286), earlier literature (e.g. Laver, 1980) included the larynx and pharynx as well. In the present study, the term *articulatory settings* is taken to denote the muscular adjustments of the vocal tract organs, including those of the oral (lips, tongue, velum, mandible) and laryngeal articulator (larynx and pharynx), in contrast with the *phonatory settings* which refer to the activity of the vocal folds.

By different positioning, the organs interrupt the airflow through the vocal tract, modifying its shape and dimensions, thus affecting the quality of the produced sound. In addition to the muscular walls of the pharynx that define its shape, its shape can also be modified by the root of the tongue. Furthermore, the shape of the oral cavity is altered by the front part of the tongue in conjunction with the lips and lower jaw. Finally, lowered velum allows the air to escape through an additional branch in the vocal tract – the nasal cavity (Hewlett & Beck, 2006: p. 286).

[8] For a detailed description of the anatomy of the vocal tract, the reader is advised to refer to Atkinson and McHanwell (2018), Hewlett and Beck (2006: p. 16-27, 239-255), whereas the relationship of speech production to the central nervous system is explained by Ackermann and Ziegler (2010), Smith A. (2010) and Wrench and Beck (2022: p. 12-14). The detailed descriptions of the musculature of the supralaryngeal vocal tract including the oral cavity skeletal framework are available in Hewlett and Beck (2006: p. 283-293) and Wrench and Beck (2022: p. 20-31)

Laver (1980: p. 23) differentiates between three types of articulatory settings, which deviate from the neutral reference setting by modifying the vocal tract's length (longitudinal settings), cross-section (cross-sectional settings), and position of the velum (velopharyngeal settings). For the analysis of the supralaryngeal vocal profile, Laver et al. (1981) introduce another dimension, that of articulator range, which can vary from neutral to narrow (minimised) or to wide (extensive) (Laver, 1994: p. 415-416). The term articulatory gesture is sometimes used to denote a movement of a single speech organ, or the coordinated movements of different articulators, in the production of a speech sound (Browman & Goldstein, 1992; Hewlett & Beck, 2006: p. 285). Below, we will analyse the articulatory settings through articulatory gestures of individual organs in the vocal tract. Table 3-1 is structured to summarise the articulatory settings, depicting the relationship between the articulators, vocal tract modifications and range, whereas Table 3-2 lists the key segments susceptible to the given setting in English and Serbian.

Table 3-1
Classification of articulatory settings

Articulator	***Type of Setting***			***Range***
	Longitudinal	***Cross-sectional***	***Velopharyngeal***	
Lips	protrusion	spreading	/	narrow labial range, wide labial range
	labiodentalisation, rounding			
Mandible	/	close jaw, open jaw	/	narrow mandibular range wide mandibular range
Tongue tip/blade	/	advanced, retracted	/	/
Tongue body	/	advanced, retracted, raised, lowered	/	narrow lingual range wide lingual range
Tongue root	/	advanced, retracted	/	/
Pharynx	pharyngeal constriction		/	/
Velum	/	/	nasal, denasal	/
Larynx	raised, lowered	/	/	/

Note: The table is adapted after articulatory setting description as provided in Laver (1994)

LABIAL SETTINGS Lips are a complex of muscles located immediately at the mouth opening, the most significant for labial settings being a ring muscle named orbicularis oris, which is in coordination with the mentalis muscle responsible for lip protrusion, and zygomaticus, risorius and buccinator, responsible for lip spreading. When the inner part of orbicularis oris is contracted, lips are protruded, which results in an elongated vocal tract and

reduced frequency of all of the acoustic resonances (higher formants in particular) associated with the vocal tract (Esling et al., 2019: p. 26-27; Hewlett & Beck, 2006: p. 289; Laver, 1980: p. 31-32, 40; Wrench & Beck, 2022: p. 29-30) – the setting known as *labial protrusion* (Laver, 1981: p. 31). Labial protrusion, however, seldom occurs without lip rounding, which affects both the length and the cross-section of the vocal tract; therefore, in Vocal Profile Analysis, the two settings are merged into one (Laver et al., 1981). In their description of voice quality, Esling et al. (2019: p. 26) include the settings of *open rounded* and *close rounded voice*. A gesture opposite to lip rounding would be *lip spreading*, in which case the segments that typically have a round lip position would assume the lip position of [e] or, in more extreme cases, of [i] (Esling et al., 2019: p. 26; Laver, 1980: p. 38; Laver, 1994: p. 408). The acoustic effect of lip-spreading is the rise of formant frequencies; however, one must be aware that multiple lip adjustments may occur simultaneously and, therefore, affect the resonance differently (Laver, 1980: p. 41).

As both Laver (1980) and Esling et al. (2019) recognise, lip protrusion (rounding) and spreading do not fully capture the range of movements that the lips can achieve. Laver (1980) identifies eight different settings that result from the combination of the horizontal and vertical lip parameters (p. 35-37). In addition, the upper and lower lip can be contracted independently – such is the case for the production of labiodental fricatives where the lower lip is retracted (Hewlett & Beck, 2006: p. 289). This articulatory gesture may become the articulatory setting of *labiodentalisation* (labiodentalised voice) if employed habitually by a speaker (Laver, 1980: p. 45; Laver, 1994: p. 407). Even though Laver (1994) classifies labiodentalisation as a longitudinal setting, he acknowledges that it affects both the length (shortens) and the cross-section of the vocal tract (p. 407). Labiodentalised voice is most prominent on the segments nearest to the lips, such as dental and alveolar fricatives or bilabial oral and nasal stops, then pronounced as labiodental. The acoustic correlates are reflected in lower formant frequencies, especially for higher formants, as in lip constriction due to protrusion. Moreover, the alveolar fricatives exhibit lowered fricative noise, whereas, for dental fricatives, the lower limit is raised (Laver, 1980: p. 33-34).

Labial range is considered narrow or minimized when the lips barely move from a neutral position. In contrast, if there is a substantial movement of the lips from the neutral position, with upper and lower teeth frequently visible, the speaker is considered to exhibit a wide or extensive labial range (p. 415-416). It is important to note, however, that the range of lip movement, even for the highest degrees of deviation in articulation, stays well within its maximum range (Hewlett & Beck, 2006: p. 381). Lip movement is inextricably tied to mandibular movement as well; for instance, when the lips are sealed, the lower jaw is raised,

and when the lips are protruded, the jaw is fronted (Esling et al., 2019: p. 27; Hewlett & Beck, 2006: p. 289).

MANDIBULAR SETTINGS The lower jaw, or mandible, is a horizontal U-shape attached in front of each ear with temporomandibular joints, which allow its vertical and lateral movements (Hewlett & Beck, 2006: p. 289). Since it provides attachment points for the muscles of the floor of the mouth and the tongue, both lingual and labial gestures are performed in coordination with the jaw movement (Wrench & Beck, 2022: p. 25). The muscles responsible for jaw closure are the internal pterygoid, masseter, and temporalis muscles, whereas jaw opening is aided by the external pterygoid, the geniohyoid, the anterior belly of the digastricus, and the mylohyoid (Laver 1980: p. 65–67). In neutral speech, the jaw remains slightly open, with a visible gap between the upper and lower teeth (Hewlett & Beck, 2006: p. 291; Laver, 1980: p. 65; Laver, 1994: p. 408). The gap between the lips gradually disappears in a *close jaw* setting, while in an *open jaw* setting, it becomes wider (Laver, 1980: p. 67; Laver, 1994: p. 408). Speaking through entirely clenched teeth is considered an abnormal adjustment (p. 408), even though, as Hewlett and Beck (2006) state, the speech can still be perfectly intelligible (p. 291). The openness of the jaw corresponds to the increase and range of the first formant. Higher formants also rise with the degree of openness, yet, they are less affected (Laver, 1980: p. 67). Protruded/retracted and lateral jaw adjustments are also possible; however, since they do not constitute standard settings of accent communities but rather idiosyncratic, speaker-specific adjustments (Laver, 1994; p. 409), they were not elaborately discussed by Laver (1980). The Vocal Profile Analysis protocol by Laver et al. (1981) includes the protruded jaw setting, and Esling et al. (2019: p. 25) write that this setting contrasts with labiodentalised voice (mandibular retraction). *Mandibular range* is relatively wide and, as with lips, not entirely exploited during speech (Hewlett & Beck, 2006: p. 289). A narrow or minimised mandibular range is characterised by restricted lower jaw movement; the lower jaw seldom parts from the upper jaw to reveal the tongue. A wide mandibular range implies large vertical movements of the lower jaw so that, on open vowels, it is possible to see the surface of the tip, blade and front of the tongue (Laver, 1994: p. 416).

LINGUAL SETTINGS The tongue is a muscular hydrostat – therefore, its volume remains constant even though its shape changes. The intrinsic muscles of the tongue are responsible for its shapes and movement: the longitudinal muscles thicken and shorten the tongue along the longitudinal axes, the transverse muscle can cause it to elongate and become thinner while the verticalis muscle flattens and widens it (Hewlett & Beck, 2006: p. 286-287; Wrench & Beck, 2022: p. 26). The extrinsic muscles responsible for the retraction and fronting

of the tongue include the hyoglossus and genioglossus muscle groups, while the styloglossus muscles are responsible for pulling the tongue back and up, i.e. raising (Esling et al., 2019: p. 23). The back of the tongue is attached to the body, but the tip and blade of the tongue can move regardless of how the rest of it is positioned. Accordingly, in Laver's (1980) voice quality model, lingual articulatory settings are grouped by *THE TONGUE TIP OR BLADE*, *THE TONGUE BODY*, and *THE TONGUE ROOT*.

Lingual articulatory adjustments affect the cross-section of the vocal tract. The *advanced tip/blade* setting is characterised by the tip of the tongue protruding between the teeth in pronunciation of dental segments, the passive articulator being the biting edge of the teeth instead of the inner surface. The *retracted tip/blade* setting displaces dental segments toward the alveolar ridge and alveolar segments to the post-alveolar space (Laver, 1980: p. 47; Laver, 1994: p. 410). Both Laver (1980: p. 50) and Esling et al. (2019: p. 21) describe the *retroflex* setting, where the tongue is curled backwards so that the tip articulates against or near the back of the alveolar ridge or, in more extreme cases, the underside of tip/blade of the tongue uses the hard palate as the passive articulator. The former case is acoustically reflected in F4 approaching F3, whereas the second case corresponds to the lower third formant approaching the second (p. 55). According to Laver (1994), the neutral position of the lingual body settings corresponds to that of the English vowel [ə], where "the surface of the tongue body is convex and regularly curved, with the vocal tract as nearly as anatomy allows in equal cross-section along its full length" (p. 410).

The body of the tongue can exhibit a *fronted* or *backed* setting, and a *raised* or *lowered* setting. Laver (1980: p. 45-46) proposes that radial movements of the location of the centre of mass of the tongue result in a range of secondary articulations such as palatalised voice (the tongue-body is advanced and raised), pharyngealised voice (the tongue-body is retracted and lowered), velarized voice (retracted and raised tongue-body). The shift of the tongue body in each direction is reflected in the compression of the vowel space and the displacement of articulation of the relevant consonants in the same direction (Laver, 1980: p. 47; Laver, 1994: p. 410). All things being equal, the settings that involve a fronting component exhibit a greater distance between the first two formants – the second formant is high in palatalised voice but lowers as the tongue approaches dentalisation, the third formant remaining high throughout. In contrast, the settings involving a backed tongue body should exhibit higher first and lower second formant, with the most prominent effect on front vowels (Laver, 1980: p. 55). Nonetheless, as it is challenging to discern between the nuanced locations of tongue displacement (Laver, 1980: p. 46), the standard protocols for voice perception adhere to the

four primary directions of tongue body movement. On the other hand, in line with the Laryngeal Articulator Model, Esling (2005) and Esling et al. (2019) propose fronted, raised and retracted tongue body settings while claiming that mandibular movements are responsible for tongue lowering. The proposed settings "reflect main contractile directions of the extrinsic lingual musculature" (p. 22).

Finally, Laver (1980) introduces tongue root settings, which can be *advanced* or *retracted*, resulting in expanded or constricted pharynx volume (p. 51); however, since the tongue root settings are strongly dependent on pharyngeal settings, they do not constitute part of standard perceptual protocols. Regarding *lingual range*, the tongue body is considered most responsible for this dimension, therefore, tongue tip/blade and tongue root are not associated with range settings. The lingual range is reflected in the general vowel space dimension – a narrow range setting implying that the tongue primarily remains around the centre of the vowel chart, whereas a wide range setting means that the tongue is more mobile in the mouth and moves within a more extensive area of the vowel chart, reaching towards the periphery (Laver, 1991: p. 416). Initially, the lingual range was understood under the terms lax and tense voice (Laver, 1980: p. 49).

VELOPHARYNGEAL SETTINGS The velum or soft palate is composed of connective tissue and muscles. It continues from the hard palate to the back of the pharynx, ending in the uvula and around and down at either side of the mouth. The velopharyngeal port is an opening between the nasal cavity and the pharynx that appears when the velum is lowered toward the root of the tongue (Hewlett & Beck, 2006: p. 291). The velum lowering mechanism consists of two paired muscles, palatoglossus and palatopharyngeous, which act as a pair of slings directed downwards (Laver, 1980: p. 70). The chief muscles involved in velum raising are the palatal tensor, the palatal levator, the superior pharyngeal constrictor, and some fibres of the upper part of the palatopharyngeus (p. 74). Laver (1980) notes that different speakers choose different mechanisms for velum closing and that intra-speaker variability is also observable on a day-to-day basis. The *nasal* setting can be observed on all segments except stops that originate below the velum (glottal stop) (Laver, 1994: p. 413). The speech is considered non-neutral when segments that do not have nasality as a distinctive feature are pronounced with a drop in velic height below the critical level (Laver, 1980: p. 87). Laver (1980) reports that a common acoustic correlate of nasality is a drop in the intensity of the first formant, sometimes followed by the same feature of the second formant, whereas the third formant may exhibit a lowering of both the intensity and frequency (p. 92). *Denasal* setting can be observed when the segments that in a particular language use nasality contrastively are pronounced orally, with an increased closure

in the velopharyngeal port. Such a phenomenon in speech is sometimes termed hyponasality (Laver, 1980: p. 69, 88; Hewlett & Beck, 2006: p. 292) and is exclusively related to the perception – were the listeners not aware of the nasal quality of the segments in question due to the background knowledge of the language or the nature of language in general, they would not be able to detect the denasal setting (Laver, 1994: p. 413). A velopharyngeal setting that is listed in the Vocal Profile Analysis (Laver et al., 1981) and described by Mackenzie Beck (1988) is *audible nasal escape* (also audible nasal emission, Kummer et al., 1992) – a fricative airflow through the nose that is most discernible on voiceless segments which require the maintenance of high oral air pressure, such as /s/ or /f/. It does not constitute a phonetic feature of any known accent and is considered pathological as it often characterises speakers with the cleft palate or velopharyngeal insufficiency (Kummer et al., 1992; Mackenzie Beck, 1988; Sundström & Oran, 2019). It should be noted, however, that stating that non-nasal, neutral speech has entirely raised velum without any nasal airflow would be an immense oversimplification (Laver, 1980: p. 78-80). Instead, nasality should be observed on the velic scale with a critical value for the velopharyngeal opening above which the raising of the velum results in degrees of denasal voice and below which in different degrees of nasalisation (Esling et al., 2019: p. 19-20; Laver, 1980: p. 88).

PHARYNGEAL SETTINGS The pharynx is a fibromuscular tube which forms part of the vocal tract from the oesophagus to the uvula (laryngopharynx and oropharynx) and continues through the velar port to form the posterior part of the nasal tract (nasopharynx) (Wrench & Beck, 2022: p. 21). It is encircled with U-shaped constrictor muscles that form a sphincter around it, connecting it to the root and body of the tongue. While in some earlier studies, in addition to the retraction of the body or the root of the tongue, the pharyngeal sphincteric mechanism was considered to be the main element in *pharyngeal constriction* setting (Hardcastle, 1976; Kaplan, 1960, as cited in Laver, 1980: p. 58-60), later research established that there is a strong relationship between the pharyngeal and laryngeal behaviour and that it is the aryepiglottic sphincter rather than pharyngeal constrictor muscles that induce pharyngeal constriction (Esling, 1996; Esling, 1999). The research ultimately resulted in developing the Laryngeal Articulator Model (Esling, 2005; Esling et al., 2019), a theory which implies that the manners of articulation articulated at the larynx and pharynx are inextricably linked to the mechanism for producing phonation type. Laryngoscopical research has shown that the same muscular adjustments encountered in pharyngealised voice (engagement of the aryepiglottic sphincter mechanism, retraction of the tongue root and elevation of the larynx) are also assumed

by the vocal tract in what Laver (1980) described as a raised larynx setting (Esling, 1999)[9]. In contrast, the lowered larynx setting and pharyngeal expansion have an open laryngeal vestibule, stretched aryepiglottic folds, and a lowered larynx position. Despite shared articulatory configuration, pharyngealised and raised larynx voice do not share the same perceptual correlates, the former appearing in a lower pitch and the latter in a higher (Esling, 1999; Esling & Moisik, 2022: p. 243). Fant (1957, as cited in Laver 1980: p. 62) predicts that the acoustics of pharyngeal constriction is reflected in the higher first and lower second formant, whereas pharynx expansion should be reflected in the lower F1. In addition, considering that pharyngeal constriction employs high tension, narrower formant bandwidths are expected (Laver, 1980: p. 62).

LARYNX HEIGHT SETTINGS As explained above, there is a fundamental connection between the position of the larynx, pharynx and the adjustment of aryepiglottic and glottal folds. Laver (1980) classifies the *raised* and *lowered larynx* voice settings as the longitudinal changes of the vocal tract and describes the acoustic correlates of the raised larynx voice as similar to those of the pharyngalised voice – the first formant slightly rises, whereas the second and the third exhibit lower frequency than in a neutral larynx setting (p. 27). Raised larynx voice is accompanied by a rise in the fundamental frequency and, therefore, perceived as higher in pitch. If the neutral or lower pitch is maintained while keeping the larynx raised, the obtained auditory quality is described as pharyngalised voice (Esling et al., 2019: p. 16). The muscles responsible for larynx elevation are the suprahyoid group and thyrohyoid muscles (the muscles used in swallowing). Larynx raising is not solely the product of lifting the laryngeal cartilages; it also involves engaging the aryepiglottic constrictor and retracting the tongue (Esling et al., 2019: p. 16-17). *Lowered larynx voice*, on the other hand, entails contracting the opposite set of muscles – infrahyoid and sternothyroid muscles. The thyroid cartilage is pulled towards the sternum by the sternothyroid muscles, and, for some speakers, the sternohyoid and omohyoid muscles may touch (Esling et al., 2019: p. 18; Laver, 1980: p. 29). Lowered larynx voice is accompanied by a low pitch (Laver, 1980: p. 30). The auditory quality that results if the higher pitch is maintained while the larynx is lowered is sometimes termed faucalised voice (Esling et al., 1994; Esling et al., 2019: p. 18). However, since the faucalised voice quality can also be observed as a lowered larynx falsetto, it does not constitute a distinct setting in standard voice quality assessment protocols such as Laver et al. (1981). Due to the downward expansion and

[9] The connection between the raised larynx voice and pharyngeal constriction was recognised by Laver (1980: p. 27) and acknowledged during the development of the Vocal Profile Analysis protocol in Laver et al. (1981).

increased volume of the epilaryngeal and upper pharyngeal cavity, the acoustics of lowered larynx voice results in decreased formant frequencies (Esling et al., 2019: p. 19). By analogy, the raised and lowered larynx setting are best observable on voiced segments.

Table 3-2 summarises the key segments in English and Serbian, listing the ones most susceptible to the given articulatory setting. The column relating to English was recreated according to the descriptions provided by Mackenzie Beck (1988) and Laver (1980; 1994), whereas the segments in Serbian were provided by the analogy of the manner and place of articulation as described in Subotić et al. (2012), taking into account the non-neutral segmental features available in the Articulation Test[10] (Kostić et al., 1983; Vladisavljević, 1981), considering that there is not available literature that describes the key segments according to the VPA protocol in the Serbian language.

Table 3-2
Key segment susceptibility to articulatory settings

Setting	***VoQS[11]***	***English key segments***	***Serbian key segments***	***Explanation***
Lip rounding/ protrusion	Vʷ/Vœ	/i/; [s], [z], [θ]; /r/, /ʃ/, /tʃ/, /dʒ/	/i/, /e/; [s], [z], [ts]; /tɕ/, /dʑ/, /tʃ/, /dʒ/	unrounded vowels are rounded; “pitch” of the friction sounds lower; optional rounding often present;
Lip spreading	V̟	/u/, /ɔ/, /w/; [s], [z], [θ]; /r/, /ʃ/, /tʃ/, /dʒ/	/u/; [s], [z], [ts]; /tɕ/, /dʑ/, /tʃ/, /dʒ/	rounded vowels and semi-vowels are less rounded; “pitch” of the friction sounds higher; optional rounding not present on consonants;
Labiodentalisation	Vᵛ	/p/, /b//, /m/; [s], [z]; /r/, /w/, /u/	/p/, /b/, /m/; [s], [z], [ts];	The onset and offset of bilabials; “pitch” of the friction sounds lower;
Close jaw	J̩	[aɪ], [aʊ]; front consonants	[a]; front consonants	Minimised vertical travel for diphthongs and front consonants;

[10] *Test za analitičku ocenu artikulacije srpskog jezika* – AT (Test for the Analytical Assessment of the Articulation of Segments in the Serbian language) is a protocol that lists potential non-neutral features of Serbian segments grouped by the manner of articulation (Kostić et al., 1983; Vladisavljević, 1981). The protocol involves marking a particular non-neutral feature for presence/absence and calculating the number of non-neutral segments, given their susceptibility.

[11] The transcription is taken from the revised Voice Quality Symbols chart, an extension of the IPA chart for voice quality description (Ball et al., 2016; 2018). The original Voice Quality Symbols chart was copyrighted in 1994 (Ball, 1996; Ball et al., 1995). Considering that the lingual settings do not occur solely on a single axis (for instance, they combine raising and protrusion or lowering and retraction), the transcription system does not support the componential analysis presented in the VPA protocol. For this reason, some settings in Table 3-2 are not followed with a transcription symbol. The Voice Quality Symbols for some common voice qualities that combine multiple tongue tip/blade and tongue body settings are the following: linguo-apicalised (V̻), linguo-laminalised (V̺), dentalised (V̪), alveolarised (V̠), palato-alveolarised (V̠ʲ), palatalised (Vʲ), velarised (Vˠ), uvelarised (Vʶ), pharyngealised (Vˤ); laryngo-pharyngealised (V̰ˤ) voice. Audible nasal escape does not have a transcription symbol as it is not considered a linguistic feature of any known language (Mackenzie Beck, 1988).

				compressed vowel space;
Open jaw	J̩			The diphthongs and front consonants either show extensive vertical travel, or fail to reach the usual articulatory end-point targets; overall expanded vowel space;
Protruded jaw	J̩	/s/, /ʃ/	/s/, /ʃ/	"darker" fricatives; all lingual articulations are fronted; dentolabialisation possible;
Advanced tip/blade		/θ/, /ð/; /t/, /d/, /l/, /n/ /s/, /z/	/t/, /d/, /ts/, /z/, /s/; /l/, /r/, n/	Dental segments are pronounced as interdental; alveolar segments are pronounced as denti-alveolar or dental;
Retracted tip/blade				Dental segments are pronounced as denti-alveolar; alveolar segments are pronounced as post-alveolar;
Retroflexion	V˞			The tongue tip moves toward the retroflex position; the tongue curls back;
Advanced tongue-body		/k/, /g/, /ŋ/, /l/, [ɹ], /n/, /w/, /j/, /ʃ/, /ʒ/, /ʧ/, /ʤ/, /s/, /z/; /i/, /u/, /ɑ/, /ɔ/	/k/, /g/, /x/, /l/, /r/, /n/, /j/, /ʃ/, /ʒ/, /ʧ/, /ʤ/, /tɕ/, /dʑ/, /s/, /z/, /ts/; /i/, /u/, /o/	Fronted place of articulation for the susceptible consonants; vowel space pushed toward the front of the mouth;
Retracted tongue-body				Retracted place of articulation for the susceptible consonants; vowel space pushed toward the pharynx;
Raised tongue-body		vowels and vowel-like segments	vowels and vowel-like segments	Vowel space pushed toward the palate
Lowered tongue-body				Vowel space expanded downwards
Pharyngeal constriction	Vˤ, V̴ˤ			Vowel space pushed toward the back
Nasal	Ṽ	all vowels and continuant consonants	all vowels and continuant consonants	Nasality is present on a segment even though it is not a distinctive feature
Denasal	V̰	/m/, /n/, /ŋ/	/m/, /n/, /ɲ/	Segments with nasality as a distinctive feature are pronounced orally
Audible Nasal Escape		/s/, /f/; all segments	/s/, /f/, /ts/; all segments	Audible fricative airflow from the nose, most prominent on fricatives but possible on all segments
Raised larynx	L̝	vowels	vowels	Perceived pitch is higher; vowel formants affected;
Lowered larynx	L̞			Perceived pitch is lower; vowel formants affected

Articulatory settings, as described above, do not occur independently of each other. As seen in the tongue-body movement example, some co-occurrences are frequent and commonly encountered in combination. Moreover, due to the physiological structure of the vocal tract, certain articulatory gestures involve the involuntary movement of other articulators, resulting in the articulatory settings sharing perceptual and/or physical properties. In addition, it cannot be overstated that "different speakers may achieve auditorily (and perhaps articulatorily) similar results by physiologically different means" (Laver, 1980: p 72). The human vocal apparatus is plastic and capable of performing magnificent compensatory adjustments to achieve specific articulatory goals - no rules specify which muscles must be employed in producing each specific segment (Wrench & Beck, 2022). To illustrate, McMicken et al. (2017) describe a speaker with congenital aglossia who could learn to speak without a tongue, articulating all segments quite intelligibly. The possible combination of various articulatory settings and the potential array of speaker-specific realisations of each setting makes voice quality a valuable forensic speaker comparison parameter.

3.1.4. Phonatory settings

The term "phonation type", as defined by Catford (1964: p. 27), refers to "any laryngeal activity which is not initiatory in its phonic, or sound-producing function – whatever its phonological function may be" (p. 27). whereas Abercrombie (1967) borrowed the term "register" from music studies to denote "different qualities of sound arising from differences in the action of phonation" (p. 99). Laver (1980) introduced the notion of "phonatory settings", whereas some recent studies seem to prefer the term "phonatory quality" (e.g. Esling & Moisik, 2022).

The larynx framework comprises the epiglottis, thyroid, cricoid, and arytenoid cartilages. The most important organs for phonation within the larynx are the vocal folds, which are muscular and capable of finely-tuned adjustments. Vocal folds consist of the thyroarytenoid muscle, along which connective tissue layers (vocal ligaments) are attached to the thyroid cartilage in the front and arytenoid cartilage in the back. The medial part of the thyroarytenoid muscle, also known as the vocalis muscle, contributes to controlling of the effective mass and stiffness of the vocal folds. The vocal fold length is changed by the movement of the cricothyroid joints and contraction of the cricothyroid muscle. In contrast, the movement of the arytenoid cartilage and engagement of the cricoarytenoid, thyroarytenoid and interarytenoid muscles controls the rear part of the vocal folds resulting in their *abduction* – the position when the vocal folds are pulled apart, and air passes freely through the space between them (glottis),

and *adduction* – the vocal folds are brought together preventing the airflow. In addition, some fibres of the thyroarytenoid muscles run upward into the folds which join the arytenoids with the edges of the epiglottis, i.e. the aryepiglottic folds. Ventricular or false vocal folds constitute the portion of the thyroarytenoid muscles above the vocal folds covered with mucous tissue and have a different composition than the (true) vocal folds. The opening between the vocal folds that runs along the vocal ligaments is referred to as ligamental glottis, whereas the opening along the stretch where the arytenoid cartilages are located is called cartilaginous glottis (see Hewlett & Beck, 2006: p. 260-261; Hirose, 1999: p. 138-139; Laver, 1980: p. 99-108; Wrench & Beck, 2022: p. 18-19).

When the vocal folds are placed close together, and a flow of air is pushed between them from the lungs, they rapidly perform repeated, vertical and lateral, opening and closing movements – they vibrate. A widely accepted model of vocal fold vibration is the aerodynamic myoelastic theory (van den Berg, 1958), according to which vibration is achieved as the result of two sets of opposing forces alternately gaining predominance – the myoelastic force of the vocalis muscles and the subglottal air pressure. If the muscular tension is stronger than the air pressure, the folds cannot be pulled apart; conversely, if the air pressure is greater than the muscular tension of the folds, the glottis will not close (Esling et al., 2019: p. 46; Hewlett & Beck, 2006: p. 266-267; Laver, 1980: p. 95-96; Wrench & Beck, 2022: p. 18). As recent research has shown, however, vibration is not limited to the vocal folds – it can be generated throughout the vocal tract in combination with other structures above the vocal folds (Esling & Moisik, 2022). Hirose (1999) classified the laryngeal adjustments for basic phonetic conditions into four groups: (1) abduction and adduction of the vocal folds, (2) constriction of the supraglottal structures, (3) adjustment of the length, stiffness, and thickness of the vocal folds, and (4) larynx lowering and elevation. Following Esling and Harris's (2005) views on the states of the glottis, which, according to them, involve two primary levels of laryngeal operation (glottal and arytenoid), Edmondson and Esling (2006) offered a model that could account for different phonation types based on the manipulation of six "valves" of the throat: (1) vocal fold adduction and abduction, (2) ventricular incursion, (3) aryepiglotto-epiglottal constriction (sphincteric compression of the arytenoids and aryepiglottic folds), (4) epiglotto-pharyngeal constriction (retraction of the tongue and epiglottis), (5) laryngeal raising/lowering, and (6) pharyngeal narrowing (due to the sphincteric action of the superior/middle/inferior pharyngeal constrictors) (p. 159). Engagement of one or more valves results in different phonation types, whereby abduction and adduction of the vocal folds are responsible for the distinction of various degrees of breathy phonation between respiration and modal voice; ventricular folds are

involved in the harsh phonation, supraglottal constriction with the open glottis is observed in whispered phonation and, with the closed glottis, in creaky voice. Vertical larynx adjustment is also relevant for the lowered and raised larynx setting and pharyngealisation (see Edmodson & Esling, 2006; Esling & Harris, 2005; Hirose, 1999).

The involvement of supraglottal elements in different phonatory settings was indeed also acknowledged by Laver (1980), who proposed a theory to demonstrate how several muscle groups act in coordination to produce three parameters of laryngeal control: adductive tension, medial compression, and longitudinal tension (p. 108-109). Laver (1980) defines *adductive tension* as the tension of the interarytenoid muscles, which bring the arytenoid cartilages together, closing both the ligamental and cartilaginous glottis. *Medial compression* closes the ligamental glottis by exerting pressure on the attachment points of the vocal folds and the arytenoid cartilage, utilising the lateral cricoarytenoid and the lateral parts of the thyroarytenoid muscles, whereas the *longitudinal tension* is the tension of the vocal folds achieved by contraction of the vocalis and the cricothyroid muscles. Table 3-3 summarises the phonatory settings in terms of the parameters of muscular control as described by Laver (1980).

Table 3-3
Parameters of muscular control in phonatory settings

Phonatory setting	***VoQS***	***Adductive tension***	***Medial compression***	***Longitudinal tension***
Voice	V	Moderate	Moderate	Moderate
Falsetto	F	Moderate	Moderate	High
Creak	C	High	High	Low
Whisper	W	Low	High	Moderate
Breathy voice	V̤	Low	Low	Low
Harsh voice	V!	Extreme	Extreme	High

Below, we will explore different phonatory settings and explain the configurational changes that occur due to the parameters of laryngeal control described by Edmondson and Esling (2006) and Laver (1980). Once again it should be underlined that in the present study, despite their well-established relation to phonation, larynx height settings are observed as articulatory settings of the vocal tract and will not be analysed further in this section. Laver (1980: p. 111-118) groups phonatory settings in three major categories: (1) settings that can occur on its own – simple types – and in combination with other settings – compound types – but not in combination with one another (such as modal voice and falsetto), (2) settings that can occur on its own, in combination with the settings from the first group and/or in combination

with each other (whisper and creak), and (3) modificatory settings that can occur only in compound types and never stand on their own (harshness and breathiness).

Modal phonation, modal voice, or simply "voice", is a neutral phonatory setting produced when there is a moderate degree of longitudinal tension, adductive tension, and medial compression, and the fundamental frequency is towards the lower end of the range of the speaker. Ideally, in modal voice, vocal folds are adducted along their entire length; thus, there is no fricative airflow between them – their vibration is regular and periodic (Laver, 1980: p. 111; Laver, 1994: p. 414). However, research has shown that incomplete glottal closure is not uncommon in what perceptually corresponds to modal voice (Gobl & Ní Chasaide, 2010: p. 400-401). Therefore, linguists warn that modal voice should not be understood as corresponding to "normal" voice but should instead be regarded as a "default" voice, that is, "the baseline against which to compare other types of phonation" (Hewlett & Beck, 2006: p. 274). Esling et al. (2019) demonstrate that "the state of voice has some characteristics that relate it to the first stages of engaging laryngeal constriction" (p. 44).

Falsetto is the term used to describe phonation in which the longitudinal tension of the vocal folds is much greater than in modal voice, resulting in the stretched vocal folds and low-amplitude, rapid, high-pitched vibrations (Esling, 2013: p. 116; Hewlett & Beck, 2006: p. 274; Laver, 1980: p. 118). On the other hand, the adductive tension and medial compression remain similar to that in the modal voice. This type of phonation is often described as having "pure" or "thin" quality (Hewlett & Beck, 2006: p. 274). In falsetto, the non-vibrating portion of the glottis is often left very slightly open so that there is some airflow leakage from the lungs, which Laver (1980; 1994) describes as accompanying whisperiness, while Esling et al. (2019) call it breathiness. Namely, according to Laver's (1980) theory, breathiness and falsetto cannot co-occur due to the incompatible amount of medial compression they require (p. 133). On the other hand, Esling et al. (2019) hold that such a combination of phonation types is expected, considering that both falsetto and breathiness are the functions of glottal rather than laryngeal constrictor adjustment (p. 61). The acoustic correlates of falsetto are reflected in the increased pitch and, therefore, greater distance between harmonics, as well as the steeper spectral slope compared to the modal voice (Laver, 1980: p. 119-120). The falsetto range, however, may overlap with the upper part of the modal pitch range (Hollien & Michel, 1968). An example of falsetto phonation may be observed in many of the songs by the Canadian singer Abel Makkonen Tesfaye, popularly known as "The Weekend".

As opposed to falsetto, *creak*, also called *vocal fry* or *glottal fry* (Hollien et al., 1966; Hollien & Michel, 1968) or *laryngealisation* (Ladefoged, 1971). is characterised by high

levels of adductive tension and medial compression and relatively low longitudinal tension, resulting in a thick cross-section of the vocal folds with slow vibration at a very low frequency (below 100 Hz) and usually with aperiodic pulse cycle (Hewlett & Beck, 2006: p. 274; Laver, 1980: p. 122-126; Laver, 1994: p. 194-195). A single isolated burst or set of aperiodic bursts is referred to as creak, whereas when they occur in succession, in combination with voice, they are interpreted as creaky voice. A compound creaky falsetto phonation is also possible despite their incompatible frequency adjustments because high-pitched phonation can be produced with pulses of bursts detected at intervals, in which case the auditory effect is that of creaky falsetto (Esling et al., 2019: p. 64-66). Both Laver (1980) and Esling et al. (2019) agree that ventricular folds can couple with vocal folds during creaky phonation (see Moisik et al., 2015), similarly as in harsh phonation. Nonetheless, in contrast to the creaky voice, harsh voice usually appears in the fundamental frequency above 100 Hz (Laver, 1980: p. 122). An essential aspect of creaky phonation is the engagement of the arytenoids and aryepiglottic folds, which constrict and almost entirely close the glottis beneath (Edmondson & Esling, 2006; Esling et al., 2019). In a recent paper, Klug et al. (in press, as cited in Klug, 2023), who observe voice quality space as a continuum propose a further classification of creaky voice in adjacent non-creaky voice quality space. According to the authors, creaky voice has four subcategories, whereby clean, harsh, and breathy creaky voice are produced by amplitude damping effects, whereas aperiodic creak is characterised by aperiodically spaced glottal pulses (p. 32). The authors further nuance the scale by acknowledging transitions into creaky voice (e.g., modal voice – clean creaky voice) as well as transitions between creaky voice modes (e.g., harsh-breathy creaky voice) (p. 29).

Whisper is the creation of fricative noise in the posterior glottis and does not involve vocal fold vibration. It is characterised by high medial compression and low adductive tension, which result in a narrow epilaryngeal tube between the aryepiglottic folds and the tubercle of the epiglottis. When the airflow passes through this narrow tube, it creates turbulence and noise that we perceive as whisper (Esling, 2013: p. 117-118; Hewlett & Beck, 2006: p. 277; Laver, 1994: p. 190). Whisper differs from the state of breath due to laryngeal constriction in the aryepiglottic constrictor mechanism above the glottis, which is disengaged and open for breath and compressed for whisper, usually with a raising of the larynx, retraction of the tongue and minimal vocal fold adduction (Esling, 1999; Esling, 2013: p. 118; Esling et al., 2019). Whisper phonation may be observed only on the sounds that would typically be voiced; the phonation of voiceless sounds is referred to as voicelessness (Laver, 1994: p. 191). In addition, whisper should not be considered interchangeable with the conversational act of whispering, which may

also be realised in the state of breath (Esling et al., 2019, p. 54-55). According to Laver (1980), whisper may co-occur with voice (*whispery voice*) or falsetto (*whispery falsetto*), in which case there is a greater amount of interharmonic noise. Perceptually, whispery phonation is characterised by more noise, whereas, in breathy phonation, the periodic component is dominant (p. 133-134). Another term found in linguistic literature to denote whispery voice is murmur (Ladefoged 1971: 12-14; Ladefoged & Maddieson, 1996: p. 58-62); however, Esling et al. (2019) warn that it has a broader meaning since it may refer to any of the positions along the continuum from breathy to whispery voice (p. 59).

Breath is one of the two basic states of an unstopped larynx, the other being voice (Esling, 2013: p. 114). Breath phonation is encompassed by the term voicelessness (Laver, 1994: p. 189), and, as opposed to voice, it is characterised by relatively long V-shaped glottis with abducted vocal folds so that there is slightly turbulent airflow but with a relatively lower amplitude of vibration and less fricative energy than in whispery voice since laryngeal constriction is not present (Esling, 2013: p. 114; Esling et al., 2019: p. 43; Hewlett & Beck, 2006: p. 279; Laver, 1994: p. 189). According to Laver (1980), breathiness is characterized by low adductive and longitudinal tension and low medial compression and cannot combine with falsetto owing to their incompatible longitudinal tension settings. More recent studies, however, allow for both *breathy voice* and *breathy falsetto* (see Esling et al., 2019: p. 61; Hewlett & Beck, 2006: p. 279-280). Laver (1980) remarks that there is an inverse relationship between breathiness and intensity and that it usually appears in low pitch because of the shortened length of vocal fold vibration and their relative separation and looseness (p. 133). This is in accordance with his observation that breathy voice and lowered larynx are auditorily and physiologically related (p. 31). Compared to the whispery voice, in which the epilaryngeal channel is constricted, adding friction and noise, the airflow in the breathy voice is less turbulent since the laryngeal mechanism is open and enables more linear airflow through the pharyngeal space (Esling et al., 2019: p. 56).

According to Laver (1980), *harshness* in voice is achieved through a high level of muscular tension, with extreme adductive tension, medial compression, and additional adduction of the ventricular folds (p. 130). Harshness cannot be observed in isolation – it always co-occurs with either voice (*harsh voice*) or falsetto (*harsh falsetto*), depending on the longitudinal tension. It is a result of noise and irregularity in vocal fold vibration that can be perceived in a waveform as either variation in amplitude (shimmer) or period (jitter) (Hewlett & Beck, 2006: p. 278-279; Laver, 1980: p. 127). Esling et al. (2019) describe the harsh voice as resulting from the tightening of the aryepiglottic constrictor mechanism and ventricular

adduction (p. 67), indicating the difference between harsh and creaky voice is due to the subglottal pressure. Furthermore, they challenge Laver's (1980) position that longitudinal tension is not crucial in harsh voice, indicating that harsh voice usually appears in a low pitch, mirroring the mechanism employed for creaky voice – arytenoid fronting due to the contraction of thyroarytenoid muscles and vocal fold shortening (Esling et al., 2019: p. 68). In addition, they explain the previous observations that harshness is more prominent on open vowels (Rees, 1958, as cited in Laver, 1980: p. 128) with the complementary tongue backing due to the constriction (Esling et al., 2019: p. 68). Laver (1980) describes *ventricular voice* as a "physiologically more explicit synonym for severely harsh voice" when the "ventricular folds become involved in phonation, pressing down on the upper surface of the true vocal folds" (p. 130). Esling et al. (2019), however, pose that the ventricular voice, as opposed to harsh or creaky voice, does not involve vocal-ventricular fold coupling (see Moisik & Esling, 2014) and that ventricular folds are engaged in "self-sustaining oscillation simultaneously with vocal fold vibration below" (p. 71). Another form of harsh voice often described in the literature is *pressed* or *strained voice*. When the laryngeal constriction of the supraglottic area is applied simultaneously with longitudinal tension of the vocal folds, their combined effect produces an isometric tension (shortening vs lengthening) that keeps the glottis closed unless a forceful airstream is used to generate phonation. The voice appears in high pitch, similar to falsetto, yet, unlike falsetto, the airway is closed due to the constrictor above (Esling, 2013: p. 120-121; Esling et al., 2019: p 15-16).

As illustrated above, phonation types do not always occur in isolation. In some cases, speakers can alternate between phonation types in a single stretch of speech, depending on their emotional and physical state; in others, phonation types may co-occur, yielding compound phonatory settings (Laver, 1980). A combination of phonatory settings, however, ensues with certain limitations imposed by the anatomy of the larynx. Depending on the available literature on the physiology of the vocal tract at the time and their definition of the particular setting, different authors have illustrated various constraints on the combinations of settings. For instance, Laver (1980) and Hewlett & Beck (2016) agree that falsetto and modal voice cannot co-occur because of their incompatible tension requirements and that harshness and breathiness cannot appear in isolation. On the other hand, while Laver (1980) allows for breathy voice to combine only with modal voice (*breathy voice*), Hewlett and Beck (2006) and

Esling et al. (2019) describe *breathy falsetto*[12]. In addition, Laver (1980; 1994: p. 414-415) states that creak and whisper can occur in isolation, modifying each other (*whispery creak*), or as various degrees of modification of modal voice or falsetto (*creaky voice, creaky falsetto, whispery voice, whispery falsetto*), all of which can further be modified by harshness. By introducing the Laryngeal Articulator Model, Esling (2005) offered a novel perspective on the correlation, and therefore common co-occurrence, of whisperiness, creakiness and harshness, identifying their articulatory relationship as a result of the aryepiglottic constriction mechanism. Since phonation implies voicing, the susceptible segments for the phonatory settings are all the segments that carry the phonetic feature of voicing (Laver, 1994: p. 414). Finally, it should be underlined that the separation of articulatory and phonatory settings in the present study is for the convenience of analysis, and the author is fully aware of the inextricable physiological and phonetic relationship of the larynx and the supralaryngeal vocal tract.

3.1.5. Overall muscular tension and prosodic settings

In his phonetic description of voice quality, Laver (1980) includes the settings of the overall muscular tension relating to both the laryngeal and supralaryngeal vocal tract. However, the muscular tension settings do not occur in isolation; instead, these are strongly interdependent with the articulatory and phonatory settings described above (Laver, 1994: p. 416). *Lax* articulatory setting is usually accompanied by the narrow labial, lingual and mandibular range, employing minimal muscular tension of the vocal tract with a slightly lowered larynx, without pharyngeal constriction and with moderate nasality. In contrast, *tense* articulatory setting commonly coincides with wide-range settings of the tongue, lips and jaw, employing higher levels of muscular tension, slightly raised larynx, laryngeal and pharyngeal constriction and no nasality. In addition, the tongue exhibits prominent radial movements and assumes a convex-surfaced shape in segmental articulation (Laver, 1980: p. 154-155). With regard to the laryngeal tension, as seen in Table 3-3 above, the highest levels of muscular tension result in what is described as harsh phonation, whereas on the opposite end of the spectrum is breathiness (Laver, 1980: p. 144-146). According to Laver (1980), whispery phonation may be on a scale toward the lax larynx setting, but tense whispery voice is also possible (p. 146). The tense larynx setting includes the "anterior voice" of "tight" or "hard" quality, also described as a "metallic" voice (Catford, 1977, as cited in Laver, 1980: p. 145-146), which implies that there is a degree of laryngeal constriction; the arytenoid cartilages are

[12] In the present study, the expert listeners were not specifically instructed how to observe breathy and whispery voice.

clamped together, and only the ligamental glottis participates actively in phonation. It should be noted, however, that according to Esling et al.'s (2019) account of the physiological adjustments for the whispery voice, laryngeal tension is always present in whispery voice; hence, it is more characteristic of a tense than a lax setting.

In the perceptual protocol for annotating an individual speaker's profile, Laver et al. (1981) extend the notion of settings to include the prosodic features: pitch, consistency and loudness, whereby the first and the third category are observed through mean, range and variability parameters. However, similarly to the muscular tension settings, prosodic settings of pitch and loudness do not exist separately from the rest of the vocal apparatus; what is more, they are mere perceptual correlates of the vibration rate and amplitude achieved through different muscular adjustments of the vocal folds and supralaryngeal vocal tract. For this reason, neither will be included in the perceptual analysis of voice in the present study. The notion of consistency is related to the coordination of respiratory and phonatory processes, which can result in auditory perception of tremor if broken. Mackenzie Beck (1988) defines *tremor* as "the occurrence of audible fluctuations in pitch and/or loudness, which typically occur at a rate of 1-3 per syllable" (p. 175). Vocal tremor is considered a pathological condition, and it can occur on its own or in combination with other neurological conditions such as laryngeal dystonia, where the entire epilaryngeal tube can be involved in uncoordinated contractions, or Parkinson's disease, where it is associated with asymmetric muscle contractions (Esling et al., 2019: p. 232-234). Vocal tremor will be included in the perceptual analysis in the present study as a parameter of laryngeal irregularity.

3.2. Measures of Voice Quality

Articulatory and phonatory settings of voice quality can be assessed physiologically, acoustically, or perceptually. As Mackenzie Beck (1988) noticed more than three decades ago, all of these aspects have advantages and disadvantages, and "they should be seen as complementary rather than competing strands of voice quality research" (p. 128). Namely, technological innovations and advancement have been crucial in understanding the physiological aspects of speech production, voice quality included. However, the importance of the oldest instrument for auditory assessment – the human ear, should not be underestimated, even in the era of biomechanical modelling and 3D printing. In the fields such as forensic speech science, where an expert has only a voice recording at their disposal, research based on acoustic and auditory analysis plays a central role in detecting and describing the voice-quality features in a sample. In the following sections, we will briefly explore some of the available

techniques employed in voice quality research, paying particular attention to the acoustic and auditory analysis chosen as instruments in the present study.

3.2.1. Instruments for measuring articulation and phonation

Techniques for measuring the physiology of articulation and phonation that have been employed in speech production research can be grouped into three main categories: (1) direct-imaging techniques such as video recording, ultrasound, radiography, laryngoscopy, magnetic resonance imaging (MRI), (2) point-tracking including x-ray microbeam, electromagnetic articulography (EMA), velotrace or optical tracking, and (3) techniques giving indirect evidence about speech production based on aerodynamics, electropalatography (EPG), electroglottography (EGG) or static palatography (see Esling & Moisik, 2022: p. 244-247; Hewlett & Beck, 2006: p. 270-271, 293-302; Hirose, 1999; Lin, 2022; Stone, 2010; Wrench & Beck, 2022: p. 34-35). In addition, numerous computational biomechanical simulation models of the larynx or the entire vocal tract have been devised to test the hypotheses regarding neural and muscle engagement in speech production.

A video camera is one of the widely available and oldest tools for visual inspection of the articulators such as the tongue, lips and jaw. As a case in point, Honikman's (1964) descriptions of language-specific settings based on films are still commonly quoted in contemporary research. Next, ultrasound has been used to image the long-term adjustments of the tongue and larynx, such as larynx height (e.g. Moisik et al., 2014) or the accompanying movement of the aryepiglottic structures during the tongue retraction for pharyngeal/epiglottal articulations (e.g. Meluzzi et al., 2017, as cited in Esling & Moisik, 2022). Videofluoroscopy or cineradiography, using the modified barium swallow, has been exploited in swallow and speech research for the detection of various disorders; however, it also found application in voice quality research, such as the assessment of anatomical and functional voice quality correlates in patients after laryngectomy (e.g. van As-Brooks et al., 2005) or for studying pharyngealised speech sounds (e.g. Esling et al., 2019). Furthermore, laryngoscopic examinations, which are performed by inserting an endoscope through the oral or nasal cavity, have been essential in understanding the relationship between the larynx and supralaryngeal vocal tract in the production of different phonation types (see Esling, 1999; Esling et al., 2019; Esling & Clayards, 1999), states of the glottis (Esling & Harris, 2005) or for studying linguistic phonation contrasts in different languages (e.g. Brunelle et al., 2010; Edmondson et al., 2001). Magnetic Resonance Imaging (MRI) is a non-invasive technique that can capture the entire vocal tract; it uses a magnetic field and radio waves to image a section of soft tissue while bones

and air spaces are displayed as a uniform black. MRI has been employed in studying the physiology of phonation (e.g. Esling et al., 2019; Moisik et al., 2015), articulatory settings (e.g. Ramanarayanan et al., 2013) and for studying voice in clinical research (e.g. Derlatka-Kochel et al., 2021; Gracco et al., 1994; Schlamann et al., 2009).

Point tracking comprises a set of tools that track specific points along a speaker's vocal tract. A technology popular throughout the eighties and in the early nineties was the X-ray microbeam – a set of metal pellets attached to the accessible active articulators whose movements are tracked by narrow-targeted X-rays. A 50-speaker x-ray microbeam database (Westbury et al., 1990) is still available for speech research today. Next, velotrace (Horiguchi & Bell-Berti, 1987), also an older technology, observes the vertical motion of the velum. Electromagnetic articulography is a more recent technology that involves glueing sensors to the mobile articulators as well as the forehead and teeth for immobile reference points and tracking their motion (see Hewlett & Beck, 2006; Lin, 2022). Similarly, optical tracking can be used for studying lip and jaw movement. It involves attaching markers to the jaw and lips and using a camera to track and measure their displacement in three dimensions during speech. Point-tracking methods, however, have mainly been applied in phonetic studies concerning segmental articulation, coarticulation or speech development and seldom in voice quality research.

The indirect methods are termed so because they help quantify "the consequences of the actions of articulators" and do not observe or track the movement of the articulators themselves (Lin, 2022: p. 375). Electroglottography (EGG), also known as electrolaryngography, is used for studying glottal movement by placing two surface electrodes on either side of the larynx and releasing a small current between them. The obtained waveform correlates with states of the glottis, where peaks represent maximum closure, and troughs represent the open phase (Titze, 1990). The technology has been employed to study phonation in speech (e.g. Borsky et al., 2017; Burin, 2018a; Heinrich et al., 2004) and singing (e.g. Dong & Kong, 2021; Selamtzis, 2018), as well as paralinguistic (e.g. Bone et al., 2010; Burin, 2018b; Leykum, 2021) and linguistic (e.g. Esposito, 2005; 2012; Khan, 2010; Kuang, 2010) voice quality. In addition, the multichannel electroglottograph (Rothenberg, 1992) has been used for assessing the changes in the vertical position of the larynx and its acoustic correlates (e.g. Pabst & Sundberg, 1993). An extensive contemporary review of electroglottographic research in various fields is provided by Herbst (2020). Static palatography and electropalatography (EPG) (Hardcastle & Gibbon, 1997) can provide information concerning contact between the tongue and the hard palate. Static palatography is rather impractical for studying long-term adjustments because it requires painting the participants' tongues with an emulsion of charcoal in oil and

photographing transfer markings on the palate. On the other hand, EPG requires that the participant wears a custom-made artificial palate with embedded electrodes that enables the patterns of contact to be tracked on a computer. Considering the cost of the artificial palates and the fact that each palate can be worn only by the person for which it was moulded, it is not surprising that EPG mostly found its application in clinical research and descriptive projects focusing on the articulation of sounds of specific (endangered) languages (Lin, 2022: p. 377).

Finally, biomechanical modelling is often used to test the hypotheses of how combinations of muscle contractions produce various articulatory shapes, speech sounds and voice qualities. Numerous 3D vocal tract models have been developed either for didactic or research purposes. An extensive review of the available models is provided by Calvache et al. (in press), who grouped them into four categories: those representing the "source" (vocal folds), "filter" (vocal tract), source-filter and airflow-source interaction models.

In conclusion, because of the costliness of the equipment, the specialised training the researchers need to have in order to be able to use it, and the invasiveness of some of the methods, the research involving the technology described above primarily includes case studies and is seldom performed on large groups of speakers.

3.2.2. Perceptual analysis of voice quality

As seen so far, the concept of voice is inextricably linked to perception. Thirty years ago, Kreiman et al. (1993, as cited in Mackenzie Beck, 2005: p. 292) identified 57 different voice perception schemes used in the United States, and there have been many more since then. Perceptual analysis of voice has had the most prominent application in speech pathology, detection of speech impairments and various neurological conditions. As a result, perceptual frameworks have mainly been developed for these purposes. The selection of an appropriate voice perception method should depend on the expert's goal – for instance, detecting a degree of abnormality in voice, measuring the response to therapy or voice comparison (Mackenzie Beck, 2005). In San Segundo's (2021) survey on the application of voice quality in clinical and forensic practice, most phoneticians declared that they rely on the Vocal Profile Analysis Scheme – VPAS (Laver et al., 1981). In contrast, most clinicians opted for some version of GRBAS (Hirano, 1981). Other responses included The Stockholm Voice Evaluation Approach – SVEA (Hammarberg, 2000), Consensus Auditory Perceptual Evaluation – CAPE-V (Kempster et al., 2009) and two in-house solutions, whereas no one selected Buffalo III Voice Profile (Wilson, 1987).

The GRBAS scale evaluates the pathological deviation of 5 phonatory voice components (Grade, Roughness, Breathiness, Asthenia, Strain) on a 4-point scale. Versions of this scale may include additional parameters such as instability (see Dejonckere et al., 1996). In addition to the description of phonation, protocols such as SVEA, CAPE-V and Buffalo Voice Profile include prosodic features such as loudness and pitch, and in the case of the BVP, tempo and nasal resonance (see Hammarberg, 2000; Kempster et al., 2009; Wilson, 1987). All of the mentioned protocols have been developed primarily for characterising voice disorders and assessing pathological voices, which renders them a less sensitive instrument for describing healthy voices as most features would seldom deviate from "normal".

In the present study, Vocal Profile Analysis developed following Laver's (1980) phonetic description of voice quality is selected as an instrument due to its applicability in describing the voice quality of any speaker with normal anatomy through a large number of components or settings. The original VPA consisted of vocal quality (supralaryngeal and laryngeal), prosodic and temporal organisation features and comments regarding breath support, rhythmicality and diplophonia. The subsequent versions have been modified several times (e.g. Laver, 1994: p. 154; Laver, 2000: p. 44-45; San Segundo & Mompeán, 2017), partly due to theoretical considerations and in part due to the growing experience in the protocol's usage (Mackenzie Beck, 2005).

In VPA, each of the groups of settings is observed through a scalar degree that represents a deviation from the neutral reference setting, Laver et al. (1981) introduced a 6-point scale for all of the settings, except for modal voice and falsetto, which are marked either for their presence or absence and temporal organisation features (continuity and rate), which are marked for their inadequacy on a 3-point scale. In addition to the scalar degree explanations for each setting, Mackenzie Beck (1988) also provides a general description of scalar degrees:

- Scalar degree 1 is used when the presence of a setting is just noticeable.
- Scalar degree 2 suggests that the judge is fairly confident about the presence of a setting, but that there is only moderate deviation from neutral.
- Scalar degree 3 can be taken as the strongest degree of a setting which could reasonably be expected to act as a regional or sociolinguistic marker for a hypothetical community, although there are exceptions to this rule.
- Scalar degree 4 indicates that there is no doubt at all about the presence of a setting, and that it is beyond the limits of widespread use amongst accents marking membership of a sociolinguistic community.
- Scalar degree 5 represents almost the maximum strength of deviation of which the normal vocal apparatus is capable.
- Scalar degree 6 is reserved for the auditory effect which corresponds to the most extreme adjustment of which the normal, non-pathological vocal apparatus is capable.

Mackenzie Beck (1988: p. 149)

The boundary between a scalar degree 3 and scalar degree 4 is considered to be the boundary between normal and abnormal adjustments (Laver et al., 1981; Mackenzie Beck, 1988), but not in the sense that scalar degrees 4 to 6 are reserved for pathological voices – instead, the labels refer to statistical norms (Mackenzie Beck, 1988: p. 147-149). Namely, if the prominence of a setting is higher than would be expected in any linguistic community and is rarely encountered in the general population, it should likely be graded with a scalar degree higher than 3 (p. 147). For the intermittent appearance of a setting (on more than 10% but less than 90% of the susceptible segments), the experts are advised to add (i) (Mackenzie Beck, 1988: p. 150). While Laver et al. (1981) and Mackenzie Beck (1988) maintain that the 3/4 boundary is crucial in voice perception, Laver (1994) suggests using a 3-point scale for the description of normal voices and extending the scale for application in speech pathology, where settings could take more extreme deviation from the neutral reference (p. 153, 400). In order for the protocol to be objective and the results obtained with this analysis to be comparable, Laver (1980) warns that "any judgement of scalar degrees has to be made on absolute grounds, not grounds relative to the accent of the speaker's speech community, nor any other relative measure which is not general to all anatomically and physiologically normal human beings" (p. 88).

The initial protocol was designed to be completed in two passes; the first involves marking a setting for its (non)neutrality and the second assigning scalar degrees to the non-neutral settings. Since settings differ in the amount of perceptual evidence needed to identify them correctly, whereby phonatory settings usually require fewer syllables than articulatory settings (Laver 1994: p. 400), experts need to consider the length of the samples that are going to be analysed. For example, the minimal amount of connected speech for establishing a speaker's vocal profile is 40 seconds (Hiller et al., 1984, as cited in Mackenzie Beck, 2005: p. 301). Furthermore, an important issue to consider is the training of the judges that perform the perceptual analysis (Shewell, 1998). Laver et al. (1981) and Mackenzie Beck (1988) describe the training and assessment procedure that has yielded successful results. Due to the specialised training the VPA protocol requires, even 40 years after its design, there are very few experts (compared to the overall number of phoneticians and speech pathologists) who are able to use it.

A possible quantification of settings is to observe the mutually exclusive settings (such as lip rounding and spreading or tongue fronting and backing) as part of the same continuum, whereby, on a 6-point scale, tongue fronting would exhibit 13 values (Mackenzie Beck, 1988: p. 179-180). Correspondingly, on a 3-point scale, there would be seven possible

values. The reliability of the voice analysis with VPA protocol is typically assessed through the measures of inter-rater and intra-rater agreement by calculating the percentage of agreement or with statistical procedures such as $\chi 2$ test (e.g. Mackenzie Beck, 1988) or Cohen's, Fleiss' or Linear weighted kappa (e.g. San Segundo et al., 2019; San Segundo & Mompeán, 2017). The percentage agreement between the judges can be absolute – the raters have assigned the same scalar degree for the particular setting, within one scalar degree and within two scalar degrees, provided that the values are on the same side relative to the neutral value (Mackenzie Beck, 1988: p. 180). Therefore, the final and most lenient criterion may not be appropriate for a 3-point scale.

Vocal Profile Analysis protocol has successfully been applied in speech pathology research and practice to characterise the vocal features associated with specific disorders and to assess the effectiveness of therapy (e.g. Carmago & Canton, 2019; Clary et al., 1996; Fraser et al., 1998; de Lima Silva et al., 2017; Mackenzie Beck, 1988; 2010; San Segundo & Delgado Hernández, 2021; Pessoa et al., 2012; 2014; Shewell, 1998; Webb et al., 2004; Wirz, 1991); in sociolinguistic studies to evaluate the accent characteristics of a specific language community (e.g. Esling 1978, 2000; Sóskuthy & Stewart-Smith, 2020; Stuart-Smith, 1999); personal identity and gender (e.g. Camargo et al., 2012; Mackenzie Beck & Schaeffler, 2015); mother-child interaction (e.g. Marwick et al., 1984); as well as in forensic speech science to characterise speaker-specific aspects of voice (e.g. French et al., 2015; San Segundo et al., 2019). A more detailed discussion on forensic usage will ensue in Chapter 3.4.1.

3.2.3. Acoustic analysis of voice quality

It has already been established in Chapter 3.1.3. above that the changes in the articulatory adjustments of the vocal tract affect resonant frequencies. Nolan (1983: p. 162) investigated the formant values of the articulatory settings in Laver's (1980) recordings, grouping them into three categories according to F2 values:

(1) settings with a raised F2 (palatalised, velarised, palato-alveolarised, alveolarised and dentalised)
(2) central settings (modal, lip spreading, neutral, uvularised and open rounding)
(3) settings with a lowered F2 (retroflex, close rounding, lowered larynx, raised larynx, pharyngalised and laryngo-pharyngalised).

Whereas the settings in the first and the second group have similar F1 values, the third group exhibits substantial variation with lowered larynx voice and close rounding having the lowest first formant (Nolan, 1983: p. 162-164). In addition, the settings of the second group

can be differentiated on the basis of F3 since this formant reflects lip rounding. Nolan's (1983) results concerning raised larynx voice (F1 similar to neutral, F2 lowered) contradict previous studies which predicted raised larynx voice to maintain higher formant values, opposite of lowered larynx voice (Sundberg & Nordström, 1976, as cited in Nolan, 1983: p. 164). However, recent research combining laryngeal ultrasound recording and acoustic measures has confirmed that raised larynx voice is associated with lowered F2 and F3 (Moisik, 2013: p. 293-308). Such results corroborate the hypothesis that this setting is accompanied by aryepiglottic constriction of the epilaryngeal tube and tongue retraction (Esling et al., 2019: p. 17).

Regarding the acoustic analysis of phonation, it is evident that the sound produced by the vocal folds cannot be isolated from the modifications imposed by the rest of the vocal tract. However, by applying the inverse filtering to the speech signal, it is possible to obtain the glottal waveform, and creating a spectrum from this waveform (i.e. cepstrum) provides insight into the spectral composition of the sound that originated in the vocal folds (Hewlett & Beck, 2006: p. 271; Gobl & Ní Chasaide, 2010: p. 382). The parameters derived from the glottal waveform include fundamental frequency, amplitude, jitter – perturbation of the fundamental frequency, and shimmer – perturbation of the amplitude (Hewlett & Beck, 2006: p. 272). Fundamental frequency (f_0) or pitch largely depends on the degree of tension, the mass, and the length of the vibrating vocal folds, and it directly depends on the rate of vocal fold vibration. Since the mass of the vocal folds cannot be voluntarily changed in the given speaker, the variation in pitch is achieved through the manipulation of the vocal fold tension (Hewlett & Beck, 2006: p. 269). Amplitude correlates with the perceptual impression of loudness and is determined by the amplitude of glottal vibration, whereas perturbations of f_0 and amplitude are conditioned by the irregularity in the vibration of vocal folds (p. 272). Aperiodic frequency and variability in amplitude are associated with harshness (Esling et al., 2019: p. 15), whereas falsetto is characterised by high fundamental frequency and steep spectral slope (p. 62). These measures can also be obtained from the speech waveform; however, as it is, to some extent, influenced by source-filter interaction effects, there may be some errors (Gobl & Ní Chasaide, 2010: p. 392).

Spectral measurements based on the speech output signal are also often employed as acoustic parameters of voice quality in linguistic research. The long-term average spectrum (LTAS) provides information on the spectral distribution (including the energy peaks) of the speech signal over a period of time (Nolan, 1983: p. 142-155; Esling & Moisik, 2022: p. 242). The level of tension in the vocal tract can be determined by the spectral tilt, which represents the balance between high- and low-frequency energy (Hewlett & Beck, 2006: p, 273). For

instance, the comparison of the amplitude level of the first harmonic with the level of the second harmonic (H1-H2) is a spectral balance measure used to determine how evenly the sound is distributed across the frequency range and is said to correspond to Open Quotient – the period of the openness of the glottis in proportion to the glottal cycle (Gobl & Ní Chasaide, 2010: p. 394; Hanson, 1997), but a low-frequency measure (H2-H4) is also used (Esposito & Khan, 2020: p. 7; Kreiman et al., 2007: p. 596). Furthermore, the measures of amplitude between the first harmonic (H1) and the amplitudes of formants (e.g. H1-A1, H1-A2, H1-A3) are reported to be an accurate indication of source spectral tilt and to correlate with the abruptness of vocal fold closure (Esposito & Khan, 2020: p. 7; Hanson, 1997: p. 469; Keating & Esposito, 2007: p. 86). In addition, the amplitude of F1 relative to that of the first harmonic (H1-A1) may be an indication of bandwidth, that is, provide indirect information regarding the "degree to which the glottis fails to close completely during a cycle of glottal vibration" (Hanson, 1997: p. 470-471). Kreiman et al. (2014) also include the measures of spectral slope from the fourth harmonic to the harmonic nearest 2 kHz in frequency (H4-2 kHz) and from the harmonic nearest to 2 kHz to the one nearest 5 kHz in frequency (2 kHz-5 kHz). In all of these parameters, a higher value is associated with breathier phonation, whereas a lower value indicates a creakier one. This was experimentally tested by Klug et al. (2019), who explored the acoustic correlates of breathy voice quality. They found correlations between the perception of breathiness and the intensity difference between the lowest two harmonics (H1*-H2*), the intensity difference between the lowest harmonic and the harmonic closest to the first formant (H1*-A1*).

Gobl and Ní Chasaide (2010) remind us that spectral measures are susceptible to various factors and cannot be applied in all circumstances. For instance, since the frequencies of the formants affect their amplitude levels, a comparison of H1 and F1 levels across different vowel qualities would not be appropriate. Moreover, when F1 is low and/or f_0 is high, the levels of H1 or H2 may be amplified depending on their proximity to the F1 peak, in which case these two harmonics are influenced by both the source and the filter and, therefore, could not be considered to indicate the mode of phonation reliably (p. 394). A few solutions have been proposed to neutralise the differences in vowel quality and cancel the effects of the vocal tract filter, in which case the corrected measures (H1*, H2*, A1*, A2*, A3*) are used for further analysis (see Iseli et al., 2007; Hanson, 1997).

Another phonation index is interharmonic noise, which can be obtained by isolating the periodic component from the noisy component of the speech waveform. The separation can be done through spectral- or cepstral-based analysis (Hanson, 1997). Harmonic-to-Noise Ratio (HNR) is the quantification of noise in the signal which reflects the airflow friction at the glottis

or elsewhere in the vocal tract and is reported to correlate with perceived hoarseness (Yumoto et al., 1984). An additional measure of periodicity applied in voice quality research is cepstral peak prominence (CPP) – the amplitude difference between a peak in the cepstral power spectrum and the value of a trend line at the same quefrency. Correspondingly to HNR, a highly periodic signal that can be observed in modal voice exhibits a well-defined harmonic structure and a more prominent cepstral peak than a less periodic signal such as whispery or creaky voice (Esposito & Khan, 2020: p. 7; Hillenbrand et al., 1994: p. 772). Klug et al. (2019) found that CPP is in correlation with perceived breathy voice.

Acoustic correlates of phonation have been explored by many researchers in linguistics and speech pathology (e.g. Cleveland, 1991; Hillenbrand & Houde, 1996; Hillenbrand et al., 1994; Jerotijević Tišma, 2020b; Keating et al., 2015; Klug et al., 2019; Kreiman et al., 2012). The latest developments in voice quality research have focused on implementing neural networks in the automatic classification of phonation types or voice disorder detection (e.g. Bilibajkić et al., 2014; Chanclu et al., 2021; Furundžić, 2018; Han et al., in press; Xie et al., 2016). Studies focusing on the acoustic analysis of voice quality in forensic speaker comparison will be explored in greater detail in Chapter 3.4.2.

3.3. Functions of Voice Quality

In order for a parameter to be used in forensic speaker comparison, it is of utmost importance to understand how it is manifested in the relevant languages and be aware of its sources of variability. With this in mind, in the present section, we describe voice quality in terms of its communicative and informative function, whereby the communicative function primarily focuses on linguistic and paralinguistic use of voice quality settings across languages. Furthermore, the sub-section on the informative function of voice quality encompasses the extralinguistic elements related to the organic and sociolinguistic variation of voice quality. Finally, we reflect on the significance of communicative and informative functions of voice quality for forensic speaker comparison and reflect on the implications for the present research.

3.3.1. Communicative function

Linguistic function

Both articulatory and phonatory adjustments of the vocal tract can be part of the phonetic quality of segments. Moreover, Laver (1980) did name the articulatory settings of voice quality after the categories recognised as place and manner of articulation of consonants. However, since, in Laver's terms, the settings are "by definition non-segmental" (Laver, 1994:

p. 184), he observes the linguistic function of voice quality on units larger than segments, giving the example of Sundanese (a language of Java), where nasality is used to mark verb forms (p. 185). In Serbian and English, neither articulatory nor phonatory (long-term) settings have a contrastive linguistic function (apart from the pitch adjustments to signal intonation and tone in the case of Serbian pitch accents); however, the phenomenon is observed in many languages of the world. The linguists who define voice quality in the narrow sense – as the result of phonatory effort (see Esposito & Khan, 2020; Keating & Esposito, 2007), on the other hand, often explore phonatory contrasts on a segmental level. While not essential for the present study, an overview of such linguistic use of phonation could be invaluable to the linguists exploring cross-language forensic speaker comparison in languages other than English and Serbian; therefore, below, we will provide a short overview of such contrasts found on consonants, vowels and syllables alike.

Vowels can carry phonatory and pitch information, yielding languages with the tone, register, or tonal register contrasts (Esling, 2013: p. 122). For instance, some Nilotic languages (Ateso and Lango) differentiate between creaky and modal vowels (Ladefoged, 1971: p. 15), whereas some Indo-European languages (Hindi, Sindhi, Marathi, Bengali, Assamese, Gujarati, and Bihari) contrastively use modal and whispery phonation, or what Ladefoged (1971: p. 12-14) referred to as murmur (Laver, 1994: p. 200). Voiceless or whispered pronunciations of vowels are also encountered in some North American languages of the Amerindian (Comanche and Cheyenne) and Nootka (Ditidaht) language family (Canonge, 1957; Pike, 1963, as cited in Laver, 1994: p. 189). Furthermore, in !Xóõ, a Khoisan language spoken by Bushmen in southern Africa, modal and whispery vowels can occur with additional creakiness, to give creaky voice and whispery creaky voice, or with an additional strident quality, which involves the narrowing of the aryepiglottic folds, pharyngeal constriction and epiglottis backing (Garellek, 2019; Traill, 1985, as cited in Gobl & Ní Chasaide, 2010: p. 404). Phonatory contrasts are seldom realised on an entire vowel; instead, they usually occur on a portion (Gobl & Ní Chasaide, 2010: p. 404).

Contrastive phonatory types can also be observed in consonants. Esling (2013: p. 122) notes that a modal-voiced stop can contrast with a breathy-voiced stop (or what Laver, 1980 would describe as whispery) and with a creaky-voiced stop or with a stop with other phonatory qualities. For example, some languages of West Africa (Hausa, Bura and Margi) employ either modal or creaky phonation (Ladefoged, 1971: p. 15), whereas Niger-Congo (Shona and Tsonga) and Benue-Congo (Ndebele and Zulu) languages have whispery/murmured and modal phonation contrasts in consonants (Ladefoged (1971: p. 12-14). Many Indo-Aryan languages (e.g. Nepali, Gujerati, Hindi) contrast whispery, modal, voiceless unaspirated, and

voiceless aspirated stop (Dixit, 1989, as cited in Gobl and Ní Chasaide, 2010: p. 404). Phonatory contrasts have also been reported on nasals, liquids, and approximants (Ladefoged, 1971: p. 14-15; Gobl & Ní Chasaide, 2010: p. 404). It is important to note, however, that while the contrasts mentioned above may be observed on the consonant itself (as is the case with sonorants), they are primarily realised on the onset or offset of neighbouring vowels (Gobl and Ní Chasaide, 2010: p. 404).

At a syllabic level, breathy, whispery, creaky and different kinds of harsh-voiced syllables can combine with a range of pitch targets to produce phonological systems with tone along one dimension and phonatory register along another (Esling, 2013: p. 122). For instance, in some varieties of Vietnamese, a phonation type of creak co-occurs with a high-rising lexical tone to distinguish it from a regular high-rise without a creak (Cruttenden, 2014: p. 302). Phonatory quality is also said to distinguish the seven tones of Green Mong, a Hmong dialect spoken in Southeast Asia (Andruski & Ratliff, 2000). Similarly, previous research has shown that creaky voice accompanies some falling Mandarin tones (Belotel-Grenié & Grenié, 2004). Syllable categories can also be contrasted with pharyngeal resonance and oral/nasal qualities (Esling & Edmondson, 2011, as cited in Esling, 2013: p. 122). Kirk et al. (1984, as cited in Gobl and Ní Chasaide, 2010: p. 404) suggest that voice quality contrasts in the Jalapa de Diaz dialect of Mazatec, Mexico, can be observed to occur on a syllabic level, in which case the language is described to have modal voiced, breathy (whispery) voiced, and creaky voiced syllables, or it can be described to have a contrast between modal and breathy (whispery) voiced vowels, and of modal and creaky voiced consonants.

Apart from contrastive voice quality observed on segments and tone in some languages, different phonatory settings often accompany specific phonetic changes that are not considered phonologically distinctive. Such phonatory adjustments can be perceived in the phenomena such as assimilation, coarticulation or segment harmony (see Laver, 1994: p. 394-397). In addition, creaky or breathy (whispery) phonation is encountered in many languages to signal phrase boundaries (Gobl and Ní Chasaide, 2010: p. 407-409). Nonetheless, contrastive phonatory settings are realised neither in the same manner nor to the same degree across languages. Keating et al. (2010) have compared the contrastive phonation types across four languages (Gujarati, Hmong, Mazatec and Yi) to find that each category (breathy, creaky, modal, lax/tense) differs from language to language on multiple acoustic measures, concluding that language/speaker differences in voice quality are more significant than phonation category differences. An extensive overview of contrastive and allophonic phonation types across

languages of the world can be found in Ladefoged (1971), Ladefoged and Maddieson (1996), Gordon and Ladefoged (2001) and Esposito and Khan (2020).

Paralinguistic function

A much broader scope of voice quality is in the paralinguistic domain in all languages, although the exact realisation is highly culture-specific. For instance, voice quality may be used to convey some communicative paralinguistic behaviour (e.g. pleading, sarcasm, humour, teasing, soothing); it can have a discourse function such as signalling attitude toward information (surprise, annoyance, sorrow, impatience) and regulating discourse (turn taking); or it can be indexical, that is, used to signal some evidential markers of the speakers, such as their social or cultural background, psychological state (upset, exhilarated, angry) or attitude toward the interlocutor without their actual intent to do so (Esling, 1978; Gobl and Ní Chasaide, 2010: p. 409-411; Hewlett & Beck, 2006; Laver, 1994). The features of voice that are maintained temporarily, for as long as a particular attitude or emotion is being conveyed, reflect a speaker's tone of voice (Crystal, 1975; Rose, 2002: p. 291). In the paralinguistic domain, a setting can span over a part or the entire utterance pronounced in a particular tone of voice (Laver, 1994: p. 397). For instance, a speaker may assume the articulatory setting with spread lips used to signal a smile (p. 115) or could employ whispery phonation to convey confidentiality (p. 153).

Brown and Levinson (1987: p. 267) described falsetto in Tzeltal, a Mayan language, as an honorific device that can span over an entire formal interaction. Grimes (1959, as cited in Suárez, 1983: p. 48) noted that in Huichol (Uto-Aztecan Mesoamerican language of Mexico) falsetto is used to express excitement, whereas speakers of Shona in Zimbabwe resort to falsetto when mocking someone who is considered to be boasting (Laver, 1994: p. 197–198). In English (RP), creak is reported to signal bored resignation (Laver, 1980: p. 126). It is also often encountered in falling intonation contours at the end of utterances when it is interpreted to have a discourse function of signalling the end of a turn in conversation (Laver, 1994: p. 196). Brown and Levinson (1987: 267) write that the creaky voice in Tzeltal signals commiseration and complaint, as well as an invitation to commiserate, while in the Totontepec dialect of Mixe (an Otomanguean language of Central America), it is used to communicate apology or supplication (Crawford, 1963, as cited in Suárez 1983: p. 48). In many cultures, whisper conveys confidentiality or secrecy, whereas breathy voice may mark intimacy (Laver, 1980: p. 122, p. 135). On the other hand, in Totontepec, whispery or breathy voice quality signals excitement or emphasis (Crawford 1963, as cited in Suárez 1983: 48). Harshness is usually understood to

denote aggression or anger, although it has also been identified as a habitual feature in some accents (Cruttenden, 2014: p. 303; Esling, 1978; Stuart-Smith, 1999). Harshness with constriction and high pitch, or the so-called "pressed voice", appears in 'rikimi' voice quality in Japanese, indicating the emotional or attitudinal state of the speaker, such as surprise, admiration and disgust (Ishi et al., 2010).

The interplay of affective states and voice quality has been studied in Serbian and English through several acoustic parameters. Protopapas and Lieberman (1997) found that the f_0 mean and maximum values correspond to perceived emotional stress in American English, whereas the f_0 range and increased jitter do not seem to play a role. By testing listeners' reactions to utterances synthesised with different voice qualities, Gobl and Ni Chasaide (2003) found that, for Southern Irish English, voice quality changes can evoke differences in speaker affect. However, there is no one-to-one mapping between voice quality and affect; instead, a given quality tends to be associated with a cluster of affective attributes. For instance, lax creaky voice may be associated with boredom, intimacy, relaxedness and contentment (Gobl & Ni Chasaide, 2003). Furthermore, Yanushevskaya et al. (2018) found that the stimuli with voice quality manipulations or the combination of voice quality and f_0 were more likely to evoke affect than the stimuli based on f_0 manipulation only, although examined languages differed in terms of the range and strength of affective responses and in terms of specific stimulus-to-affect association. Similarly, Yanushevskaya et al. (2013), who studied the interplay of loudness and voice quality to signal affective states, conclude that "loudness per se does not seem to be the major determinant of perceived affect", but it can enhance signalling of high activation states in combination with tense or modal voice quality.

In Serbian, Rajković et al. (2005) describe spectral tilt and energy distribution measures as valuable discriminants between the affective state of excitement and sorrow. Furthermore, Đorđević and Rajković (2004) underline that parameters related to f_0 values (f_0 mean, min and max) are most important for distinguishing between different pairs of emotions, with HNR and shimmer distinguishing fear-coloured speech from emotionally neutral speech and sorrow, respectively. Similarly, Grozdić et al. (2011) found that affective states of excitement and anger are reflected in jitter values, relative average perturbation and HNR measured in stressed syllables. Kašić and Ivanović (2011), in their study of the auditory and acoustic aspects of voice quality in emotional speech conveying sorrow, conclude that emotionally coloured words are often characterised by creaky voice and tremor. Similarly, from H1-H2 measures, Jerotijević Tišma (2020a) concludes that female speakers exhibit breathy phonation in the speech conveying sorrow and creaky phonation in anger, whereas male

speakers exhibit creaky voice in excitement and breathy in anger (p. 310) In addition, affective speech of sorrow exhibits lower f_0 and f_0 variability and lower intensity in both individual segments and sentences (Jerotijević Tišma, 2020a; Kašić & Ivanović, 2011). Intensity as an acoustic correlate of emotionally coloured speech was also confirmed for other emotions, increasing in segments taken from utterances conveying anger, excitement and fear and decreasing significantly in utterances conveying sorrow (Ivanović & Kašić, 2011a). Jerotijević Tišma's (2020a) results corroborate the previous findings concerning the intensity in the emotions of anger and excitement, but the researcher notes a lower intensity in fear for both male and female speakers.

The information on paralinguistic aspects of voice quality has been chiefly contributed through impressionistic research and native speaker intuition. The relatively limited number of quantitative studies on paralinguistic voice quality most likely lies in the fact that it is difficult to isolate it from the environmental influence, habitual voice quality and the context in which the speech is recorded. It is important, however, to be aware of the communicative functions of voice quality because they can be a source of within-speaker variability in forensic speaker comparison. Linguistic knowledge of the phonology of the language in question and its paralinguistic features are vital for selecting speaker-specific variables in both single-language and cross-language comparisons (Rose, 2002: p. 291). For instance, if a native speaker of French employs negative transfer when speaking English and pronounces some of the vowels with a nasal quality, the nasalised speech in the foreign language should not be misinterpreted as habitual voice quality. Similarly, speakers tend to transfer the paralinguistic system for the affect or attitude of their native language when speaking a foreign language, which may lead to misinterpretation if the analyst is not aware of the functions that voice quality may have in the given languages.

3.3.2. Informative function - habitual voice quality

Finally, a fruitful area of voice quality research lies in its informative, extralinguistic function, which is of significant interest for forensic speaker comparison, bearing in mind that voice quality is rich in evidential information about the speaker's identity in terms of physical, psychological, or social markers (Laver, 1980: p. 1; Laver, 1994: p. 14). Extralinguistic aspects of voice quality can be said to have the indexical function in the sense as described by Laver (1968) and provide the maximum possibility for the span of a setting, bearing in mind that "every single utterance produced by a particular speaker is phonetically coloured to some degree by his or her personal [voice] quality" (Laver, 1994: p. 397).

The habitual voice quality of a speaker is the combination of the organic component, that is, the anatomy of their vocal tract and their personal speaking style – which is in part defined by their idiosyncratic speech habits and, in part, characteristic of the accent-community the speaker belongs to (Laver, 1994: p. 398).

Organic variation

Considering that a person's physical constitution conditions the anatomy of their vocal tract, it is understandable that individual speakers have different acoustic characteristics of voice reflected in the auditory quality. In addition, bearing in mind the anatomical changes that occur during the life cycle, it is clear that the vocal tract features present not only the source of between-speaker but also within-speaker variability (Mackenzie Beck, 2010; Hewlett & Beck, 2006: p. 280). Mackenzie Beck (2010) writes that the primary organic sources of within- and between-speaker variability of voice quality include regular life-cycle changes (childhood, puberty, adulthood, senescence), genetic and environmental factors (e.g. sex, hormonal factors, nutrition, socioeconomic status, emotional disturbance), and consequences of physical trauma or disease (p. 157).

From birth to senescence, the body undergoes significant developmental changes, the key ones responsible for speech production being the changes in the respiratory system and the anatomy of the head and neck, including both the laryngeal and supralaryngeal vocal tract. Both phonation and articulation are influenced by the size and shape of the pharyngeal, oral and nasal cavity, larynx and vocal folds, the morphology of the skeletal structures, the contour of the palate, dental arches, elasticity of the cartilaginous framework, the physiological state of the muscles involved in phonation, and the state of the tissues covering the vocal folds (Mackenzie Beck, 2010; Esling & Moisik, 2022: p. 237-239; Hewlett & Beck, 2006: p. 257; Rose Y. et al., 2022). In favour of speaker-specificity of speech production goes the fact that these structures almost always differ even between genetically related members of a family (Mackenzie Beck, 2010: p. 156). Previous research has confirmed the change of resonance characteristics of the vocal tract (vowel formants and long-term average spectra), as well as differences in phonation quality (f_0, intensity, jitter, shimmer) between different age- and sex-groups (Mackenzie Beck, 2010). Genetic factors such as malocclusion (the non-standard relationship between the upper and lower teeth when biting together) or genetic disorders that may affect the physical development of the vocal apparatus (such as the case with Down syndrome) will also be reflected in both articulatory and phonatory settings (Mackenzie Beck, 2010). Furthermore, any changes to the vocal tract due to physical injury (tooth loss, scarring

of the tissue), temporary illness (inflammation of the tonsils, blockage of the nasal cavity, laryngitis, mouth ulcer), long-term disease (tumours of the tongue, pharynx, or larynx), vocal tract surgery and mental illnesses (depression, schizophrenia) may affect the speech production (Mackenzie Beck, 2010: p. 191; Gobl & Ní Chasaide, 2010: p. 414; Wrench & Beck, 2022: p. 19). For instance, due to the inflammation, the vocal folds become thicker; therefore, f_0 lowers. Additionally, if the swelling is asymmetrical, an irregular vibratory pattern will be reflected in perceived harshness. Fundamental frequency may also be notably lowered due to voice disorders caused by smoking or other long-term vocal exhaustions (Hewlett & Beck, 2006: p. 284).

Relying on the model initially established by Mackenzie et al. (1983, reprinted in Laver, 1991), Hewlett and Beck (2006) provide a model of the effects of vocal fold structure on speech variation based on four vocal fold parameters: (1) mass, (2) stiffness, (3) symmetry/asymmetry, (4) protrusion of any mass into the glottis so that it interferes with vocal fold closure and (5) length (p. 283). According to their theory, the higher the vocal fold mass, the lower the amplitude and rate of vibration; therefore, the lower f_0 – voice is perceived as lower in pitch and less loud. The higher the stiffness, the lower the amplitude, but the higher rate of vibration; hence higher f_0 – voice is perceived as higher in pitch but less loud. The asymmetry of vocal folds affects the regularity of rate and amplitude and is reflected in different jitter and shimmer values, which can be perceived in different degrees of harshness. Protrusion of any mass into the glottis causes incomplete adduction of the vocal folds, allowing for air leakage and is reflected in the interharmonic noise, its perceptual correlate being whisperiness. Finally, the longer the vocal folds, the lower the rate of vibration and higher amplitude, which is reflected in lower f_0 and the voice is perceived as lower in pitch but louder (Hewlett & Beck, 2006: p. 283). Mackenzie Beck (2010) adds that disrupted tissue layer geometry will also result in irregular vocal fold vibration (p. 190).

The model presented above is primarily created to describe variations of the vocal folds that occur in voice pathology (e.g. Mackenzie Beck, 2010: p. 190); however, they can be applied to predict the phonetic output of normal voice modifications as well. The study of organic variation and change of the vocal apparatus has been part of developmental phonetics and speech therapy/pathology, yet, the research and findings have significant implications for forensic speech science.

Sociolinguistic variation

Habitual patterns of voice quality do not only reflect the physical state of the speaker but also the norms of the sociolinguistic community the speaker belongs to (Abercrombie, 1967: p. 94; Esling & Moisik, 2022: p 242; Hewlett & Beck, 2006: p. 257; Honikman, 1964; Laver, 1980: p. 6-7). Numerous studies have shown that speakers of different languages or accent communities adopt different default articulatory and phonatory settings.

If we recollect that the term articulatory settings was first introduced by Honikman (1964), we shall not be surprised to learn that she was one of the first (if not the first) researchers to describe language-specific articulatory settings. Based on the observation of external articulators of native speakers, Honikman (1964) wrote that French is characterised by considerable mobility of the lips and jaw as opposed to English, which exhibits moderate lip and jaw movement. In addition, she observed that French has a lowered tongue and prominent lip rounding, whereas Russian is described as a language with a close-spread lip setting and palatalisation (p. 74-75). Honikman (1964) also described some of the settings in Hindi and Pakistani (open jaw, retroflexion), Turkish and Iranian (tongue-tip pronunciation) and German (lip-rounding). Moreover, (Cruttenden, 2014) writes that the speakers of Spanish tend to hold their tongue more forward compared to the speakers of English. In contrast, the speakers of Russian habitually retract the tongue even more to the back of the mouth (p. 302). With regard to vocal tract tension, British English is often described as lax, while French or German are described as tense (Cruttenden, 2014: p. 302-303). As far as the English language is concerned, Liverpool and West Midlands (e.g. Birmingham) accents in Britain are characterised with velarised voice of denasal quality (Abercrombie, 1967: p. 94-95; Wells, 1982: p. 93), a setting also observed in the Bronx accent in New York, and of some types of Houston accents in Texas (Esling & Dickson, 1985, as cited in Laver, 1994: p. 411). In contrast, many speakers of Australian, New Zealand, some regional varieties of American English, and British accents with RP are often characterised by nasalised speech (Laver 1994: p. 398). In addition, Wells (1982) describes Texan and Canadian male voice quality as lowered larynx voice, while the speech of working-class Norwich is qualified as raised larynx voice (p. 93). Finally, he describes the accents of lowland Scottish people as exhibiting tense and southern Americans as having lax voice quality (p. 93).

The interplay of phonatory settings and sociolinguistic variation has yielded a fruitful field for phonetic variation research. Cruttenden (2014) gives an example of speakers of Danish and Dutch, who are usually described as having breathy voice. Many accents of English, including in some parts of North America, Received Pronunciation and the Scottish

speech are often characterised by creakiness, whereas the Glasgow accent is also said to have a degree of whisperiness (Hewlett & Beck, 2006: p. 275, 277; Stuart-Smith, 1999). Esling (2013) notes that creakiness or laryngealisation is also prominent in Germanic languages such as Swedish (p. 124), and Loakes and Gregory (2022) found this phonatory quality in male speakers of Australian Aboriginal English (p. 6). In addition, speakers of Scottish English and Cockeny are often perceived to have harsh (ventricular) voice (Cruttenden, 2014: p. 303). Wagner and Braun (2003), who compared the acoustic correlates of voice quality in Polish, German and Italian, concluded that Polish speakers exhibit the highest values in HNR and could thus be perceived as having a "bright" voice, whereas the measures for the Italian speakers are indicative of perceived "roughness" (p. 654).

Finally, distinctive articulatory and phonatory settings may be adopted as habitual voice quality by people from different socioeconomic backgrounds within the same culture or used to signal identity. For example, Trudgill (1974) notices that, in Norwich speech, the working class tends to use creaky phonation, while this phonatory type is seldom employed by the middle class (p. 186). In contrast, in Edinburgh, male speakers who habitually employ creaky voice are associated with higher socioeconomic status, whereas speakers who use whisperiness and harshness are associated with lower socioeconomic status (Esling, 1978). A similar observation was made in Copenhagen Danish, where habitual creaky phonation in voiced segments may function as a social marker of upper-class speech (Laver 1994: p. 196). Furthermore, a recent study on the chronological change of the Glasgow accent showed a continuous increase in the presence of the tongue-body height setting over time (Sóskuthy & Stewart-Smith, 2020). The changes in voice quality patterns have also been noticed in recent studies exploring voice quality and identity, in which female speakers have been reported to be considerably more creaky than male speakers (Podesva, 2013; Yuasa, 2010).

As defined in this paper, voice quality is viewed as a long-term adjustment of the vocal tract encompassing both articulatory and phonatory settings, and, as such, it is seen as extralinguistic. It is a powerful, informative tool, considering that listeners rely on it to infer various information regarding the speaker, including their age, physique, mental and physical health, and, as seen above, regional background. Even though, in the present study, we aim to explore the habitual voice quality, a review of some typical linguistic and paralinguistic functions was necessary as these may play a key role in explaining potential variations.

3.4. Previous Research on Voice Quality

3.4.1. Voice quality in forensic speech science

That voice quality is a robust index of one's identity is a fairly old notion (see Garvin & Ladefoged, 1963; Laver & Trudgill, 1979, reprinted in Laver, 1991). Laver (1994) writes that a person's voice "identifiability" is based on organic foundations of the speaker's anatomy on one hand and personal style on the other. The personal style of voice quality of an individual speaker is reflected in the dominant (articulatory and/or phonatory) settings that are in part determined by the sociolinguistic (accent) community to which the speaker belongs and partly represent idiosyncratic habits (p. 398). It is the organic and idiosyncratic aspects of voice quality that are crucial in forensic speech science. According to Ladefoged (1982, as cited in Gobl & Ní Chasaide, 2010: p 405-406), language-specific voice quality features should not outweigh the intrinsic differences between speakers. For instance, if a particular dialect employs breathy/modal contrast in pronouncing some consonants, the speaker with an intrinsically breathy voice would be expected to increase the degree of breathiness to achieve linguistic contrast.

Nolan (2005) opens a debate about whether Laver's (1980) voice quality framework could find its application in forensic speaker comparison bearing in mind that Laver (1980) himself warned that his descriptive system is not designed to consider the organic type of influence on voice quality (p. 10) and that the settings are "learnable", thus any speaker with a healthy vocal tract could imitate them (p. 9). However, considering its vast application in speech pathology and usage for the description of speech disorders (e.g. Carmago & Canton, 2019; Mackenzie Beck, 1988; 2010; San Segundo & Delgado Hernández, 2021; Webb et al., 2004; Wirz, 1991), Nolan (2005) concludes that Laver's descriptive framework is capable of capturing the anatomic differences between speakers as they "perceptually replicate the effect of the relevant settings" (p. 91). Another argument for using voice quality framework in forensic speech science is that our characteristic "auditory colouring" is not merely a result of our anatomy but of how we habitually use it (Nolan, 2007: p. 113). As Nolan (2007) exemplifies, even twins with similar vocal tracts who speak the same dialect may differ in their habitual articulatory or phonatory adjustments (p. 113). The flexibility of Laver's (1980) framework, as Nolan (2005; 2007) explains, lies in the fact that it does not require that we know whether the auditory impression we have about a specific voice quality component is a result of the speaker's anatomy or their habitual vocal tract adjustment.

The scope of application of voice quality in forensic speech science is challenging to determine, considering that there is only a number of small-scope practice reviews that can give insight into it. According to the casework review[13] by Nolan (2005), until the beginning of the 21st century, there were few instances of componential voice quality description in forensic speaker comparison reports (p. 394). Nolan (2005), among other reasons, attributes the scarce application of this framework at the time to the lack of training of the phoneticians, the time-consuming nature of the task, the nature of the samples that are being compared, with particular regard to their quality and stylistic variation, and the increasing importance of the acoustic analysis in the presentation of evidence (p. 394-404). As he exemplifies, due to high within-speaker variability, sometimes the application of componential voice quality in forensic casework should be avoided for a good reason; nonetheless, the benefits of having such a system as Laver's voice quality framework at one's disposal are numerous (see Nolan, 2005). In the survey on forensic practices by Gold and French (2011), 94% of the respondents who include an auditory perceptual analysis in their casework (either in isolation or in corroboration of the acoustic analysis) reported that they examine voice quality as part of their overall procedure; and 61%[14] of these experts rely on a recognised scheme, such as Laver's (1980) voice quality framework, or a modified version of such a scheme. A more comprehensive survey focusing specifically on the application of voice quality in forensic and clinical practice was conducted by San Segundo (2021), who surveyed 42 experts from 20 different countries (24 forensic speech scientists, 18 voice therapists, and three experts working in both fields). Almost all forensic practitioners (96%) reported considering voice quality in their professional activity, the majority using a combination of the auditory and acoustic approach (42%). Furthermore, almost half of the practitioners (46%) reported observing only phonatory features, whereas laryngeal and supralaryngeal features are considered by only 19% of the experts – the remainder (35%) opted for the view that voice quality encompasses more than laryngeal and supralaryngeal features. Of the participants who reported evaluating auditory voice quality (either in combination with the acoustic analysis or individually), 72% relied on established protocols or modified versions, the most common being VPA (9 out of 13 responses). In addition, six experts reported assessing prosodic aspects either as part of or in addition to the VPA protocol. As for the experts who consider the acoustic voice quality, the most commonly

[13] The author notes that all reviewed cases took place in the British Isles between 1988 and 2002 (Nolan, 2005: p. 391).

[14] As the survey is not specific about the exact number of respondents who employ auditory perceptual analysis in their casework, considering that there were 36 participants, it can be inferred that no more than 20 (out of 36) experts employ such a scheme in their casework analysis.

reported measures include long-term average spectra, followed by jitter, shimmer, harmonic-to-noise ratio and specific software to measure laryngeal and supralaryngeal features – fewer experts marked using long-term formant distribution (San Segundo, 2021). Judged from the available practice surveys, we can conclude that the use of voice quality in forensic speech science has increased over the past two decades; however, a number of issues, such as lack of professional training in standard protocols, difficulties in adapting the protocols to different languages and non-consistent use of labels across different approaches (see San Segundo, 2021) still hinder its broader application in forensic practice.

The application of perceptual and acoustic measurements of voice quality in casework will likely increase in the future, considering a growing amount of research on its discriminatory power. Recently, San Segundo et al. (2019) proposed a methodological framework for the successful application of the VPA protocol in forensic speaker characterisation. Using a modified 32-feature version of VPA employed in the JP French Associates forensic laboratory in the UK, the researchers assessed the voices of 99 speakers of Standard Southern British English, comparing three methods of inter-rater agreement evaluation (absolute percentage agreement, agreement within one scalar degree and Fleiss' kappa). The results indicate that the inter-rater agreement is highly setting dependant. However, strong results can be achieved provided that there is a calibration session between raters (San Segundo et al., 2019). The researchers also examined the correlation between the individual settings – bearing in mind that correlation between the parameters should be avoided in FSC (Gold, 2014; Nair et al., 2014; Rose, 2006; 2013b), detecting the most apparent positive correlation between the raised larynx and tense larynx settings and the most prominent negative correlation between lax and tense vocal tract. However, they conclude that the correlations are not strong enough to collapse the correlated settings into one (San Segundo et al., 2019). Mackenzie Beck (2005) warns that forensic speaker comparison based on VPA alone does not yield strong evidence (p. 310-311). Namely, in an earlier case (Mackenzie Beck, 1988), the vocal profiles of a questioned and a known speaker were compared to 50 other vocal profiles to determine the likelihood of any two speakers having equivalent levels of similarity. The outcome was that around 14% of comparisons of different-speaker pairs yielded contrary-to-fact results (p. 238). Notwithstanding, numerous recent studies have confirmed that VPA can be used to corroborate other forensic analyses. French et al. (2015) compared the performance of MFCCs, LTFDs and VPA ratings in speaker discrimination and explored the relationships between the three sets of parameters. According to their results, all three systems performed relatively well, with MFCCs exhibiting perfect performance in same-speaker pairs, which, as

the authors explain, is in line with the previous research (French et al., 2015). The distance scores between speakers were correlated for each of the measures, indicating a stronger relationship between MFCCs and LTFD than between either of the two and the VPA. According to the authors, this indicates that the auditory VPA offers different information as opposed to MFCCs and LTFDs with regard to speaker characterisation, and, as such, it can be used to complement the acoustic (and automatic) analysis and improve a forensic speaker comparison system performance (French et al., 2015). That voice quality analysis can contribute to the overall system performance was also suggested by Gonzalez-Rodriguez et al. (2014) and Hughes et al. (2017), who found that false acceptance errors in different-speaker comparisons in an i-Vector- and MFCC-based ASR could be explained by auditory analysis. Gonzalez-Rodriguez et al. (2014) conclude that phonation (particularly creak) was the most helpful diagnostic. In Hughes et al. (2017), the corpus was tagged in terms of laryngeal and supralaryngeal settings; thus, the researchers reveal that the speakers for whom the contrary-to-fact likelihood ratios were obtained could be differentiated on the basis of, for instance, lip-spreading or a close-jaw setting and more clearly based on phonation.

In addition to forensic speaker comparison, voice quality has found application in speaker profiling and the construction of voice parades to estimate the salient features of the voices that will be presented to the earwitness during a voice identification task (see Nolan, 2005; San Segundo, 2021: p. 22). When constructing a voice parade for an earwitness to identify the suspect, it is necessary to ensure the "fairness" of the experiment by choosing the voices in such a way that none of them would potentially bias the listener (in either direction) due to some dominant feature, such as nasality (de Jong et al., 2015; Nolan, 2007; San Segundo et al., 2018). Screening the samples through the voice quality framework could help identify those voices that the earwitness is likely to discard instantly, thus effectively reducing the number of foils in the line-up (Nolan, 2005: p. 409). San Segundo et al. (2018) propose annotating the voice databases with VPA information prior to voice parade design as it would enable the automatic selection of similar foils. The researchers showed that the non-hierarchical k-means method separated the 99 age- and dialect-matched speakers in two clusters – lowered larynx, lax larynx, creaky and breathy phonation v. raised larynx, tense larynx, harsh and whispery phonation (San Segundo et al., 2018).

The most recent research that reveals the importance of voice quality for naïve listener judgments has been undertaken by McDougall and her colleagues at the Cambridge University, most of which has been conducted as part of the IVIP project – "Improving Voice Identification Procedures" (McDougall, 2023). Nolan et al. (2011) and McDougall (2013a)

assessed the correlation of acoustic and voice quality parameters with naïve listener similarity ratings to understand on which features listener rely in speaker identification. It was found that the most important properties for perceived voice similarity are mean fundamental frequency, creaky voice, larynx height, larynx tension and pharyngeal expansion (Nolan et al., 2011; McDougall, 2013a). When the research was later extended to multiple varieties of British English, it was found that for different accents, different acoustic features may be crucial in voice similarity judgements, with f_0 and F1 being predominant correlates of high similarity ratings (McDougal, 2021), irrespective of the sample duration (McDougall et al., 2022).

3.4.2. Acoustic analysis of voice quality in FSC

Regarding the acoustic measures of voice quality in forensic speaker comparison, formants have been most widely investigated. Long-term formant distribution (LTF) is a global representation of vowel formant frequencies over a longer speech sample (Nolan & Grigoras, 2005). Compared to segment-based formant values, long-term formant frequencies are independent of linguistic information to a great extent (Jessen, 2010; Nolan & Grigoras, 2005) and are often described as a valid parameter in forensic speaker comparison (Gold 2014; Nolan & Grigoras, 2005). In addition, unlike segment-based formant frequencies, they are easier to extract and measure as no segmentation is required. Finally, LTF values are reported to be language-independent (Jessen & Becker, 2010) and are, therefore, widely explored in cross-language forensic speaker comparison (see Chapter 2.4.2.).

Nolan and Grigoras (2005) employed LTF measurements and distribution shape of the first and second formant to compare the anonymous telephone recordings to the suspect voice, revealing that the suspect had significantly higher LTF2 than the voice in the recordings.

Moos (2010) analysed LTF1, LTF2 and LTF3 values of read and spontaneous speech of 71 speakers of German transmitted over a mobile phone. The long-term values of the third formant (LTF3) emerged as more beneficial for voice comparison due to the lowest within-speaker variability (p. 19-20). In addition, for most speakers, this formant had the least notable difference in values in read-out and spontaneous speech (p. 15).

Gold et al. (2013) tested the performance of the long-term distribution of the first four formants of 100 speakers of Southern Standard British English under the Likelihood Ratio framework. According to their results, LTF3 has the highest percentage of correct same-speaker and different-speaker comparisons, as well as the lowest equal error rate (17%). It is followed by LTF4 (EER = 22.4%), LTF1 (EER = 28.06%), and finally, LTF2 (EER = 31.65%) (Gold et al. 2013: 4). These authors, however, find that LTF3 has the highest C_{llr} score (1.0731). The

EER score improves significantly with the combination of parameters: for the combination of the first three formants, EER is 11.47%, and for all LTFs, EER is 4.14%. C_{llr} score, in their research, is the lowest for LTF4 (0.8085), and it improves slightly with the combination of all parameters (0.5411) (p. 4).

Asadi and Dellwo (2019) employed the linear model to explore long-term formant features and long-term f_0 of 12 male speakers of Persian in two non-contemporaneous sessions. Their results confirmed that LTF3 and f_0 are speaker-specific and that variability across recording sessions is not significant for these parameters for most speakers.

Hughes et al. (2018) tested LTF values in mismatched recording conditions under the LR framework, finding that the mismatch has a detrimental effect on the overall performance of all parameters. The authors noticed a significant variability in the individual behaviour of speakers; however, they were unable to predict which speakers would perform well or badly neither from the mean formant values nor from the auditorily-judged voice quality features (p. 231).

Lo (2021a) explored LTFDs and formant bandwidths in Canadian English and French by modelling the data using the GMM-UBM approach and calculating likelihood ratios. In both languages, formant-based comparisons yielded mean C_{llr} between 0.61 and 0.74 and EER between 18.8% and 27.2%, while including bandwidths improved the system performance, C_{llr} 0.40-0.51 and EER 10.8%-14.6%. (p. 204). The author notes that the first formant performed the best, whereas the second formant consistently produced the highest C_{llr} and EER scores, with minor differences, depending on the language (p. 205).

Few studies employ likelihood ratio calculations on the basis of the acoustic parameters of the glottal source. Nonetheless, numerous studies have explored the robustness of these parameters within and between speakers. A separate section will be devoted to the variability of laryngeal voice quality in bilingual speakers (see Chapter 3.4.4.), whereas here, we will focus on more general forensic implications of voice quality research.

Harmegnies and Landercy (1988), who investigated within-speaker variability of LTS in French speakers, concluded that while LTS is a relatively robust parameter overall, its application in speaker recognition depends on the subjects. Namely, some speakers exhibit higher variability both within the same and across different texts.

More recently, using the principal component analysis, Lee et al. (2019) analysed within- and between-speaker variability of voice quality parameters of 100 American English speakers (50 males and 50 females). The acoustic measures (1) fundamental frequency, (2) formant values, (3) harmonic source spectral shape, (4) interharmonic source/spectral noise and

(4) variability were performed on vowels and approximants. According to their results, most of the variance was induced by the balance between higher harmonic amplitudes and inharmonic energy (females = 20%, males = 22%), followed by formant frequencies and their variability (12%). The remaining variance appeared largely idiosyncratic, indicating that individuals have speaker-specific voice space.

Using the same methodology as the previous study, Lee and Kreiman (2019) explored speaker variability across two tasks (spontaneous speech and reading), revealing that speakers' voice spaces do not differ significantly. The only feature that emerged as different was fundamental frequency variance, which accounted for more variability in spontaneous speech.

Furthermore, Vaňková and Skarnitzl (2014) assessed within- and between-speaker variability of various spectral amplitude measurements (H1*-H2*, H2*-H4*, H1*-A1*, H1*-A2* and H1*-A3*) across different speaking styles in Czech. According to their results, H1*-H2*, H1*-A1* and H1*-A2* are not only stable across speaking styles for one speaker, but they also exhibit high between-speaker variability, outperforming formant values (p. 1081).

For the purposes of FSC, Enzinger et al. (2012) explored the voice source features (several jitter and shimmer values, glottal-to-noise excitation ratio, mucosal wave cepstra, frequency and amplitude measures) in three different transmission channels, comparing the results to a baseline MFCC GMM-UBM system. The voice source features could not match the baseline system performance with the exception of the mobile phone-to-landline recordings; thus, the authors conclude that the measures are irrelevant to forensic speaker comparison. However, it should be noted that the voice source feature extraction was performed solely on the nasal /n/ because it was the most represented segment in the corpus, and the authors aimed to use a sustained segment production. Had the researchers used vowels or all voiced segments, they would have been able to extract more long-term information regarding the laryngeal voice quality.

Our point is corroborated by recent research by Cardoso et al. (2019), who also performed the acoustic analysis of long-term laryngeal voice quality, comparing the results to an MFCC-based ASR system across four channels. The best performance when the entire voice quality system was observed was found in high-quality recordings, with EER between 5.8%-12.2% and a C_{llr} between 0.26-0.63. Separate calculations of the parameters based on additive noise and spectral tilt yielded a slightly weaker performance (mean EER 17.6% and 13.1% and mean C_{llr} 0.61 and 0.54), with spectral tilt measures being unaffected by the transmission channel. The results not only confirmed that the acoustic analysis of laryngeal voice quality

could be a valuable parameter in FVC but also showed that such an analysis could improve ASR performance, especially in degraded-quality channels such as low-quality mobile phone conversation (Cardoso et al., 2019).

Most recently, Holmes (2023) explored the discriminatory power of fundamental frequency, formants, intensity, HNR, auto-correlation and several jitter and shimmer measures using a top-down approach. The system baseline performance was calculated using the GMM-UBM likelihood ratio and contribution of individual parameters to the system was assessed by removing one feature at a time. If C_{llr} decreased, the feature was seen as detrimental to the system. Parameters that Holmes (2023) considers integral for speaker characterisation include f_0, intensity, higher formants, whereas the performance of lower formants (F1 and F2) largely depends on the accent. The author also rejects HNR, mean autocorrelation, jitter and shimmer on the basis that the system yields higher C_{llr} scores after the removal.

3.4.3. Voice quality and telephone transmission

In a typical forensic speaker comparison case, the unknown sample provided for the analysis is a recording of a telephone conversation, whereas the known sample is a recording of the police interview with the suspect (Künzel, 2001; Nolan, 2005; Rose, 2003). Such a mismatch has always posed a challenge to forensic practitioners due to the recording quality and speaking style mismatch encountered in the two contexts. In order to approach the realistic scenario, the present study is performed on the corpus assembled over the mobile phone; therefore, it is necessary to acknowledge the known effects of the transmission channels on voice quality.

Nolan (2005) believes that one of the main reasons linguistic-phonetic voice quality has not found its proper place in forensic speaker comparison casework is precisely due to the limitations imposed by the telephone. It is now widely accepted that landline transmission limits speech signal in such a way that the sound energy below 300 Hz and above 3,500 Hz may be lost, and the distortions of the spectral shape could be encountered near these frequencies (Künzel, 2001; Nolan, 2007; Rose, 2003). Laver et al. (1981) noted that even for the auditory analysis of vocal tract features, it is necessary to have good-quality audio as some settings are prone to distortion by poor recordings. For instance, the hiss or background noise can interfere with the perception of whisperiness, breathiness or audible nasal escape (Laver et al., 1981), while the lost frequencies outside the mentioned bandwidth can have a perceptual effect on articulatory settings, such as palatalisation or nasalisation (Laver et al., 1981; Nolan, 2005; 2007). Concerning laryngeal settings, the first harmonic of a male voice, which is a known

acoustic correlate of voice quality when compared to the second (or some other) harmonic, could be as low as 75 Hz and therefore affected by the degraded signal, much in the same way as the high-frequency aperiodic energy of breathy and whispery settings (Nolan, 2007: p. 119). All things considered, Nolan (2007) believes that, while it is not impossible to perceptually "reconstruct" what the sample would have sounded like had it not been passed through the telephone, assuming that a regular (good quality) and a telephone recording are a good match in terms of voice quality presents a serious "leap of faith" (p. 120).

An influential paper by Künzel (2001) demonstrated the existence of what he referred to as "the telephone effect" on lower vowel formants. In his study, with 20 speakers recorded face-to-face and over a standard digital telephone line (ISDN), for every single subject, there were significant differences in the first formant values, which always proved to be higher in the telephone recording condition (p. 87). The difference was the largest for close vowels such as [i] and [u], medium for vowels such as [e] and [o] and negligible for open vowels like [ɔ, a] (p. 89). Künzel (2001) warns that /i:/ can be wrongly perceived as /ɪ/ because of the higher F1 due to the loss of low-frequency energy (p. 94). However, as the author himself acknowledges, not all the speakers are equally affected by the telephone transmission (p. 89); what is more, in some cases, formant centre frequencies are lower in the telephone condition (p. 93). The results were replicated by Rose (2003), who investigated the vowels of Broad Australian in realistic forensic conditions. The researcher suggests excluding F1 measurements of /i/, /u/, /o/ and /ə/ from such comparisons (Rose, 2003: p. 99-5107). Similarly, Lawrence et al. (2008) investigated the acoustic and perceptual effects of landline telephone transmission on the vowels /iː/, /æ/ and /uː/ in Standard Southern British English, confirming the effects on *f*1 for the close front and back vowel (p. 170), and minor effects on F2 and F3 on the back vowel alone (p. 171). However, the differences between the direct and telephone recordings as perceived by the trained listeners were not significant for the close vowels – only /æ/ was described as different in terms of backness and height (p. 180). Similar findings were presented by Byrne and Foulkes (2004) for the speech transmitted over a mobile phone. According to their results, however, the effect of mobile phone transmission on F1 values is even more severe than in the landline recordings. Guillemin & Watson (2008) investigated the effect of AMR codec in the GSM network on fundamental frequency and vowel formants. Their results suggest a significant difference in frequency distribution between the source recording and the one transmitted through the chosen GSM codec, reflected in a very high standard deviation. The authors suggest that the codec increases the voicing probability for individual frames, with a 10% higher number of voiced frames in the codec-influenced speech (p. 208). However, as with

all previous studies, the effect of telephone transmission seems to be highly speaker-specific. Significant between-speaker inconsistencies were found for formant-tracking of F2 and F3. The formant tracking issue was confirmed by Carne (2015), who performed likelihood ratio comparisons of Japanese diphthong /ai/ in direct and mobile phone recordings. The author notes an 18% reduction in system validity when the mismatched conditions apply, proposing that F3, in particular, should be excluded when the system cannot correctly track it in high vowels (p. 3474).

Nolan (2002) argues that, despite the notable variation of F1 in Künzel's (2001) data, vowel formants should not be disregarded in forensic speaker comparison. He draws attention to the stability of F2 (cf. Künzel, 2001; Lawrence et al., 2008; Rose, 2003), which is more sensitive to the nuances in anatomical and articulatory differences (Nolan, 2002: p. 77-78), concluding that, not unlike any other acoustic parameter, vowel formants should be analysed with caution in forensic speaker comparison cases where one of the samples is a telephone recording. Several years later, he reinforces his view by stating that "formant values reflect the interaction of three potentially identifying sources: the linguistic accent, the anatomy of the individual's vocal tract, and the speaker's acquired articulatory strategies" (Nolan, 2007: p. 115-116) and are thus of great importance to forensic speaker comparison.

Let us now turn to the more recent research. Passetti and Constantini (2019) used the VPAS adapted for Brazilian Portuguese to assess the perceived voice quality of subjects with dysphonia in direct and mobile phone recordings. Their results indicate that the most significant rating discrepancies stemmed from the velopharyngeal system, respiratory support, laryngeal tension, speech rate, and creaky voice. Despite the noted distances in the ratings between the two recording contexts, the authors conclude that VPA is "a relevant scientific tool" to apply in forensic casework. Furthermore, Pommée and Morsomme (in press) explored the effect of mobile phone transmission on the acoustic and perceptual parameters (GRBAS) of voice quality on spontaneous speech and sustained vowels. They found that the frequency cut-off is below 100 Hz and above 3,700 Hz. The most stable measures across the recording conditions were local jitter, the harmonics-to-noise ratio, the period standard deviation and pitch measures. The acoustic voice quality index is higher in telephone recordings, while the breathiness index is lower. Regarding the perception, despite the low inter-rater agreement, intra-rater scores across recording conditions are relatively stable. The most affected parameters are reported to be breathiness, rated as lower in mobile phone recordings, and roughness, rated higher on sustained vowels for men. The authors advise against relying on the acoustic and perceptual measures of these parameters in channel-mismatched conditions. In a forensic

experiment already mentioned in section 3.4.1., Cardoso et al. (2019) assessed the discriminant ability of the acoustic correlates of laryngeal voice quality (f_0, Cepstral Peak Prominence, HNR, H1-A1, H1-A2, H1-A3, H1-H2, H2-H4) in the acoustic signal recorded over four different channels (studio recordings, landline, high and low bit rate mobile phone samples). The authors concluded that not only are the tested parameters relatively robust to the channel variation, but also their discriminatory power becomes significantly more valuable when the ASR system falters with low-quality samples (Cardoso et al., 2019). It is worth noting, however, that the experiment was not performed in channel-mismatched conditions; hence it does not refute the general concerns regarding direct and mobile phone recordings comparisons. Most recently, Klug and Niermann (2024) found that the chosen f_0 algorithm may affect the results for breathy voice quality in the mobile recording condition. They suggest using the SNACK, (Talkin, 1995, as cited in Klug & Niermann, 2024) rather than STRAIGHT (Kawahara et al., 1999) algorithm for this purpose. Relying on these results, Klug (2023) found that CPP, HNR_{05}, and the spectral tilt parameters A1-A3, and H4*-A2* showed systematic differences between breathy and non-breathy speakers under the mobile recording condition (p. 86) and suggest using these rather than low-frequency measures for speaker differentiation in mobile phone recordings.

Previous speaker comparison research based on MFCCs (Enzinger, 2014; Nair et al., 2016), GMM-UBM (Alexander et al., 2005; Zeljkovic et al., 2008) or semi-automatic formant-based measures (Hughes et al., 2020) have consistently yielded higher C_{llr} and lower credible intervals in channel-mismatched conditions. However, system enhancement and channel effect compensation are possible provided that various statistical models and neural network deep learning technologies are applied (see Enzinger, 2014; Li et al., 2020; McLaren et al., 2016; Muralikrishna & Dinesh, 2022).

Apart from the limitations due to the encoding and technical specifications of the devices, a potential source of variation when speaking over a phone could also be the speaking style or the position of the speaker and the device. For example, a phenomenon often observed in telephone-recorded material is the so-called Lombard reflex, named after Étienne Lombard, a French otolaryngologist who was the first to describe that there may be an increase in speech loudness due to background noise (see Lombard, 1911). Reviewing some of the most influential literature on the topic, French (1998) summarises that Lombard speech can be characterised by "a reduced rate of speaking (measured either in terms of syllable production or relative vowel duration), a higher average frequency for the first formants of vowels and a higher average pitch" (p. 61). Additionally, Jovičić et al. (2015) demonstrated the change in LTAS, LTF and central formant values for both male and female speakers in five conditions of mobile phone

usage (regular, with candy in the mouth, with a cigarette between lips, the phone between the cheek and the shoulder and mouth and phone covering).

From the look of the previous research, the situation regarding channel-mismatched values of voice quality is hardly optimistic – yet, this is the reality that forensic experts face in their daily work (Rose, 2003: p. 99-107). Notwithstanding, since the present research aims to assess the effects of language mismatch rather than channel mismatch, it would be inappropriate to introduce another source of variation, which is why a corpus of mobile phone recordings was created for the present study.

3.4.4. Voice quality and bilingualism

A broad definition of bilingualism is the regular use of two or more languages (Grosjean, 1982: p. 1). Soares and Grosjean (1984) distinguish between the monolingual and bilingual speech modes, two ends of the situational continuum of the everyday life of bilingual speakers (p. 380). The former involves adopting the language of the monolingual interlocutor (either the first or the second language of the given bilingual) and deactivating the other language as much as possible. On the other hand, the latter implies that they speak to bilinguals with whom they usually mix languages by choosing the base language and incorporating the elements of the other language, i.e. code-switching (p. 380-381). As far as the monolingual speech mode is concerned, however, there has been evidence that first language deactivation is never total (p. 381).

Foreign language learners can be observed as emergent bilinguals (Blake, 2018); however, whether they can maintain the monolingual mode in the foreign language depends on their foreign language competence. Namely, nowadays, foreign language learners often engage in the foreign language outside the classroom in everyday activities, including social networks where they actively participate in conversations with other native and non-native speakers of the language in question, thus becoming *language users* and not merely *learners* (Kao & Wang, 2014). In addition, even though they live in a monolingual community, many learners rely on the foreign language to communicate at work with their foreign colleagues or clients.

Despite the recommendations that voice quality research should find its application in foreign language teaching and learning (Honikman, 1964; Wilhelm, 2019), there are not many studies that explore how voice quality is affected by language switch. According to Esling (2000), “each language has its own pattern of physiological behaviour in which articulators are trained to operate in different ways based on the language’s phonetic constituent” (as cited in Ferreira Engelbert, 2014: p. 157). However, whether the differences in voice quality in

bilinguals are a matter of the acoustic structure of the phonemes or the result of language-mapping on a psychological level remains a matter of debate up to this day.

Almost 40 years ago, Harmegnies and Landercy (1985) compared speech spectra of Dutch and French bilinguals, concluding that most of the variability stemmed from speaker-specific rather than the language differences. The authors, however, warned that phoneme distribution in the two languages may have a slight influence. At the end of the same decade, Harmegnies et al. (1989) published another research on the effect of language change on voice quality, featuring 10 Catalan-Castilian bilinguals. The results once again confirmed that LTAS measure is highly speaker-specific, but the language effect was more prominent due to structured experimental conditions. The researches finally hypothesise that "various degrees of bilingualisrn result in various degrees in inter language coherence" and propose controlling the language proficiency factor in future research (p. 2491).

Toward the end of the previous century, Bruyninckx et al. (1994) explored the influence of language on voice quality of Catalan and Spanish bilinguals by measuring long-term average spectrum (LTAS) in 12 male and 12 female speakers. The between-language variability in voice quality was higher than within-language variability for each speaker, regardless of their sex or dominant language. The within-language variability, however, was on average higher in the dominant language. The researchers propose that the voice quality variability may be relative to the second language proficiency. In conclusion, the researchers express the opinion that the reasons for between-language LTAS shifts lie in the "voice quality shifts due to variation in the dominant features of the articulatory behaviour" and that the language phonemic inventory is not responsible for between-language variability (p. 28-29).

The language effect on voice quality was also confirmed by Ng et al. (2012), who studied the speech of 40 Cantonese speakers proficient in English. The LTAS measures, including fundamental frequency, mean spectral energy (MSE), and spectral tilt (ST) were found to differ across languages, while the values of first spectral peak (FSP) remained the same on average. Their results were replicated in a research by Bahmanbiglu et al. (2017), who compared the language of 32 Farsi-Qashqai bilinguals. Interestingly, in both studies, the mean spectral energy was found to be higher in the speakers' dominant language, whereas the spectral tilt was lower.

A pilot study (Ferreira Engelbert, 2014) with three native speakers of Brazilian Portuguese and one native speaker of American English compared L1 and L2 speech production, analysing intra-speaker variation in phonation types using LTAS measures, f_0, H1-H2 and noise-to-harmonics ratio (HNR). The results revealed that f_0 was lower for the speakers'

native language. In addition, H1-H2 indicated that Portuguese was spoken with more open glottis while, in English, all speakers showed tendency toward modal voice (p. 167). The similar tendency was observed by Ferreira Engelbert et al. (2016) in a more extensive research with 16 native speakers of Brazilian Portuguese who employed creaky voice when speaking English (L2) and breathy voice for native Portuguese. The authors explored the alpha measure, difference between the amplitude peaks and f_0 measures across two languages (Portuguese and English) and two tasks (reading and semi-spontaneous speech). The LTAS measures revealed significant differences across the languages in the reading task, as if the participants had a specific "reading mode" for each language (p. 45). Comparably to Bruyninckx et al. (1994), the authors found greater variability of voice quality in the dominant language. The authors, however, do not exclude the possibility of influence of paralinguistic factors and individual differences on their results (p. 45).

Using a different methodology, Pillot-Loiseau et al. (2019) studied the contact dynamics of English and French and how it reflected on voice quality of the speakers in the course of a 3-month monitored interaction. The creaky portions of speech were annotated manually and the selected measures included the percentage of creaky syllables and percentage of creaky speech. The researchers found that creaky voice was not only more frequent in L1 English than in L1 French, but also in L2 English compared to L1 French for each of the participants (p. 9-10). In addition, the authors found a significant correlation between the proportion of creakiness in L1 and L2 speech for every given speaker, which supports the hypothesis that the speaker-specificity of voice quality outweighs sociolinguistic differences.

Schwab and Goldman (2016) explored fundamental frequency in early and late bilingual speakers of English and French, English and German and French and German. According to their results, f_0 is lower in English than either in French or German, regardless of whether it is L1 or L2, whereas the speakers of French and German exhibited no differences in f_0 across languages. Furthermore, as in previous studies (cf. Čubrović, 2020; Kainada & Lengeris, 2015; Marković, 2011; Paunović, 2013; 2015; 2019), f_0 variability was found to be higher in the dominant (first) language of the speakers.

Furthermore, Schwartz G. (2019) measured the spectral balance in the linguistic contrast between tense and lax vowels of L2 English in two proficiency groups of native Polish speakers. He found that the Polish speakers with C2-level proficiency use voice quality to help maintain the distinction between tense and lax vowels, while those with a lower level of English proficiency do not.

Recently, Johnson et al. (2020) examined the connected speech of early Cantonese-English bilinguals by extracting 24 filter and source-based acoustic measurements of voice quality, including the mean and standard deviation of f_0, F1-F4, H1*–H2*, H2*–H4*, H4*–H2kHz*, H2kHz*–H5kHz*, Cepstral Peak Prominence, Root Mean Square Energy and subharmonics-harmonics amplitude ratio and subjecting them to principal component analysis (PCA). The researchers conclude that "while talkers vary in the degree to which they have the same 'voice' across languages, all talkers show strong similarity with themselves" (p. 2387), which corroborates the observations disclosed in the oldest studies on the topic (cf. Harmegnies & Landercy, 1985; Harmegnies et al.,1989).

In the latest research, relying on Johnson et al.'s (2020) and Lee et al.'s (2019) methodology, Asiaee and Asadi (2022) examined the voice quality parameters in 10 simultaneous Persian-Kurdish bilinguals. Results from t-test analysis revealed that while all f_0, formants, source spectral shape, and spectral noise parameters remained stable across Persian and Sorani Kurdish, almost all covariance measures varied significantly (p. 28). The researchers further performed the principal component analysis to extract the common voice space of each language. The results reveal that formant dispersion (FD), F4 and F3 account for 10.9% of the variance in the Persian data set and 11.7% in Sorani Kurdish. These are followed by spectral shape measures (H4*-H2kHz* and H2kHz*-H5kHz) and F2, representing 10.2% of the variance in Persian and 11% in Sorani Kurdish. Regarding speaker-specific variability, the researchers reveal that the most prominent parameters are the formants and their covariance measures in both languages (p. 30). The authors conclude that the acoustic patterns of voice are similarly structured across languages and that, in most cases, "bilingual speakers have the same voice when they switch from one language to another" (p. 31).

As can be inferred from the literature review on voice quality and bilingualism above, we have reached the full circle in the 40-year period of research. More advanced statistical analysis and methods have enabled to neutralise the language effect on individual voice quality features; however, whether the voice quality across languages remains the same seems to be largely speaker-specific. While some authors argue that it is the phonemic and prosodic structure of a language that influences voice quality (Asadi & Asiaee, 2022; Harmegnies & Landercy, 1985), others disagree (cf. Bruyninckx et al., 1994; Johnson et al., 2020). In addition, views have been proposed that "the degree of bilingualism" may affect the speaker variability across languages (Harmegnies et al., 1989). Johnson et al. (2020) propose that turning to listeners will aid to decipher what meaningful variation within a voice is and would lead to the ultimate goal – "to understand how the acoustic variability and structure of

talkers' voices maps onto listeners' organization of a voice space for use in talker recognition and discrimination" (p. 2390).

In summary, the results of the previous research support, to varying degrees, our initial hypothesis that individuals retain their voice quality when speaking a foreign language and could thus be identified based on the voice quality parameters. In order to interpret the influence of the degree of bilingualism, the present research will incorporate the assessment of the speakers' foreign language proficiency, focusing on pronunciation. In addition, to establish the interface between perceived voice similarity and acoustical measures of voice quality, the study takes into consideration the observations by expert and naïve listeners. Before proceeding to the experiments, let us compare Serbian and English vowel systems, as these are the segments where voice quality measures are most prominent.

4. Serbian and English Vowels through the Lens of Voice Quality

Considering that previous researchers have found that language phonemic system may influence voice quality parameters as well as cross-language forensic speaker comparison (see see Bhattacharjee & Sarmah, 2012; Cho & Munro, 2017; Harmegnies & Landercy, 1985; Jovičić & Grozdić, 2014; Nagaraja & Jayanna, 2013), we will compare the sound systems of the two languages explored in the present study (Serbian and English), focusing on vowels, as most of the acoustic parameters of voice quality are extracted in these segments. Furthermore, since the speakers in the present study are not simultaneous but rather sequential bilinguals (Flynn et al., 2005) who rely on their dominant language significantly more than on the second language in their daily life – and degree of bilingualism is a potential factor in voice quality retention across languages (Harmegnies et al., 1989) – it is appropriate to review some of the previous studies on English language acquisition by Serbian learners.

4.1. Comparison of Serbian and English Vowel Systems

Considering sociolinguistic variation, a vast number of dialects and a chronological change of speech, it is beyond challenging to define the sound system of a language, especially one such as English, which is officially spoken across four continents. For instance, Grieve et al. (2013), based on the multivariate spatial analysis of 38 vowel formant variables measured in 236 cities across the United States, completely redrew the dialectological map from the Atlas of North American English (Labov et al., 2005). Nonetheless, for the present research, we will limit our description to the varieties for which there is available data. The situation for Serbian is somewhat different. Namely, due to the overwhelming prescriptive trend to maintain the prestigious status of the standard language (see Paunović, 2009; Sretenović, 2015), many linguistic books and papers describe different pronunciations of specific sounds as "incorrect" (e.g. Subotić et al., 2012). In addition, apart from a number of papers focusing on foreign language acquisition and featuring a relatively small number of speakers (e.g. Bjelaković, 2018; Marković, 2012; Marković & Jakovljević, 2016; Paunović, 2011; Tomić & Milenković, 2019; 2021), there are not any recent large-scale studies that comprehensively[15] depict pronunciation of Spoken Serbian across different regions. Therefore, for the sake of comparison, the numerical

[15] Comprehensively here implies using large corpora of spontaneous speech rather than carrier sentences and without observing the pronunciation through the prism of prescriptive norms.

formant values of Serbian vowels will be drawn from the studies mentioned above, whereas the formant values of English will be drawn from recent research featuring native speakers of Standard British and American English.

Serbian can be described as a pitch-accent language whose prosodic system has two pitch accents, falling and rising. Each accent is defined by a characteristic pitch shape and stress, correlating with an increase in duration (Sredojević, 2017; Subotić et al., 2012; Zec & Zsiga, 2009). The vowel space of the Standard Serbian language includes five vowels, which, according to the openness of the mouth and position of the tongue, can be described as follows: front high slightly spread /i/, front mid unrounded /e/, central low unrounded /a/, back high rounded /u/ and back mid rounded /o/ (Simić & Ostojić, 1996: p. 178-179; Stanojčić & Popović, 1989: p. 27-28; Subotić et al. 2012: p. 44). Due to the influence of the pitch accent, in Standard Serbian, there are short and long syllables, which, in turn, influences the length of vowels on the suprasegmental level (Ivić & Lehiste, p. 2002; Simić & Ostojić, 1996; Sredojević, 2017; Subotić et al., 2002). In some dialects, vowels in long and short syllables may have different qualities. For instance, in the speech of Novi Sad (Šumadija-Vojvodina dialect), high [e] is found in long syllables, while, in short syllables, it is more common to encounter the allophone [ɛ] (Ivić & Lehiste, 2002: p. 123; Marković, 2012: p. 104; Marković & Jakovljević, 2016: p. 218; Tomić & Milenković, 2019: p. 161). All vowels in Serbian are monophthongs; nonetheless, diphthongisation may be encountered in some varieties in the north (Subotić et al., 2012: p. 45). Pitch accentuation, however, is not present in all of the varieties of the Serbian language. For instance, in the Prizren-Timok area (where the participants of the current research are from), there is no pitch accent as such – there is only expiratory stress, a result of the elimination of all quantitative and qualitative differences (Ivić, 1956), much like in English. However, recent research has indicated that younger urban speakers, while not making quantitative distinctions between long and short vowels, tend to realise the post-stressed vowel pitch contours differently in the words where rising and falling pitch accents are expected to occur (Tomić, 2020).

The number of vowels in the English language may vary depending on the dialect or sociolect of the speaker (Cruttenden, 2014: p. 96-97; Kreidler, 2004: p. 49; Ladefoged, 2001: p. 22; Roach, 1991: p. 14-22). Monophthongs that can appear in stressed syllables in the standard British English, known as RP (Received Pronunciation) or SSBE (Standard Southern

British English), are the following[16]: front high unrounded /iː/ and /ɪ/, front mid unrounded /ɛ/, front mid-to-low unrounded /æ/, central mid unrounded /ɜː/[17], central low-to-mid unrounded /ʌ/, back high slightly rounded /uː/[18] and /ʊ/, back high-to-mid rounded /ɔː/, back low-to-mid slightly rounded /ɒ/ and back low unrounded /ɑː/[19] (Cruttenden, 2014: p. 96; Ladefoged, 2001: p. 27-28; Ladefoged & Johnson K., 2011: p. 87-89; Roach, 1991: p. 14-22). In addition, Cruttenden (2014) adds another cardinal vowel /ɛː/, which is sometimes realised as a diphthong [ɛə] (Kreidler, 2004: p. 54)[20]. A monophthong that appears only in unstressed syllables in English is schwa /ə/ (Ladefoged & Johnson K., 2011: p. 42). In Standard American speech (General American English), vowels followed by /r/ have the so-called r-colouring. In some rhotic accents, including General American, speakers do not make a distinction between /ɒ/ and /ɑː/; depending on the region, /ɑː/ or /ɔː/ are used instead of /ɒ/ (Kreidler, 2004: p. 55; Ladefoged & Johnson K., 2011: p. 41; Roach, 1991: p. 240). Monophthongs in English are often classified as tense /ɑː, iː, ɔː, uː/ and lax vowels /ʌ, ɪ, ɒ, ʊ, ɛ, æ,/. Tense vowels have an inherently longer duration than their lax counterparts; however, the pairs do not differ only in length but also in quality, hence the different transcription symbols. It should be noted, however, that vowel /æ/ is usually longer than the rest of the lax vowels, although it cannot be regarded as a long, tense vowel since it can never be found in open stressed syllables (Kreidler, 2004: p. 50; Ladefoged & Johnson K., 2011: p. 98-99). Ladefoged and Johnson K. (2011) list six diphthongs for British pronunciation: /aɪ/, /aʊ/, /eɪ/, /əʊ/, /ɔɪ/, /ju/[21] with /oʊ/ instead of /əʊ/ for American (p. 90). Cruttenden (2014), on the other hand, does not treat /ju/ as a diphthong but includes /ɪə/ and /ʊə/[22] (p. 96).

[16] Despite the literature standard to use head words to represent English vowels (Wells, 1982: p. 120), we opted for the phonetic symbols for the sake of comparability with the vowels n Serbian.

[17] The vowel may be slightly rounded in rhotic dialects (Kreidler, 2004: p. 55)

[18] Even though vowels /uː/ and /ʊ/ are often described as high back vowels, native speakers tend to centralise them, which is confirmed by numerous studies (see Bjelaković, 2018; Kleber et al., 2011; Sóskuthy et al., 2015) as well as the vowel diagram in Figure 4.2. The centralisation, or rather, fronting of back vowels seems not to be limited to British English or even to native speakers. For instance, Havenhill (2019) has shown that in some varieties of American English fronting is observed for /uː/, /ʊ/ and /ɔ/ alike, whereby the back vowels tend to differ from the front ones by lip rounding. Similarly, Valenzuela & French (2023) found that Spanish learners of English are also suscepttible to the so-called "push effect" as they gravitate to accommodate their pronunciation towards present-day native speakers (p. 10).

[19] The listed vowels are found in the following words in order of appearance: BEAD, BID, BED, BAD, BIRD, BUD, BOOED, BOOK, BOARD, BOD, BARD. The transcription in the relevant literature may differ depending on the author and edition.

[20] The vowel of the word PAIR.

[21] The listed diphthongs are found in the following words in order of appearance: HIGH, HOW, HAY, HOE, BOY, CUE.

[22] The diphthongs are found in the words PEER and POOR.

The diagram in Figure 4-1 below, drawn from the data presented by Tomić & Milenković (2019: p. 159) and Bjelaković (2018: p. 186-192), compares Serbian and English vowels as pronounced by female native speakers of Serbian from Prizren-Timok dialectal region and Newscasters speaking Standard British English.[23] The diagram in Figure 4-2 depicts the comparative vowel space of the same Serbian speakers and General American English as described by Hillenbrand et al. (1995).

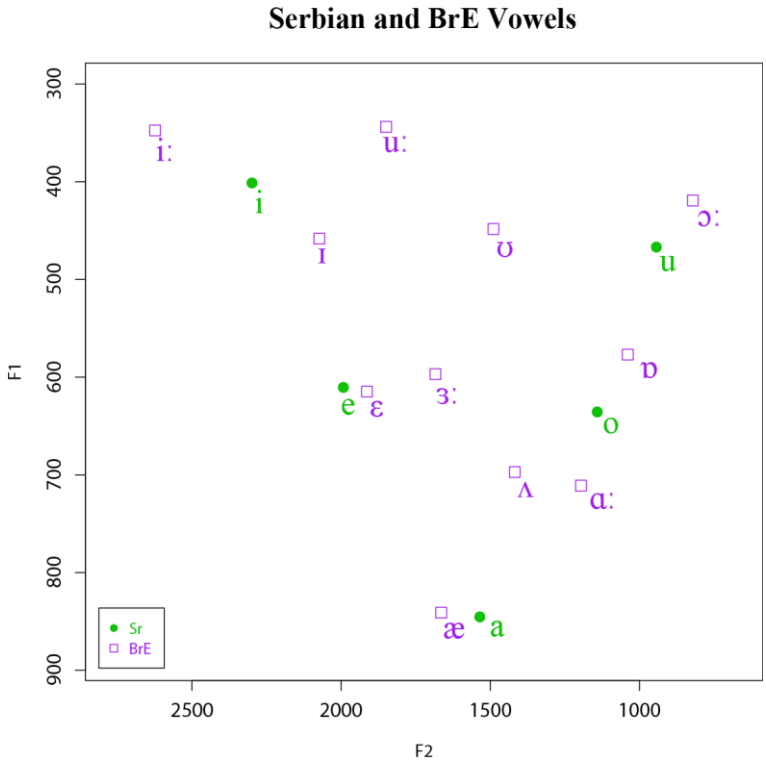


Figure 4-1
Comparative vowel space of Serbian and British English speakers

[23] The diagrams in Figures 4-1 and 4-2 were constructed using NORM, The Vowel Normalization and Plotting Suite. http://lingtools.uoregon.edu/norm/norm1.php. Vowels were plotted according to their F1 and F2 values without applying any normalisation. The data was adopted from Tomić and Milenković (2019), Bjelaković (2018) and Hillenbrand et al. (1995) as these studies offer F1 and F2 frequency values for the relevant Serbian and English vowels pronounced by female speakers.

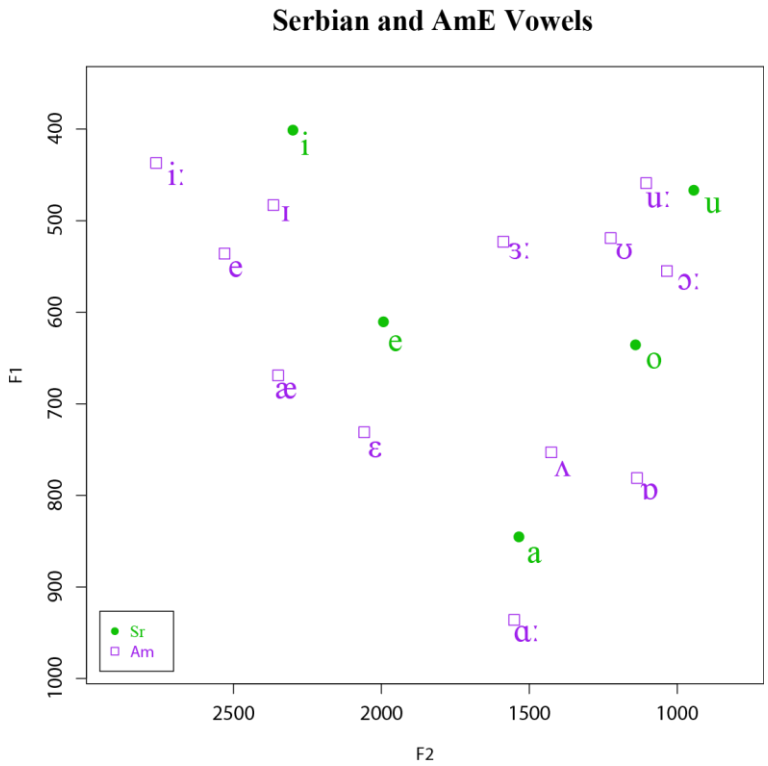


Figure 4-2

Comparative vowel space of Serbian and American English speakers

In light of the voice quality theory, it can be concluded that both varieties of English have more peripheral vowels than Serbian; therefore, this language can be expected to have an overall wider tongue range. It is not surprising, considering that Serbian has five cardinal vowels, and the distance between them is large enough; thus, the speakers need not assume extreme articulatory movements to ensure intelligibility. On the other hand, speakers of English push their tense vowels toward the periphery of the vowel space because the language distinguishes between multiple vowel categories. The peripheral positions of high and low vowels in English predict a higher degree of openness of the jaw. Another observation that can be made is the fronted pronunciation of American vowels. The centralisation of the back vowels and fronting of the entire vowel space is even more prominent in recent research by Nikolić (2016), who described the speech of two American scholars. Nikolić (2016) comments that his results "demonstrate quite an unorthodox 'image' of the vowels produced by the American participants" (p. 97). However, due to the lack of recent large-scale studies of American English vowel formants, it is difficult to confirm whether the participants are outliers or represent the language change that is bound to have occurred since Hillenbrand et al. (1995). Furthermore, while rounding is not a distinctive feature in either of the languages, it seems to be more prominent in Serbian back vowels. Finally, neither Serbian nor English vowels are followed by distinctive phonatory features or nasality.

These are some general comparisons between Serbian and English vowel systems. In section 4.3. we will explore the acquisition of the English language by Serbian speakers, and only then will it be possible to make more specific predictions regarding voice quality behaviour across these two languages.

4.2. Production of English Vowels by Serbian Speakers

Studies of English pronunciation by native speakers of Serbian from Novi Sad (Šumadija-Vojvodina dialect) indicate that vowel categories of BED and BAD share the vowel space to a great extent (Marković, 2009b: p. 261; Marković & Jakovljević, 2016: p. 222) and the contrast between the two vowels is achieved primarily by the duration component (Marković, 2009b: p. 260). Regarding the close front vowels, the vowel in the word BEAD, produced by Serbian speakers, corresponds to Serbian /i/ in syllables under long pitch accents (Marković, 2009a: p. 7,9). In contrast, the vowel of the word BID is more centralised (p. 7). In addition, Marković (2009a) remarks that the realisation of the vowel in BOOED is "halfway" between Serbian /u/ and the realisation by native speakers, while the lax vowel (BOOK) seems to be acquired better (Marković, 2009a: p. 15).

Similar conclusions were reached by Bjelaković (2018), who analysed the production of English vowels by native speakers of Serbian from Belgrade (Šumadija-Vojvodina dialect). He confirms that the contrast between BED and BAD is achieved primarily through the quantitative component and that the vowel of BED corresponds to the vowel space for /e/ under short pitch accents in Serbian (p. 161). Unlike the speakers in Marković (2009a), the speakers in Bjelaković (2018) do not pronounce the vowel of the word BEAD in the same vowel space as Serbian /i/; instead, it is more peripheral, and, therefore, almost identical to the vowel pronounced by English control speakers. In addition, the vowel of the word BID is centralised as expected. Bjelaković (2018) notes that the vowel in BUD corresponds to the target vowel of native speakers; however, the vowel in BARD appears to be "a compromise" between the expected values and vowel /a/ in Serbian (p. 132). All speakers in Bjelaković (2018) distinguish between the vowel categories of BOARD and BOD; however, the production does not correspond entirely to that of the native speakers. In addition, for a small group of participants, the pronunciation of the vowel in LOT corresponds to /o/ in Serbian (p. 136). Furthermore, both vowels in BOOT and BOOED are pronounced more central than Serbian /u/, which is in accordance with the pronunciation of native speakers. However, the contrast between the tense and lax vowel categories is not achieved by all speakers (p. 139). Similar to Marković (2009a), Bjelaković (2018) notes that the pronunciation of BOOED by Serbian

speakers appears to be “a compromise” between the native and target category. Finally, for some speakers, the vowel in the word BIRD shares the vowel space with Serbian /e/ as pronounced in the syllables under short pitch accents (p. 144).

Paunović (2011) studied the pronunciation of English vowels by Serbian speakers from Niš (Prizren-Timok dialect). She remarks that the vowel of the word BAD is the most open of all and concludes that it is assimilated in the vowel category of /a/ that exists in the participants’ mother tongue. In addition, she notices that vowels in BUD, BARD and BOD partly correspond to the category of /o/ in Serbian.

Tomić and Milenković (2019) compared the pronunciation of English by two groups of Serbian speakers from different dialectological backgrounds (Novi Sad and Niš). According to their results, and in contrast with previous research (Bjelaković, 2018; Marković, 2009b), both groups of speakers distinguish between vowel categories of BED and BAD, but the speakers from Novi Sad tend to pronounce the vowel in BAD as significantly more open (similarly as described in Paunović, 2011). On the other hand, neither group distinguishes between the vowel categories of BUD/BARD and BID/BEAD, whereas the distinction between BOOK and BOOED seems to be achieved only by speakers from Novi Sad.

Based on the previous research, it is difficult to draw general conclusions regarding English vowel acquisition by native speakers of Serbian. The differences in the obtained results most likely stem from varying proficiency levels of the participants or the elicitation tasks employed during the experiment (text reading, semi-spontaneous and spontaneous speech, isolated words). None of the studies above has reported any information on the participants’ exposure to the foreign language nor engaged native speakers to assess whether the target realisation of individual vowels was achieved. Nonetheless, reviewing the results, we can group vowel acquisition into three outcomes:

(1) The foreign language vowel category is not acquired, and the vowel is assimilated to a similar vowel category that already exists in the mother tongue
(2) The foreign language vowel category is not fully acquired, and vowel formant values are in between the target vowel and a similar category that exists in the mother tongue
(3) The foreign language vowel category is fully acquired.

Correspondingly, it can be concluded that, for those speakers who have more assimilated than acquired vowel categories, voice quality features in the foreign language are likely to remain the same (both perceptually and acoustically) as when speaking the mother tongue. Conversely, the speakers who have acquired more target vowel categories are likely to exhibit more significant within-speaker variability across languages. The third and probably the

most comprehensive group of speakers will be those whose vowel space resembles neither native Serbian nor native English. Therefore, to determine how language proficiency and pronunciation affect cross-language forensic speaker comparison based on voice quality parameters, all participants will be assessed according to the standardised test by trained judges.

5. The Present Study

5.1. Research Questions Revisited

As was established in Chapter 1.3., the present study aims to explore how individual voice quality changes depending on the language spoken and the implications of the potential variation for forensic speaker comparison. More precisely, the perceptual experiments aim to investigate the following questions:

(1) Are the same phonatory and articulatory settings equally audible when someone is speaking Serbian (L1) and English (L2)?

(2) How similar are the voices of the same/different speakers when speaking Serbian (L1) and English (L2) as perceived by naïve listeners?

(3) What is the relationship between objectively scored voice quality and voice similarity perceived by naïve listeners?

(4) What is the discriminatory power of voice quality compared to naïve listeners' performance in the cross-language context?

(5) What is the relationship between speakers' foreign language proficiency and objectively scored voice quality, that is, voice similarity as perceived by naïve listeners?

Answering the questions above will help us understand how voice quality is related to voice perception in the cross-language setting, whether there is a "language effect", and its magnitude.

The second part of the research concerns the analysis of the acoustic correlates of voice quality and testing the reliability of the selected parameters in cross-language forensic speaker comparison. The analysis will aim to answer the following questions:

(6) How do the acoustic correlates of voice quality compare across languages?

(7) How reliable are the acoustic correlates of voice quality in the FSC in Serbian (L1)?

(8) How reliable are the acoustic correlates of voice quality in the cross-language FSC (Serbian as L1 and English as L2)?

(9) Which acoustic measures are more robust in the cross-language FSC, the ones pertaining to the articulatory or the phonatory settings?

(10) Which reference population is the most suitable for cross-language FSC (L1, L2, or L1+L2)?

(11) Is a speaker's foreign language proficiency relevant to their individual performance within the FSC system?

By shedding light on the questions above, we will understand how the acoustics of an individual's voice quality changes when they speak a foreign language and, more practically, how these changes affect the forensic speaker comparison in cases of language mismatch.

5.2. Corpus Development

5.2.1. Participants – all speakers

The corpus for the present research consists of the telephone-recorded spontaneous speech of 50 native speakers of Serbian with varying English proficiency. The speakers were selected following the criteria for sex (female), age (between 18 and 27 years old, mean 21.9, SD 2.435), dialectological background (Prizren-Timok dialectological region), education (students and graduates) and the age of exposure to the foreign language in question (before age 11). In addition, the speakers who participated in the corpus creation were not students at the Department of English Language and Literature, where the English language is studied at a professional level and where students undergo intensive training regarding English pronunciation in Phonetics and Phonology courses.

In order to understand the speakers' exposure to English and their reliance on this language in everyday life outside the scope of formal education, they were asked to rate how often they use English in various circumstances. Table 5-1 summarises the self-reported scores of all the participants:

Table 5-1
Self-reported English language exposure scores by the recorded speakers

Context	***Mean***	***SD***	***Median***
Reading books	2.22	1.055	2
Reading newspaper articles and blog posts	3.5	1.329	3.5
Engaging with social network content	4.3	0.953	5
Listening to podcasts or videos	3.98	1.134	4
Listening to music	4.32	1.019	5
Watching movies/TV series without subtitles	3.66	1.222	4
Watching movies/TV series with English subtitles	3.78	1.183	4
Speaking	3.12	1.136	3
Chat via social networks	3.04	1.228	3
Writing	3.08	1.243	3
Overall exposure – average score	3.5	0.809	3.6

As it can be inferred from Table 5-1, the average English language exposure score gravitates toward 3.5 on a 5-point scale (1 - never, 2 - seldom, 3 - sometimes, 4 - often, 5 - very often). The lowest value is observed for book reading in English while engaging with social media content and listening to music have the highest values and lowest variability among the speakers. The participants in the present study can be described as using English moderately in their daily lives, the scores for active English use (speaking and writing) being on average lower than for passive exposure (listening to and watching different content).

5.2.2. Participants – narrow set

A narrow set of 20 participants was assembled from the corpus for the two listening experiments. In order to minimise the potential effects of accents on voice similarity scoring, we only included speakers born and raised in Niš, the largest urban centre of the Prizren-Timok dialectological area. The mean age of the narrow-set participants is 22.15, with SD = 2.477. The average score of their self-reported exposure to English on a 5-point scale is 3.45, with SD = 0.876 and a Median of 3.55.

5.2.3. Recording procedure

The recording was performed via the mobile phone on the far end of the speaker over the 4G-LTE network using the standard recording API implemented in the Call Recorder application with root-access requirements – skvalex v. 3.4.9 (Skvortsov, 2021). The recording device model LG G2 D802 runs on Android v. 11 – LineageOS. In order to avoid speech overlap and potential interference of the noise on the interviewer's end, only the downlink audio source was recorded. The files were stored in the wav format with the sampling rate of 44100 Hz and the bit-depth of 32 bits.

While the described recording setting has some limitations, considering that it does not control for the devices and network setup on the speakers' end, it provides nearly-realistic conditions encountered in actual forensic casework, where it is impossible to control such variables.

The speakers were instructed to stay indoors during the interview and limit potential background noise sources. In addition, they were asked to turn off the notifications on their devices and use the device in the same manner in terms of proximity and input method (earphones or device microphone) throughout the procedure. When the desired network quality was not established or deteriorated throughout the call, the connection was cancelled and re-established until satisfactory quality was obtained. Before the interview, the speakers were

asked to rate the quality of the signal on a 5-point scale, the average score for the present corpus being 4.62 with an SD of 0.49.[24]

5.2.4. Materials

Spontaneous speech for the corpus in Serbian and English was elicited through an interview modelled after the IELTS speaking exam[25] (Cullen et al., 2014). The interview consists of three parts:

(1) the speakers talk about personal life on familiar topics,

(2) the speakers are given a minute to prepare a short monologue on the given subject, describing an object or a place,

(3) the speakers are challenged to employ critical thinking and discuss more abstract topics which require a broader vocabulary repository.

Such a framework was chosen because not only does it enable structured conversation comparable across speakers, but it also permits the participants to use the foreign language to the extent that they feel comfortable. Finally, the spoken material gathered in this way can be scored for proficiency using the established criteria by IELTS. The discussion topics for the interviews can be found in Appendix 1.

Prior to the interview, the speakers were instructed to avoid single-word responses and try to provide longer answers, explaining their choice and giving some examples, details or reasons for the answer. Furthermore, in case they cannot remember a specific vocabulary, the speakers were encouraged to rephrase the sentence and explain what they meant in another way. If they thought they had said something incorrectly or used a wrong word, the speakers were encouraged not to dwell on the mistake but to rephrase their sentence or repeat how they felt was correct. The goal was to obtain 7 to 10 minutes of spontaneous speech in each language. The length of an interview with a single speaker, including the preparatory stage and giving instructions, was around 30 minutes.

[24] The corpus of 50 participants in the present study was selected from a broader group of 60 recorded speakers based on the optimal sound quality.

[25] IELTS speaking test description: https://takeielts.britishcouncil.org/take-ielts/prepare/free-ielts-practice-tests/speaking

5.3. Language Proficiency Scoring

5.3.1. Scoring procedure

For the purposes of the present study, language proficiency was estimated via two methods: a mock level-placement test[26] by British Council and proficiency scoring by experts following IELTS speaking criteria. The test is approximate; it is designed to assess the candidate's competence in English grammar, vocabulary and phrasing through 25 multiple-choice questions, each followed by evaluating the candidate's confidence regarding their answer on a 3-point scale (1-not sure, 3-certain). The participants were asked to acknowledge that they had completed the test alone, without consulting dictionaries, textbooks, or relying on the internet search and similar sources, and that they were aware the results were anonymous and would not be shared with anyone. In order for us to ensure that the submitted results were relevant, the level-placement test was distributed via Google Forms and the answers, including confidence values, were then entered into the original form by the researcher to obtain final scores.

Oral proficiency scoring for the present study was performed following IELTS speaking task assessment criteria (see Chapter 5.3.2.) by five ESL teachers with more than six years of teaching experience (including the author). The experiment was distributed via the custom-developed online platform (Appendix 2), where the experts could hear three recordings corresponding to the three parts of the interview for each speaker and grade the relevant parameters on a 9-point Likert scale. Hovering the mouse cursor over each point on the radio buttons would provide a description of what the candidate should have mastered for the given band score. So that we would ensure maximum quality, the raters were given a 45-minute onboarding concerning the IELTS exam and assessment criteria. In addition, the experiment contained a mandatory training session with five speakers so that the raters would become acquainted with the procedure and ask for additional instructions if needed.

In order to guarantee objectivity in proficiency grading, most of the recorded material was included in the experiment. The pauses, hesitations and occasional code-switching were not removed from the corpus. The total amount of graded material is 367.8 minutes, with an average duration of 6.97 minutes per speaker (SD = 1.3). Considering the cumbersome and time-consuming nature of the task, only the experts whose final scores exhibited a statistically

[26] Online English Level Test (British Council, 2023) https://learnenglish.britishcouncil.org/english-levels/online-english-level-test

significant correlation with the mock-test scores for the first 20 speakers (the narrow set) were asked to continue the experiment, which resulted in keeping three proficiency raters with the inter-rater agreement (Gwet's AC2) of 0.9103 and standard error of 0.013. Their grades were averaged for each scored parameter, and new average final bands were obtained by rounding the average score to the nearest .5 value. Pearson correlation between the obtained bands and the level placement test is r = 0.591, p-value = 0.006 for the narrow set, and r = 0.581, p-value < 0.00001 for the entire corpus.

5.3.2. IELTS speaking – assessment criteria

The IELTS exam utilises a 9-point band scoring system that roughly corresponds to the CEFR in the following way: band 4 – strong A2, bands 4.5 and 5 – B1, bands 5.5, 6 and 6.5 – B2, bands 7, 7.5 and 8 – C1, band 9 – strong C1 or C2 (IELTS, 2023a).

The IELTS speaking task is graded according to four criteria: fluency and coherence (F/C), lexical resource (LEX), grammatical range and accuracy (GR/A), and pronunciation (PRON); the final score is the average of the four (IELTS, 2023b). *Fluency and coherence* refers to the ability to talk continuously without notable effort and to connect ideas into coherent speech. Key indicators of fluency are speech rate and speech continuity (ideally without false starts, backtracking, hesitations, or functionless repetitions), whereas coherence is reflected in logical sequencing, appropriate usage of pauses and discourse markers, relevance, and appropriate use of cohesive devices (IELTS, 2023b). *Lexical resource* refers to the range of vocabulary the candidate can apply, affecting the range of topics they can discuss and the precision of meaning they can convey. Key indicators of lexical resource are the variety of words, their appropriateness in terms of style, collocation or referential value, and the ability to paraphrase the utterance in case of missing the appropriate vocabulary (IELTS, 2023b). *Grammatical range and accuracy* reflect the range of complexity of grammatical resources and, therefore, propositions the candidate is able to express, as well as the appropriateness and accuracy of the used grammar. Key indicators are the length of the sentences, subordination, the complexity of phrases (pre- or post-modification, verb phrase complexity), error density and the effect of errors on intelligibility (IELTS, 2023b). Finally, *pronunciation* refers to the "accurate and sustained use of phonological features to convey meaningful messages" (IELTS, 2023b: p. 3). Key indicators of appropriate pronunciation are the accurate reproduction of phonemes, employment of connected speech features, word stress, sentence stress, rhythm and intonation, as well as the overall effect of the accent on intelligibility.

The band scores for the described criteria were assigned after the instructions on the IELTS website (IELTS, 2023c), reproduced in Appendix 3.

5.3.3. Proficiency scores and instrument validity

According to the level placement test, 20 participants were classified as upper-intermediate or above language users (test scores above 80%) and 40 as intermediate (test scores between 55% and 80%). Table 5-2 summarises the test scores and confidence values.

Table 5-2
Average level-placement test scores and confidence values

Participants	***Average test score***	***SD***	***Average confidence value***	***SD***
All	75.66%	8.6	2.42	0.29
Upper-intermediate	83.8%	3.4	2.64	0.18
Intermediate	70.23%	6.4	2.27	0.27

IELTS-based evaluation has yielded 35 independent and 15 proficient language users. For the purposes of the present research, the former will be considered to roughly correspond to the intermediate while the latter to the upper-intermediate level. The average scores for individual criteria are summarised in Table 5-3.

Table 5-3
Average IELTS-based proficiency scores

Participants	***F/C***	***LEX***	***GR/A***	***PRON***	***Av_band***	***SD_band***
All	6.48	6.1	5.94	5.8	6.22	0.97
Independent	6	5.63	5.43	5.29	5.71	0.65
Proficient	7.6	7.2	7.13	7	7.4	0.42

Cohen's kappa agreement for the level placement between the test and IELTS-based assessment is 0.435 (significance 0.002), rendering both instruments relatively stable and comparable. Pearson correlation for averaged IELTS scores and test percentage is presented in Table 5-4 below.

Table 5-4
Correlation between the level-placement test and IELTS-based scores - all participants

		F/C	***LEX***	***GR/A***	***PRON***	***Band***
Test score	Pearson r	.522	.584	.596	.445	.581
	Sig (2-tailed)	.0001	.0000	.0000	.001	.0000

In the narrow set, six of the 20 speakers were classified as upper-intermediate (proficient users) and fourteen as intermediate (independent users) by both the level-placement

test and IELTS-based scores. However, since only 3 of them were recognised as upper-intermediate users by both instruments, Cohen's kappa was calculated to estimate the reliability of the scores, obtaining the agreement of .286 and approximate significance of .201. Such results are not statistically significant, so the Pearson correlation was applied to observe whether there is a general trend in speaker ranking. The correlation values are presented in Table 5-5 below.

Table 5-5
Correlation between the level-placement test and IELTS-based scores - narrow set

		F/C	*LEX*	*GR/A*	*PRON*	*Band*
Test score	Pearson r	.519	.491	.642	.532	.591
	Sig (2-tailed)	.019	.028	.002	.016	.006

As it can be inferred from Table 5-5, averaged values across each of the scored criteria exhibit a statistically significant correlation with level-placement test results. The strongest correlation is observed for grammatical range and accuracy, which is reasonable, considering that the test predominantly focuses on assessing grammar.

Considering that the level-placement and IELTS speaking tests do not focus on grading the same language aspects, slight disagreement in final scores is acceptable. Finally, both instruments were chosen to obtain as comprehensive a perspective on the English language proficiency of the participants as possible.

5.4. Voice Quality Scoring

5.4.1. Expert listeners

Within phonetic sciences, voice quality is often described as "obscure" (Hewlett & Beck, 2006: p. viii) and is "not considered part of traditional auditory-phonetic training"; therefore, worldwide, there are not many phoneticians proficient in this approach (Rose, 2002: p. 289). For the present study, four expert listeners volunteered to score 40 samples by 20 speakers (the narrow data set) across two languages. The experts are experienced phoneticians who have studied voice quality for academic purposes and have substantial experience with the Vocal Profile Analysis Scheme. The phoneticians have reported different mother tongues: English, German, and Czech. Moreover, they all reported strong English competence, while only one declared slight familiarity with Serbian phonology. In addition, when asked whether they speak any Slavic languages, two of the participants reported familiarity with Polish and Russian, respectively. While an ideal choice of participants would include experts equally

familiar with both Serbian and English, the present selection is valid because, as previous studies have shown, expert familiarity with and training in the use of Vocal Profile Analysis are of greater importance to the quality results than the listeners' language background (San Segundo et al., 2019). In addition, Laver (1980) underlines that the VPA is an objective tool that can be applied to the analysis of voice "on absolute grounds, not grounds relative to the accent of the speaker's speech community" (p. 88); therefore, the selected experts are considered competent to perform the analysis.

5.4.2. Truncated VPA protocol

The present research utilises a modified VPA protocol modelled after Laver et al. (1981). It is a truncated, 28-feature version of the original protocol employed in forensic speaker comparison casework by JP French Associates Acoustic Laboratory, also recently used by San Segundo et al. (2019) to propose a VPA-based methodological framework for forensic speaker characterisation. The list of analysed articulatory and phonatory features is available in Appendix 4.

The protocol employed in the present research describes the features in three scalar degrees as follows:

(1) slight – although you are confident that the setting is audible, it is not prominent. It is hearable but might be missed if you were not specifically listening for it;

(2) marked - the setting is easily noticeable and is a distinguishing feature of the voice;

(3) extreme - the setting is highly prominent. The degree of prominence is unusual and verging on abnormal.

For neutral settings, the experts were advised to leave the fields blank.

Since the movements of lips and tongue are greatly affected by the position of the jaw (Esling et al., 2019: p. 27; Wrench & Beck, 2022: p. 25), and the "jaw can be seen to move in sympathy with the articulations of the body of the tongue" (Laver, 1980: p. 63), the labial and mandibular range settings are not included in the truncated version of the protocol used in the present study. In addition, jaw protrusion was not considered as it correlates with a close jaw and lip protrusion setting (Esling et al., 2019: p. 25). Next, considering that audible nasal escape is not a phonetic feature of any known accent and is considered pathological, it was not expected to be encountered in the present corpus; thus, it was not included in the protocol. Furthermore, since all of the settings involving pharyngeal constriction are attributed to retraction of either the body or the root of the tongue, or both, tongue-root and pharyngeal

settings are, by convention, seldom included in the same descriptive protocol (see Laver, 1994: p. 412). As with the tongue-root and pharyngeal constriction, pharyngeal expansion and lowered larynx settings are not separately listed in the present experiment. Finally, bearing in mind that the present research relies on a 3-scalar-degree scheme instead of the original scheme that utilises six scalar degrees, a more extreme tongue tip retraction known as retroflexion is listed as an additional setting.

Phonation features in the present research are observed both in isolation (creak, whisper) and in combination with voice (creaky, whispery). As explained in Chapter 3.1.5., overall muscular tension and prosodic features are not considered in the present study.

5.4.3. Scoring procedure

The experiment was distributed online via a custom-made platform (see Appendix 5). The expert listeners were asked to provide background information regarding their familiarity with the Serbian and English language, including the phonological system, as well as to elaborate on their previous experience with the VPA scheme. The experts were instructed to identify settings that run through the speech chain that are present all, or at least most, of the time. They were discouraged from marking settings that are only occasionally heard, for example, if the speaker uses creaky voice, but only briefly and at the end of utterances when speaking on a low pitch. The recordings could be replayed as many times as needed.

Twenty speakers from the corpus (the narrow set) were selected for voice quality analysis in Serbian and English. Per recommendations from previous studies (see Mackenzie Beck, 1988: p. 144), the recordings were 40 seconds long, which amounted to 26 minutes of listening material. The listeners were first presented with the recordings in the foreign language (English) and then with the recordings in native Serbian, but in a mixed order.

Prior to the analysis, the settings that form a continuum (e.g. fronted-backed tongue body, raised-lowered larynx) were transformed into a single range of seven scalar degrees (cf. Laver et al., 1981), where neutral was graded as 4. For the rest of the features (e.g. labiodentalisation, retroflexion), neutral was marked as 1, slight as 2, marked as 3 and extreme as 4. Modal voice, falsetto, whisper and creak had only two values. Even though the previous studies have used Kappa statistics for determining inter-rater reliability, the measure is considered inappropriate in the present research due to the Kappa paradox and the assumptions this analysis requires to be met. Namely, it is common for Kappa statistics to yield lower values than simple percentage agreement calculation, the phenomenon known as the Kappa paradox. Another consequence of the paradox is that it is impossible to know the number of subjects

required to obtain a standard error below 0.5 (see Gwet, 2021). In addition, Fleiss' Kappa, often employed in comparing the ratings by more than two raters, assumes that the categories of the response variable are mutually exclusive; however, the VPA scalar degrees are part of a continuum. Furthermore, Fleiss' Kappa assumes that each variable has the same number of categories, which is not the case in our data. Finally, this method assumes non-unique raters, while our study employs the same raters for each sample (see Fleiss, 1971; Fleiss et al., 2003; Laerd Statistics, 2023). For these reasons, the statistical measure employed to assess the inter-rater reliability is Gwet's AC2 (Gwet, 2021). Gwet statistics was calculated in RStudio using "irrCAC" package (Gwet, 2019).

5.4.4. Inter-rater reliability and instrument validity

Gwet's AC2 statistics can treat the dataset as either categorical (also Gwet's AC1) or interval. Table 5-6 lists the reliability scores between all expert listeners and between every two pairs of listeners, respectively.

Table 5-6
Inter-rater reliability for VPA scores

	All experts	***Min***	***Max***	***Phonatory***	***Articulatory***
Gwet's AC1	.587	.510	.700	.648	.547
s.e.	.014	.021	.019	.020	.019
Gwet's AC2	.940	.916	.960	.924	.932
s.e.	.004	.007	.004	.007	.005

The inter-rater reliability is higher if the data is treated as interval. Notwithstanding, both Gwet's AC1 and AC2 yield satisfactory results and low standard error. The agreement seems to be stronger for phonatory features if the data is treated as categorical: however, if it is observed as interval, the agreement for the articulatory features is stronger. Minimal observed agreement between two raters is .510, whereas the maximum obtained agreement for two raters is 0.7. The "true" value for each scored setting will be obtained by calculating the median of the four scores. The assigned VPA scores are interpreted as correct and valid, considering the moderate interrater agreement.

The procedures outlined above and the results obtained therein constitute part of the methodology of multiple experiments in the thesis. They will be relevant both to the listening experiments presented in Chapter 6 and the acoustic analysis and likelihood ratio calculations in Chapter 7.

6. Part 1 – Perceptual Experiments

6.1. Experiment 1 – Vocal Profile Analysis

6.1.1. Study design

The primary goal of this experiment is to objectively assess voice similarity across languages by relying on an established protocol and trained judges. The underlying aim is to examine the usability of the Vocal Profile Analysis protocol in cross-language speaker comparison on the example of native Serbian and foreign English. The recordings of twenty female native speakers of Serbian (the narrow set, see Chapter 5.2.2.) were subject to Vocal Profile Analysis by four phonetic experts. The procedure for selecting expert listeners, VPA settings, rating scales, as well as the distribution of the experiment itself and instrument validity are explained in Chapter 5.4.

Speakers' vocal profiles are compared quantitatively by calculating Euclidean distances (d) and Cosine similarity (S_c), two standard measures used for comparing vector variables, whereas qualitative analysis was performed by observing the number and percentage of non-neutral settings that were noted in either language and kept constant across languages. Speaker discriminatory value of the VPA protocol in cross-language speaker comparison was established by identifying close matches (with varying thresholds) in same-speaker (SS) and different-speaker (DS) pairs (cf. French et al., 2015). Quantitative analysis in the listening experiments was performed by using Microsoft Excel 2016 and IBM SPSS Statistics v. 26.

6.1.2. Results

Quantitative analysis

Average between-speaker distances and similarities are calculated for each speaker in Serbian and English, respectively, while within-speaker distances/similarities are observed across languages. Statistical analysis has confirmed that within-speaker distances across languages are lower than average between-speaker distances, both in Serbian and English. Consequently, within speaker similarity across languages is higher than average between-speaker similarities in either language. In addition, paired t-test comparisons have not revealed any differences in average voice distances or similarities in the foreign language and the mother tongue. The distance and similarity values are presented in Table 6-1, whereas Table 6-2 provides the results of the statistical analyses.

Table 6-1
Average between-speaker and within-speaker Euclidean distances and Cosine similarities

	Euclidean distance			***Cosine similarity***		
	Between speakers		***Within***	***Between speakers***		***Within***
	Serbian	*English*	*Sr-En*	*Serbian*	*English*	*Sr-En*
Mean	1.732	1.808	1.541	.988	.989	.992
Variance	0.039	0.086	0.199	.000	.000	.000
Min	1.543	1.461	1	.982	.980	.980
Max	2.172	2.504	2.345	.991	.993	.997

Table 6-2
Paired t-tests and Pearson correlation of the compared distances and similarities (two-tailed)

	Euclidean distance			***Cosine similarity***		
	Between-Between (Sr-En)	*Within-Between (Sr)*	*Within-Between (En)*	*Between-Between (Sr-En)*	*Within-Between (Sr)*	*Within-Between (En)*
t-test	-1.150	-2.198	-3.544	0.161	2.565	3.161
p-value	.264	.041	.002	.874	.019	.005
Pearson r	.338	.488	.654	.346	.401	.624

Furthermore, we performed Pearson correlation to examine whether the speakers with the largest average distances in Serbian also stand out in the foreign language; however, the hypothesis was disproved. As a matter of fact, the results indicate that the more similar a speaker is rated across languages, the more similar that speaker is to the others in the dataset. Figure 6-1 displays the results of multidimensional scaling based on square Euclidean distances. As it can be inferred from the graphs, the speakers who are rated as most distinct in Serbian are not necessarily rated as most distinct when they speak English.

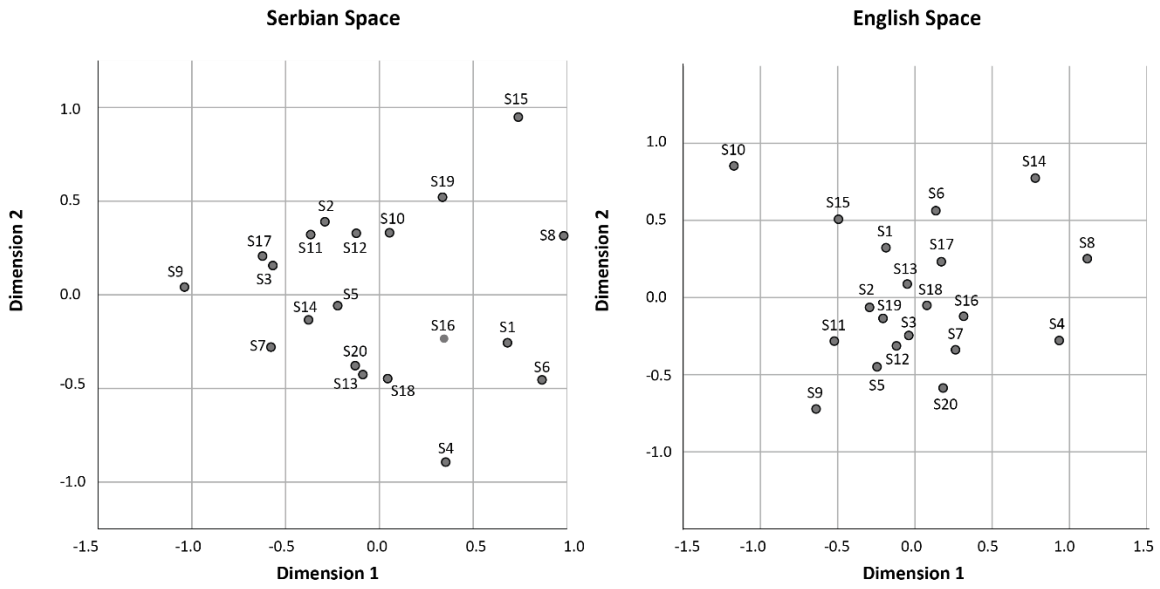


Figure 6-1

Multidimensional scaling based on squared Euclidean distances

One of the hypotheses of this study was that voice quality in a foreign language may depend on the speaker's proficiency in that language. To investigate the hypothesis, we correlated within-speaker distances and similarities with language proficiency scores; however, no correlation was found. The speakers were then grouped into categories (intermediate/upper-intermediate) and the intra-speaker distances and similarities were compared with Welch's t-test. A slight difference in cosine similarity was detected for the two groups of speakers categorised according to the IELTS criteria ($t = -1.902$, $p = .074$). Namely, it can be said that the intermediate level speakers exhibit lower within-speaker similarity across languages than upper-intermediate speakers, with the confidence interval of 90%. Such a finding is on the very opposite end of our initial hypothesis, in which we assumed that higher proficiency would result in lower within-speaker similarity. However, due to weak statistical evidence, we will not embrace a firm stand on the finding. Weak statistical evidence is most likely the result of a small number of participants who constitute a rather homogenous group of foreign language learners/speakers with the same sociolectal background, whose both vocal profile and language proficiency scores are very similar to begin with.

Considering that roughly a third of the VPA settings concern the phonatory features, less likely to be affected by language proficiency, we performed the same analyses as above on articulatory and phonatory settings, respectively. Average Euclidean distances and Cosine similarities for articulatory settings are summarised in Table 6-3, whereas the results based on the phonatory features are presented in Table 6-5.

Table 6-3
Average between-speaker and within-speaker Euclidean distances and Cosine similarities – articulatory settings

	Euclidean distance			***Cosine similarity***		
	Between speakers		***Within***	***Between speakers***		***Within***
	Serbian	*English*	*Sr-En*	*Serbian*	*English*	*Sr-En*
Mean	1.325	1.306	1.227	.9928	.9937	.9942
Variance	0.059	0.082	0.190	.0000	.0000	.0000
Min	1.055	0.967	0.500	.9851	.9873	.9848
Max	1.921	1.957	1.936	.9958	.9964	.9992

Compared to the values obtained from the entire Vocal Profile Analysis, both within- and between-speaker distances based solely on the articulatory settings have reduced significantly. Comparatively, the similarities have increased (t-test results available in Table 6-4). Strong correlation results indicate that the speakers who are most distant from others based on the entire VPA also tend to be most distant in the articulatory features. Likewise, those with the highest within-speaker similarity across languages remain so even after the phonatory features are removed from the protocol.

Table 6-4
Comparison of Euclidean distances and Cosine similarities when assessed from the entire VPA and articulatory settings in isolation

	Euclidean distance			***Cosine similarity***		
	Between-speaker (Serbian)	*Between-speaker (English)*	*Within-speaker*	*Between-speaker (Serbian)*	*Between-speaker (English)*	*Within-speaker*
t-test	20.903	12.393	7.753	-20.168	-11.022	-5.603
p-value	.000	.000	.000	.000	.000	.000
Pearson r	.944	.806	.917	.946	.829	.937

Furthermore, as opposed to the results obtained with the entire VPA, within-speaker distances no longer exhibit statistical significance in comparison to average distances between speakers. There is still some evidence that within-speaker (cosine) similarity is higher than average between-speaker similarity in the mother tongue ($t = 1.771$, $p = .093$); however, the confidence interval is much lower. Such results indicate that the articulatory features of voice quality have a less significant impact on speaker-specificity in cross-language comparison. Finally, as with the entire protocol, the correlation statistics have not revealed any association between the within-speaker distances/similarities and proficiency scores.

Table 6-5
Average between-speaker and within-speaker Euclidean distances and Cosine similarities – phonatory settings

	Euclidean distance			***Cosine similarity***		
	Between speakers		***Within***	***Between speakers***		***Within***
	Serbian	*English*	*Sr-En*	*Serbian*	*English*	*Sr-En*
Mean	1.032	1.168	0.867	.9499	.9568	.9728
Variance	0.010	0.064	0.129	.0000	.0003	.0003
Min	0.925	0.856	0	.9227	.9159	.9418
Max	1.319	1.805	1.414	.9631	.9769	1

The analysis based on the phonatory settings has yielded significantly lower within- and between-speaker distances, but also lower within- and between-speaker similarities (t-test results available in Table 6-6). Such results oppose the results based on the articulatory settings, where similarity values have increased compared to the analysis based on the entire VPA. Such an outcome, in corroboration with weaker correlation scores than the ones obtained for the articulatory settings, indicates that phonatory settings may play a more significant role in speaker specificity, especially in the mother tongue. It should not be disregarded, however, that part of the explanation lies in the nature of the statistical tests employed. Namely, the reduced number of features in a vector is bound to result in lower Euclidean distances, which accounts for lower scores both in articulatory and phonatory settings observed in isolation.

Table 6-6
Comparison of Euclidean distances and Cosine similarities when assessed from the entire VPA and phonatory settings in isolation

	Euclidean distance			***Cosine similarity***		
	Between-speaker (Serbian)	*Between-speaker (English)*	*Within-speaker*	*Between-speaker (Serbian)*	*Between-speaker (English)*	*Within-speaker*
t-test	15.926	13.712	8.912	18.172	10.531	5.393
p-value	.000	.000	.000	.000	.000	.000
Pearson r	.258	.718	.668	.265	.747	.616

Paired t-test comparison has proven that within-speaker distances are lower than between-speaker distances in both Serbian and English ($t = -2.015$, $p = .058$; $t = -3.88$, $p = .001$), and, in contrast, within-speaker similarities are higher than average between-speaker similarities ($t = 4.479$, $p = .000$; $t = 2.871$, $p = .009$). The results indicate that phonatory features are crucial in maintaining within-speaker similarity across languages; however, the contribution of the articulatory settings to higher between-speaker distances should not be disregarded (compare Table 6-2). The conclusion is corroborated by a moderate correlation for the averaged

between-speaker cosine similarity (r = .583) and Euclidean distance (r = .513) measures in Serbian and English. To put it in plain English, considering solely the phonatory settings, the less similar a speaker is to other speakers in Serbian, the less similar she will be to these speakers in English. Such a relationship was not found when the entire vocal profile was considered.

Qualitative Analysis

For the vocal profiles in Serbian, the experts have identified between 1 and 7 non-neutral settings (mean = 4.1; SD = 1.841). Whereas individual rates occasionally opted for higher scalar degrees, median values reduced the strength of identified settings for most speakers. The settings represented in 50% of the speakers or more include nasalisation, creaky and harsh voice, whereas other frequent settings are raised and backed tongue body, raised larynx and breathy voice (more than 30%).

Table 6-7 summarises the VPA results for the 20 participants when speaking Serbian, whereas Table 6-8 displays the VPA values for their speech in English[27].

Table 6-7
Cumulative results of the Vocal Profile Analysis in Serbian

Vocal tract features		***Slight***	***Marked***	***Extreme***
Labial	Lip rounding/ protrusion Lip spreading Labiodentalisation	1		
Mandibular	Close jaw Open jaw	2		
Tongue tip/blade	Advanced tongue tip/blade Retracted tongue tip/blade Retroflexion	1		
Lingual body	Raised tongue body Lowered tongue body Fronted tongue body Backed tongue body Extensive lingual range Minimised lingual range	8 3 7 5 1		
Pharyngeal	Pharyngeal constriction	4		
Velopharyngeal	Nasal Denasal	9	2	
Larynx height	Raised larynx Lowered larynx / pharyngeal expansion	7 2		

[27] After obtaining the median values by the four speakers, 0.5 was converted to scalar degree 1 (slight), whereas 1.5 was converted to the scalar degree 2 (marked).

Phonation features	*Absent*	*Present*	
Falsetto			
Creak		3	
Whisper			
	Slight	***Marked***	***Extreme***
Creaky	10		
Whispery			
Breathy	7		
Harsh	9	1	
Tremor			

As observed in the data, the experts have detected approximately the same number of non-neutral settings for the speech in English (mean = 4.7; SD = 1.552), ranging between 2 and 7. Advanced tongue tip/blade and pharyngeal constriction are more often encountered in English than in Serbian, whereas backed tongue body, raised larynx, creaky and harsh voice seem to persist across languages for most speakers. Settings that only emerge in the English corpus are denasalised speech and tremor for a limited number of speakers (3-5 speakers), including occasional lip-rounding, retracted tongue tip/blade and whisper (1 speaker).

Table 6-8
Cumulative results of the Vocal Profile Analysis in English

Vocal tract features		*Slight*	*Marked*	*Extreme*
Labial	Lip rounding/ protrusion	1		
	Lip spreading	1		
	Labiodentalisation			
Mandibular	Close jaw	2		
	Open jaw			
Tongue tip/blade	Advanced tongue tip/blade	7		
	Retracted tongue tip/blade	1		
	Retroflexion			
Lingual body	Raised tongue body	2		
	Lowered tongue body			
	Fronted tongue body	3		
	Backed tongue body	7		
	Extensive lingual range	3		
	Minimised lingual range	1		
Pharyngeal	Pharyngeal constriction	7		
Velopharyngeal	Nasal	4		
	Denasal	5		
Larynx height	Raised larynx	7	1	
	Lowered larynx / pharyngeal expansion	3		

Phonation features	*Absent*		*Present*
Falsetto			
Creak			3
Whisper			
	Slight	***Marked***	***Extreme***
Creaky	12	1	
Whispery	1		
Breathy	5		
Harsh	12	2	
Tremor	3		

Overall, 50% of non-neutral settings persist across languages. Observed individually, only 36.5% of the articulatory settings are retained, whereas phonatory settings, creaky, harsh and breathy voice, are more consistent (73%). Such a small percentage of settings retained across Serbian and English may result from the "neutralised" profiles obtained through median values for scalar degrees; therefore, we decided to inspect the retention of settings for each expert, respectively. The results are summarised in Figure 6-2.

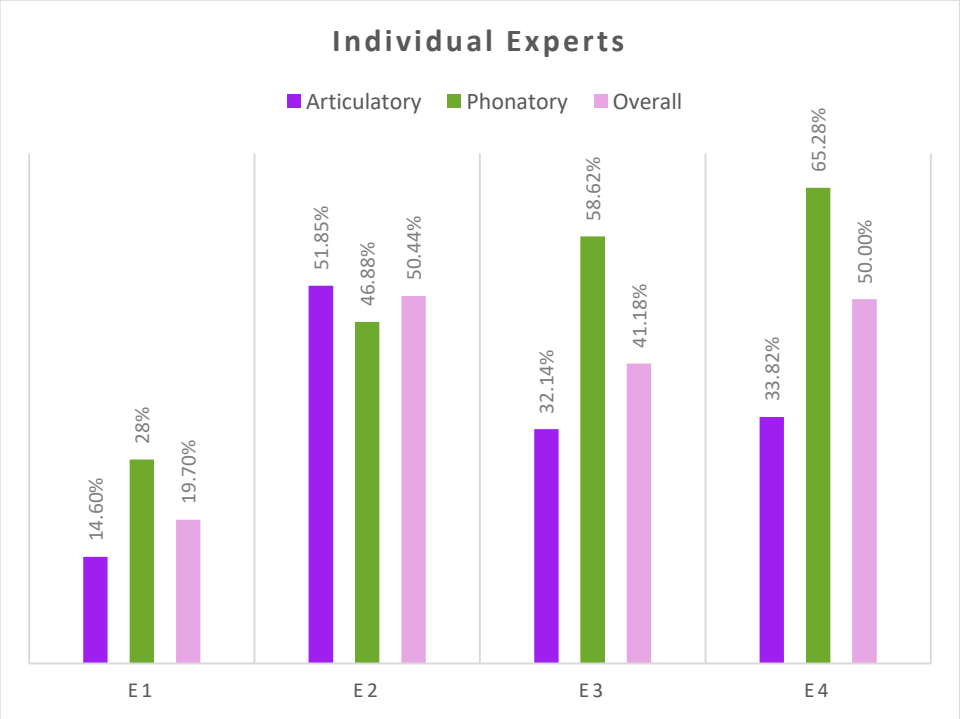


Figure 6-2
The percentage of non-neutral settings retained across languages per expert

The results reveal that the percentage of setting retention differs significantly across experts, ranging roughly from 20% to 50%. Experts 1, 3 and 4 have similar distribution of retained settings (phonatory settings are retained almost twice as more than articulatory settings), but with an increasing percentage of retention. One of the reasons for this difference could be the experience the experts have with the VPA; whereby more experienced experts are able to detect a larger number of settings. Figure 6-3 compares the setting-retention results when summarised on average between the four experts and as their final cumulative score obtained through median calculation.

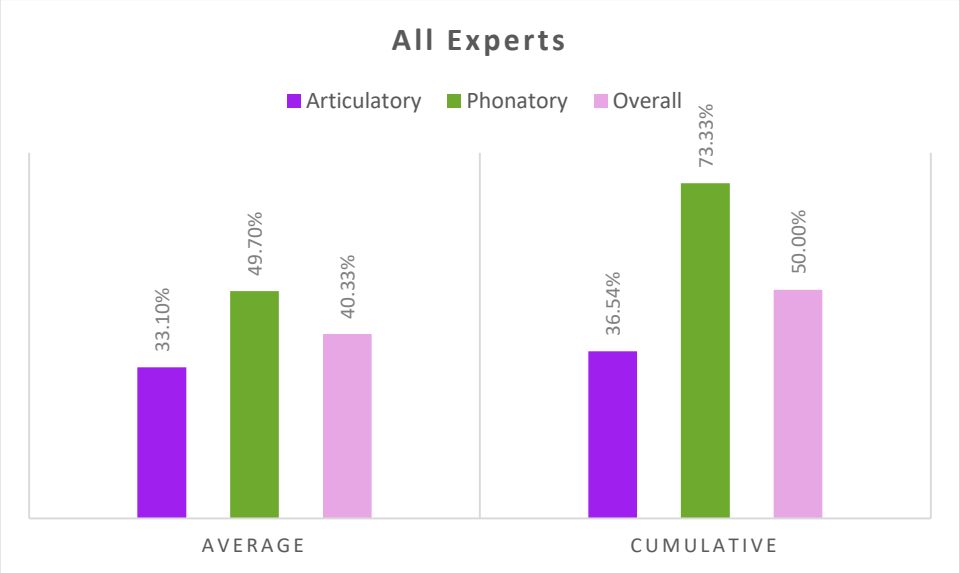


Figure 6-3
The percentage of non-neutral settings retained across languages - all experts

The retention of non-neutral settings across languages is higher when calculated through cumulative median scalar degrees than as an average between individual experts. Such a result may indicate that including a greater number of raters in this experiment would result in more detailed and more precise vocal profiles.

Speaker comparison

Same-speaker (N = 20) and different-speaker (N = 380) vocal profiles were compared between Serbian and English by calculating their Euclidean distances. Table 6-9 contains the identification results, including missed hits (MH) and false alarms (FA).

Table 6-9
Speaker comparison based on Vocal Profile Analysis

Acceptance threshold	***Cross-language***				***Same-language***	
	Same speaker		***Different speakers***		***Different speakers***	
	Correct Identification	*MH*	*Correct Rejection*	*FA*	*Correct Rejection*	*FA*
≤ 2	85%	15%	33%	67%	26%	74%
≤ 1.7	70%	30%	56%	44%	50%	50%
≤ 1.6	70%	30%	63%	37%	58%	42%
≤ 1.5	60%	40%	69%	31%	68%	32%
≤ 1.4	45%	55%	83%	17%	80%	20%

The initial threshold for the acceptance as a "close-matching" profile was d ≤ 2; however, considering that the obtained distance values were generally small, the threshold yielded strong same-speaker bias and was subsequently lowered. The speaker comparison analysis performed this way did not result in favourable error rates – with the equal error rate between 30-37% for the appropriate threshold, the system performs at less than 70% of correct identifications/rejections. Moreover, at this threshold, the number of false acceptances is higher than the number of missed hits, which has negative implications for forensic reality. Finally, the selection of the appropriate threshold is in itself problematic since it depends on numerous factors, including but not limited to the homogeneity of the dataset and the number of assessed features.

The error rates for same-language comparisons are higher than for cross-language comparisons, confirming that the comparison of samples in the same language creates a strong same-speaker bias. Put differently, voices of different speakers speaking the same language appear less distinct than voices of different speakers speaking different languages. Since we used a single sample per language in this experiment, performing the same-speaker comparison in Serbian to confirm our theory is here impossible.

6.1.3. Discussion

Considering that the present study utilises a truncated version of the VPA protocol with fewer settings and only three scalar degrees, it is not surprising that both between- and within-speaker distances are relatively low compared to previous studies. For instance, French et al. (2015) report between-speaker distances that range between 0 and 9. Consequently, cosine similarity scores are very high (above 0.98); therefore, statistically speaking, the profiles of the twenty speakers are virtually identical. In addition, the fact that the speakers for the present

experiment were carefully selected by their sex, age and sociolect renders the corpus very homogenous, contributing to the high similarity between the voices.

Nonetheless, the statistical analysis has revealed insightful information even with such a homogenous dataset. Namely, there is evidence that within-speaker similarity across languages is higher than between-speaker similarity in both the mother tongue (Serbian) and the foreign language (English). Furthermore, both quantitative and qualitative analyses have confirmed that phonatory settings play a more significant role in voice similarity across languages, even though the influence of articulatory settings is not insignificant. Based on our analysis, it can be concluded that the more features the Vocal Profile Analysis has, the more robust the results are.

Regarding the relationship between voice quality and language proficiency, the results have taken a different turn compared to the initial hypothesis, which was based on the pilot study preceding the present analysis (cf. Tomić & French, 2023). Namely, Tomić and French (2023) found a strong negative association between voice similarity and foreign language proficiency, meaning that the better a speaker is at the foreign language the less similar her voice is across languages. Such a finding seems intuitive bearing in mind that each language or dialect has its own "vocal profile" and by acquiring the pronunciation of a foreign language we also acquire the settings inherent to that sociolinguistic community. However, in the present study, we detected weak evidence that more proficient speakers have more similar voices across languages. Some of the difference in the results could stem from the nature of the dataset and the participants. Namely, in the pilot study only two raters and ten speakers were employed while the present research doubled the number of participants, rendering the new results more accurate. On the other hand, the present research has proven that phonatory features are more relevant to similarity of voice quality across languages, therefore, it seems plausible that the more fluent or proficient a speaker is in the foreign language, the more prominent their phonatory features will be. The problem regarding the relationship between voice quality and foreign language proficiency seems to be more complex and layered than we initially assumed. Further research with a larger number of participants (both speakers and expert listeners) is needed to shed light on the questions raised here. In addition, in the future analysis steps should be taken to assign equal weights to articulatory and phonatory features so that none would take supremacy in determining the speakers' voice quality.

Finally, our results indicate that performing numerical cross-language forensic speaker comparison based on low distance scores is not recommended, as the equal error rate exceeds 30%. In addition, it can be challenging to determine the acceptance threshold.

Regarding the single-language comparison, the error rates in the present study are higher than in the results presented by French et al. (2015), who reported the true rejection of 88% for "close matches" in the corpus of 100 male native speakers of English. Such a poor performance of voice quality profiling in same-language different-speaker pairs is probably the result of the low distance values obtained from the truncated protocol.

In conclusion, cross-language forensic speaker comparison may benefit from voice quality analysis based on the VPA protocol primarily from a qualitative perspective as long there is no numerically predefined distance threshold of acceptance or rejection. Our initial hypothesis that speakers retain their voice quality when speaking a foreign language was confirmed, and we have demonstrated that, even in a very homogenous dataset, within-speaker distances are lower than distances between speakers in both the mother tongue and the foreign language. Phonatory features appear to be more relevant for cross-language voice quality comparison as their retention in the foreign language is twice as high as the retention of articulatory features. Whether the results obtained here will be corroborated by the experiment involving naïve listeners remains to be examined in the following chapter. Recommendations for future research within the Vocal Profile Analysis will be presented alongside the summary of research limitations in the Conclusion (Chapter 9).

6.2. Experiment 2 – Naive Listeners

6.2.1. Study design

In the present experiment, the aim is to assess voice similarity across languages as perceived by lay listeners and to determine the robustness of cross-language naïve voice recognition in comparison to the recognition in the mother tongue. Furthermore, it will be explored how speakers' voice quality and language proficiency affect similarity and recognition rates by naïve listeners.

Naïve listeners

The listeners for the present study are 60 native speakers of Serbian, 38 female (63.33%) and 22 male (36.67%) students of English language and literature at the University of Niš. Such participants were chosen because of their basic phonetic and phonological training as well as their knowledge of English. Most listeners reported having grown up in the same dialectological area as the speakers, Prizren-Timok (N=50, 83.33%), several are from the Kosovo-Resava dialectological area (N=8, 13.33%), one from Eastern-Herzegovina (1.67%),

while one was raised abroad (1.67%). The listener-related demographic data is presented in Figure 6-4.

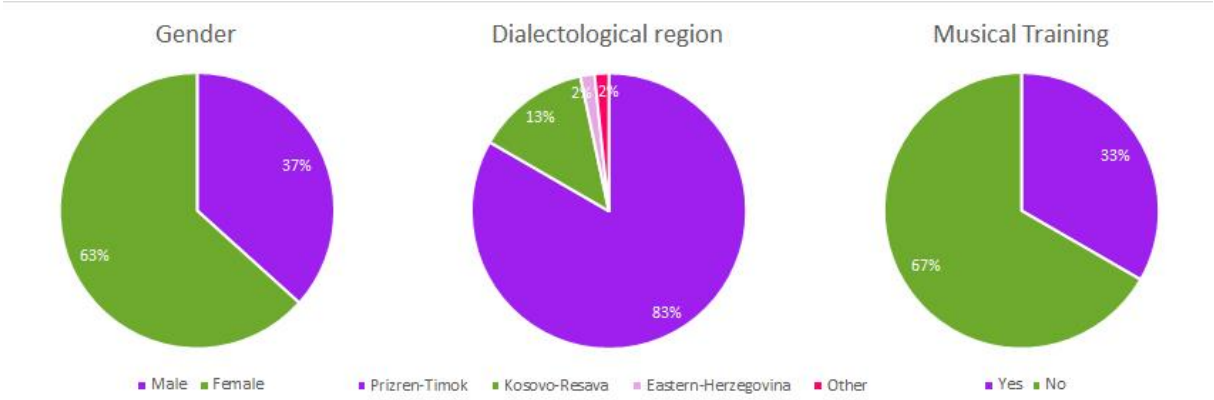


Figure 6-4
Demographic data about the listeners

At the time of the experiment, the listeners were aged 19-24 (mean = 20.52; SD 1.081); on average, this group of participants is younger than the speakers (t = 4.289, p-value = .000). They reported having started to learn English between age 3 and 11 (mean = 6.467; SD = 1.61), which is not unlike the speakers (t = 1.4640; p-value = .432). This group of participants was also asked to rate their English language exposure. Overall, the listeners can be described as more frequent users of English than the speakers (mean = 4.285, SD = 0.46; t = 5.46972, p-value = .000)), with the lowest score for speaking English (mean = 3.67; SD = 1.052) and highest for reading social media content (mean = 4.95; SD = 0.22) and listening to music in English (mean = 4.92; SD = 0.424). While the listeners' English language competence was not assessed in the same manner as the speakers', according to the CEFR level goals described in the accreditation documents for each year of study[28], majority of the listeners can be labelled competent users of English – B2 (N=48, 80%), several modest users – B1 (N=10, 16.67%) and two very good users – C1 (N=2, 3.33%). Finally, 20 listeners (33.33%) reported having had some musical training (singing classes, playing an instrument, attending primary or secondary music school) and all of the participants reported normal hearing.

[28] English language and literature at the University of Niš curriculum https://drive.google.com/drive/folders/1HZgRYMqfWkOcVj8GU5H4uvXSekOzhits

Stimuli

The corpus for the experiment was created from the mobile-phone-recorded speech of 20 female native speakers of Serbian, defined as "the narrow set" in this dissertation (Chapter 5.3.2.). The listeners were presented with 80 stimulus pairs distributed in 4 groups:

A. Serbian - Serbian, same speaker,
B. Serbian - Serbian, different speakers,
C. Serbian – English, same speaker,
D. Serbian – English, different speakers.

The stimuli were presented randomly for every listener to avoid repetition of one and the same context as well as to counterbalance the fatigue effect. The pairs consisted of 15-second long recordings that were normalized in *Audacity* in terms of gain and volume to be comparable. The presented samples were continuous stretches of speech with removed hesitation and psychological pauses. The speaker pairs in B and D were selected randomly and kept consistent in both contexts for comparability (*cf.* Nolan et al., 2013).

Procedure

The experiment was performed via a custom-developed web-based tool (Appendix 6), which was designed to save the participants' progress and allow them to complete the experiment in multiple sittings. The listeners were advised that for participating in the research they should be indoors, in a room with a minimum amount of background noise and that the questionnaire should be completed using a laptop or a desktop computer, preferably with a set of earphones or headphones. After filling out the demographic data, they were asked to disclose how exhausted and stressed out they were feeling. In the listening task, the participants were asked to score how similar the voices were on a 1-10 Likert scale and then perform the recognition. The recordings could be replayed as many times as needed; however, as recommended for the voice line-up procedure (Broeders & van Amelsvoort, 1999; Hollien, 2002), the listeners were allowed to opt out of the recognition task. Finally, the listeners were given an optional opportunity to explain if they relied on any specific speech and voice characteristics to perform the recognition.

Out of 75 volunteers who applied to contribute, only the responses from the participants who completed more than 95% of the questionnaire were considered valid. The number of completed stimulus pairs range between 77-80 per speaker and the chi-square test confirmed equal distribution (χ^2= 0.318, p = 1). Three participants with error rates equal to chance were removed under suspicion of randomly clicking through tasks, and one on the account that she declared to personally know some of the recorded speakers, which finally

resulted in 60 valid responses. When the statistical analysis required equal datasets, the missing data for each listener were filled with their average similarity scores for that particular context (A, B, C or D).

6.2.2. Results

Voice similarity and discrimination

In Table 6-10, we can observe the average similarity scores with standard deviation (SD) for same-language (SL) and cross-language (CL) comparisons for same-speaker (SS) and different-speaker (DS) pairs. In both SL and CL comparisons, there is a clear distinction in scores for SS and DS pairs; however, the scores in CL stimuli slightly lean toward the centre of the scale relative to SL stimuli, as proven by the notably higher distance between the means for Serbian-Serbian and Serbian-English comparisons (t = -52.564 vs t = -38.865).

Table 6-10
Similarity scores with t-tests and p-values between SS and DS pairs across SL and CL stimuli

	Sr-Sr	***SD***	***Sr-En***	***SD***	***t-test***	***p-value***
DS	3.611	2.634	4.427	2.862	7.2486	.000
SS	8.791	2.158	8.454	2.154	3.822	.000
t-test	-52.564		-38.865			
p-value	.000		.000			

The two-factor analysis of variance (Table 6-11) confirms that the similarity scores for same-speaker and different-speaker pairs depend on the language stimuli. Namely, the listeners are prone to using more extreme values to grade similarity/difference between voices in the same-language than in the cross-language comparisons.

Table 6-11
Two-factor ANOVA for SS and DS pair similarity scores in SL and CL stimuli

Source of Variation	***SS***	***df***	***MS***	***F***	***p-value***	***F crit***
SS vs DS	25437.65	1	25437.65	4184.518	0	3.843
SL vs CL	68.56	1	68.56	11.278	.001	3.843
Interaction	398.039	1	398.039	65.478	.000	3.843
Within	29154.84	4796	6.079			
Total	55059.1	4799				

Standard deviation was then calculated per listener; on average, it was found to be higher in DS pairs in both language contexts (SD (A) = 1.906, SD (B) = 2.175, SD (C) = 1.794,

SD (D) = 2.478). Two-factor analysis of variance proved that language context slightly influences SD in same-speaker and different-speaker pairs (Table 6-12).

Table 6-12
Two-factor ANOVA for SS and DS pair SD in SL and CL stimuli

Source of Variation	***SS***	***df***	***MS***	***F***	***p-value***	***F crit***
SS vs DS	13.643	1	13.643	23.894	.000	3.881
SL vs CL	0.546	1	0.546	0.957	.329	3.881
Interaction	2.571	1	2.571	4.502	.035	3.881
Within	134.751	236	0.571			
Total	151.511	239				

The similarity scores and SD were not found to depend on sex, dialect, age of first exposure to English, the estimated English level, or the listeners' self-reported musical training. In addition, exposure-to-English scores were correlated with similarity scores and standard deviations for each context without obtaining significant correlation results, most likely because the group of listeners is homogenous regarding their exposure to the foreign language in question.

As perceived in Table 6-13 and Figure 6-5, the distribution of correct discriminations changes with the language context for SS and DS pairs. Notably, more listeners refrained from performing the discrimination task in the cross-language context. Same-language stimuli expectedly yielded a higher percentage of correct discriminations. The most remarkable difference between two language stimuli can be observed for different speaker pairs, which indicates that this particular context poses the gravest challenge for speaker discrimination.

Table 6-13
Speaker discrimination percentage with χ2 for distribution in SL and CL stimuli

	Same-language			***Cross-language***			***χ2 test***	***p-value***
	Correct	*False*	*Not sure*	*Correct*	*False*	*Not sure*		
SS	78.98%	15.16%	5.86%	75.15%	14.01%	10.84%	19.43	.000
DS	90.95%	5.03	4.02%	72.84	14.92%	12.24%	132.1	.000
Overall	84.97%	10.09%	4.94%	74%	14.46%	11.54%	99.797	.000

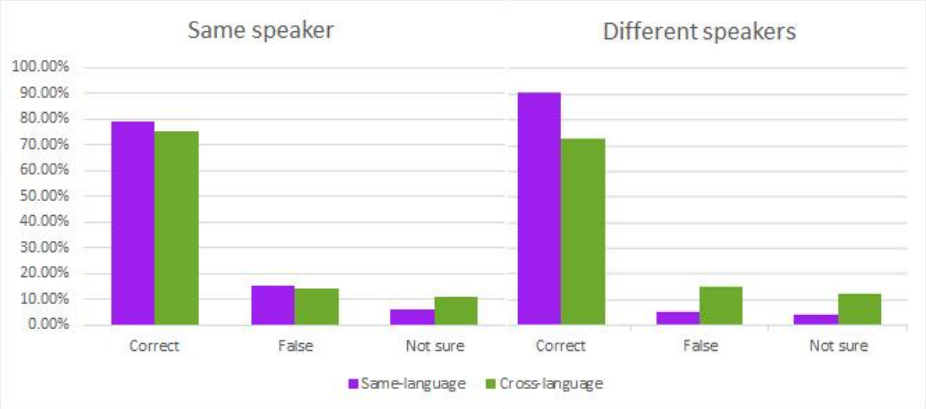


Figure 6-5
Distribution of correct discriminations across language contexts in SS and DS pairs

False alarms and missed hits were calculated per listener solely on the basis of attempted discrimination; the tokens marked as "not sure" were not taken into consideration (Table 6-14).

Table 6-14
Error rates for speaker discrimination in SL and CL stimuli with t-test and correlation scores

	False alarms	***Missed hits***	***Overall correct discrimination***	***SD***	***t-test***	***p-value***	***Pearson***
SL	5.34%	16.4%	89.37%	5.374	5.889	.000	.508
CL	17.7%	16.15%	83.64%	8.669			

As can be noted in Table 6-14, there is a slight same-speaker bias in cross-language stimuli, and the overall percentage of correct discriminations is significantly lower in this context. Two-factor analysis of variance confirmed the influence of language context on error rates (Table 6-15).

Table 6-15
Two-factor ANOVA of MH and FA for speaker discrimination in SL and CL stimuli

Source of Variation	***SS***	***df***	***MS***	***F***	***p-value***	***F crit***
MH vs FA	1337.393	1	1337.393	8.984	.003	3.881
SL vs CL	2178.215	1	2178.215	14.633	.000	3.881
Interaction	2364.762	1	2364.762	15.886	.000	3.881
Within	35130.19	236	148.8577			
Total	41010.56	239				

Furthermore, we found a linear correlation trend for correct discrimination in the same-language and cross-language contexts. In general, listeners who performed better in the former context also performed better than their peers in the latter (Figure 6-6).

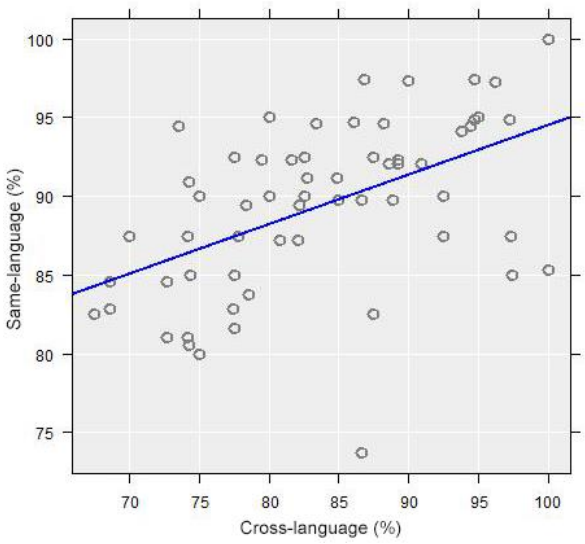


Figure 6-6
Linear correlation of overall correct discriminations in same-language and cross-language stimulus pairs

No relationship was found between the error rates and listeners' sex, dialect, age of first exposure to English, estimated level of English, exposure to English or self-reported musical training.

In participant selection, Hollien (1990: p. 205) suggests choosing only the listeners who can demonstrate that they can discriminate between same-speaker pairs at a level of 80% or better and different-speaker pairs with at least 85% accuracy. Therefore, we narrowed the participants to 37 by selecting only the listeners who fulfilled both criteria. The results obtained for the entire group of listeners were replicated with the super-recognisers but with slightly stronger similarity scores and lower error rates (Table 6-16). However, the statistical analysis did not confirm a significant difference between the results in the cross-language context for the two groups of listeners.

Table 6-16
Error rates for speaker discrimination by super-recognisers in SL and CL stimuli with t-test and correlation scores

	False alarms	***Missed hits***	***Overall correct discrimination***	***SD***	***t-test***	***p-value***	***Pearson***
SL	4.25%	10.41%	92.68%	3.16	4.740	.000	.518
CL	16%	12.06%	86.22%	7.66			

Voice quality and speaker discrimination

In order to explore the relationship between voice quality as scored by expert listeners and speaker discrimination by naïve listeners, we observed the results for four different contexts respectively (*see* Chapter 6.2.1. on Stimuli). The summary of the results is presented in Table 6-17 below.

Table 6-17
Distribution of false identification and non-identification responses across four contexts

False identifications	***A***	***B***	***C***	***D***	***Non-identifications***	***A***	***B***	***C***	***D***
Between 50%-100%	1	0	1	1	Between 50%-100%	0	0	0	0
Between 20%-50%	5	1	3	6	Between 20%-50%	1	0	3	4
Between 10%-20%	4	2	5	4	Between 10%-20%	3	4	8	8
Less than 10%	10	17	11	9	Less than 10%	16	16	9	8

The values in Table 6-17 can be interpreted in the following way: in same-language, same-speaker pairs (Context A), one speaker pair was falsely rejected as a different speaker by more than 50% of the listeners, while five other pairs were falsely rejected by more than 20% of the listeners. One speaker pair yielded a high percentage of nonidentifications in this context (more than 20%).

In order to understand the relationship between naïve listeners' similarity scores and discrimination percentage and VPA-based voice quality analysis, we correlated Euclidean distances and Cosine similarities with lay listeners' scores for each speaker pair across the four contexts, respectively. No significant correlations were found for any context, which implies that naïve listeners do not necessarily (or at least not solely) rely on equivalent features scored on the VPA protocol when assessing speaker similarity or making discrimination decisions.

Furthermore, to study the speakers' relationship to the discrimination scores, we converted the results to speaker-focused data by calculating the percentage of correct, false and non-identifications for each speaker. The values were then correlated with each other. It was found that speakers with higher false rejection in SS SL context (A) have a higher false acceptance rate in DS CL context (D) when their sample is in Serbian ($r = .461$, $p = .041$). Furthermore, speakers who tend to have fewer correct rejections in DS SL comparisons (Context B) yield a higher percentage of non-identifications in DS CL pairs (Context D) when their sample is in English ($r = -.522$, $p = .018$), and the higher the false acceptance in DS SL, the higher non-identification in English ($r = .469$, $p = .037$). This implies that, for equivalent speaker pairs, listeners are more eager to accept the non-identification option in language

mismatch conditions, while they feel more confident to perform the identification when both samples are in Serbian, even though it might result in a mistake (Figure 6-7).

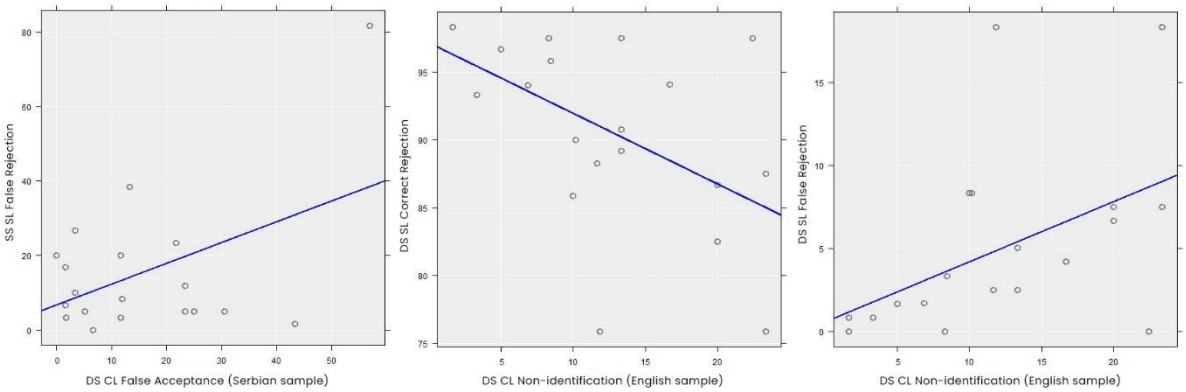


Figure 6-7

Speaker-focused correlations of discrimination results in the same-language context

In addition, a strong association was found between SS CL correct identifications (Context C) and DS CL (Context D) correct rejections when the sample is in English (r = .488, p= .029). Thus, the more similar a speaker is to herself when she speaks two languages, the easier it is to distinguish her from other speakers when she speaks English. On the other hand, speakers with fewer correct identifications in the SS CL context have a high percentage of non-identifications in the DS CL context when their sample is in English (r = .626, p = .003). Next, the higher the percentage of false rejections in the SS CL context, the lower the percentage of correct rejections in the DS CL context when this sample is in English (r = -.545, p = .013). The implication is that some speakers diverge very much from their native Serbian when speaking English to sound like someone else. Finally, the higher percentage of false rejections in the SS CL context coincides with a higher percentage of non-identifications in the DS CL context (r = .656, p = .002), indicating that, even though the language mismatch introduces a strong different-speaker bias for same-speaker pairs, when the voices are very distinct it is challenging to make a decision, particularly in the DS context (Figure 6-8).

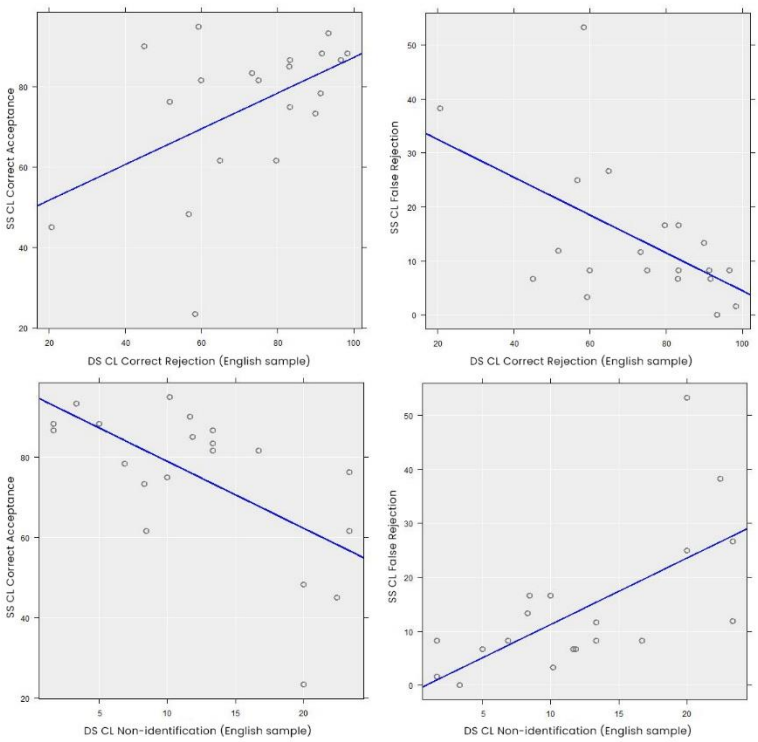


Figure 6-8
Speaker-focused correlations of discrimination results in the cross-language context

Considering that these relationships do not appear for the DS CL samples in Serbian, only for the samples in English, the conclusions should not be generalised to concern the speakers' voice in its entirety; instead, they relate to how the speakers sound when they speak English.

To explore how voice quality is related to the speaker-focused discrimination scores, we correlated VPA-based speaker average Euclidean distances and Cosine similarities and within-speaker (cross-language) distances/similarities to the discrimination percentages for each context, respectively. The results (Figure 6-9) indicate that DS CL correct rejections are more common for speakers with higher within-speaker cosine similarity ($r = .651$, $p = .002$). Put differently, if a person's voice in English and Serbian is very similar, it is easier to distinguish this person's English voice from other speakers' Serbian voices. Conversely, there is a negative association between the percentage of false acceptances and within-speaker cosine similarity when the sample is in English ($r = -.726$, $p = .000$). The opposite is true of within-speaker Euclidean distances, the lower the cross-language distance, the higher the discrimination for that particular speaker's English sample ($r = -.600$, $p = .005$; $r = .672$, $p = .001$). Another correlation was found between average between-speaker cosine similarity for samples in English and false rejections in the SS CL context ($r = -.450$, $p = .047$). It can be

interpreted that if a particular speaker sounds more similar to others when they speak English, it is easier to discriminate her against herself in the cross-language context.

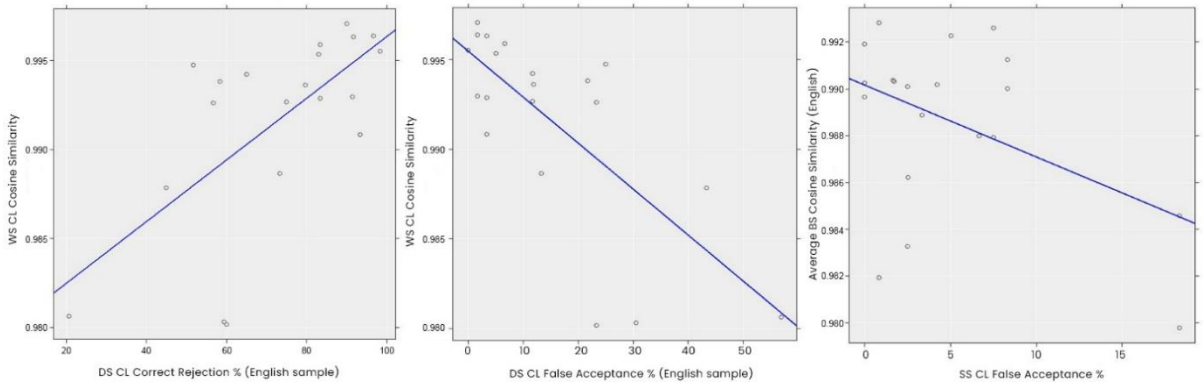


Figure 6-9
Speaker-focused correlations of discrimination results and VPA-based distance/similarity scores

Finally, we wanted to explore the relationship between foreign language proficiency and speaker discrimination; however, no correlation was found between the percentage of correct/false identifications and proficiency scores for this group of speakers.

Voice features – qualitative analysis

Eight hundred fifty (850) discrimination tasks were followed by a comment concerning the speech/voice characteristics that helped the listener decide whether the samples originated from the same person. Although the listeners have had basic phonetic and phonological training, many did not rely on the established phonetic terminology to denote specific features of voice/speech. For instance, tone of voice was often preceded by adjectives such as soft or raspy, which indicates that the term was used to denote the voice timbre. The listed features were grouped into five major categories (Table 6-18), derived from Laver's generic phonetic concepts (Laver, 1994).

Table 6-18
Classification of voice/speech characteristics listed by the listeners in the discrimination task

Category		
Accent		Accent, dialect, pitch accent (pronunciation of words)
Segments		articulation of consonants (lisp, "hard" consonants), vowels (vowel length, openness)
Suprasegmental features	Tempo	Pauses, hesitation markers (fast, slow, speed, pace)
	Prosody	Intonation (cadence), pitch (frequency, deep, low, high), loudness (volume), intensity

	Metrics	Stress (word/sentence stress), prominence (speech pattern), syllables, rhythm
Voice quality	Phonation	Vocal fry, breathiness (aspiration), harshness
	Articulation	diction, slur, muffled voice, nasality
	Voice timbre (tone of voice)	*Impressionistic labels*: mature, young, raspy, soft, clear, feminine, thin, sharp, dull, constrained, smooth, strong
Delivery		enthusiastic, confident, casual, forward, fluent

Much like Winters et al.'s (2008) participants, our listeners relied on everything at their disposal to discriminate between the speakers. The qualitative analysis revealed that the same features were mentioned in same-language and cross-language stimuli. In cross-language discrimination, the accent was often used to infer that the foreign language speaker was Serbian, which sometimes led to the wrong conclusion that the two samples were uttered by the same person. Finally, this may be a part of the explanation for the same-speaker bias present in different-speaker cross-language discrimination tasks.

6.2.3. Discussion

The results of the present study align with previous research where similarity between voices in same-language stimuli was rated higher than in cross-language stimuli (Fleming et al., 2014), particularly with regard to same-speaker pairs. That the introduction of mismatched conditions increases the dissimilarity score was also reported by Nolan et al. (2013), who found that recording pairs of mixed conditions in terms of transmission channels would yield relatively high values compared to matching-condition recordings. Curiously, however, in the present research, cross-language different speaker pairs received higher similarity scores than same-language different speaker pairs, the scores leaning toward the centre of the scale. Such neutralised similarity scores most likely reflect the listeners' uncertainty about whether the two samples originate from the same speaker.

Concerning speaker discrimination, the results of our experiment confirm previous findings that cross-language discrimination poses a greater challenge to listeners than discrimination of speakers in their mother tongue (cf. Mok et al., 2015; Wester, 2012; Winters et al., 2008), even if the listener group is narrowed down to the super-recognisers. Such results could be ascribed to the so-called "language-familiarity effect", that is, the notion that voice memory is inextricably linked to the linguistic aspects (syntax, lexicon, phonology) learned through the exposure to voices in our local community (see Perrachione, 2019: p. 520). Fleming (2014) explains that the lower discrimination in this context possibly arises from the concept that subjectively perceived similarity between different voices tends to be higher for a foreign

or unfamiliar language, analogous to the "other-race" effect in face recognition. Another observation is that cross-language discrimination tasks yielded a higher percentage of non-identifications, the number remarkably increasing for different-speaker pairs. Therefore, we can conclude that listeners feel less confident about making a decision in the cross-language context, and if given an option not to perform the discrimination, they are likely to use it.

Moreover, same-language comparisons yielded more "different" identifications, while cross-language comparisons have a higher percentage of false acceptances than false rejections. Such a result reveals that two samples are more likely to be perceived as originating from the same speaker in a cross-language than in a same-language context. A tentative conclusion could be that the listeners ascribe the difference between voices they hear to the "language effect" and therefore disregard it when making a discrimination decision between two samples in different languages.

Furthermore, we aimed to explore the relationship between voice quality and speaker similarity as perceived by naïve listeners. Unlike Nolan (2007), we were unable to find the relationship between similarity scores and VPA-based distance ratings. A consideration for future research may include naïve listeners scoring each pair of voices in order to be able to perform multidimensional scaling (cf. McDougall, 2013; Nolan et al., 2013) and compare the results with the ones obtained from the VPA scores. Nonetheless, focusing on individual speakers, we detected correlations concerning their VPA-based distances/similarities and listeners' ability to discriminate them correctly. Two general observations stem from the obtained results:

(1) speakers whose Serbian and English VPA are very similar are easier to distinguish from other speakers in cross-language comparisons when their voice sample is in English (conversely, the more distant the speaker is from herself, the more difficult it is to discriminate her in the CL DS context correctly)

(2) speakers whose VPA scores are closer to the population in English have lower false rejection scores in the same-speaker cross-language context (less distinct voices better discriminated in CL SS pairs)

The first observation reinforces the finding that the speakers with a higher number of correct identifications in the SS CL context were also the ones with the highest number of correct rejections in the DS CL context (Figure 6-8). The implication, however, may be that, although we were using different samples, the listeners may have remembered the speakers' voices throughout the experiment and were at some point able to recognise that the offered sample in Serbian was not uttered by the same person as the English sample. In order for us to

understand whether the observed correlation is the effect of voice quality or voice memorability, future experiments should use samples from a greater variety of speakers without repetitions across contexts. On the other hand, the second observation could be understood to reinforce the previous conclusion that in the cross-language context, listeners ascribe the distinction they hear to the language effect and therefore tend to produce more false acceptances. However, if a voice is more typical, the listeners assume that the language effect is less prominent, making it easier for them to identify the speakers correctly.

6.3. Perceptual experiments – Discussion

In this chapter, we presented two perceptual experiments on the same corpus of speakers. The first experiment involved four expert listeners who assessed the Serbian and English samples on a Vocal Profile Analysis protocol, while, in the second experiment, sixty naïve listeners were engaged to score voice similarity and perform speaker discrimination in four conditions (same speaker, same language; different speaker, same language; same speaker, different language; and different speaker, different language). In this interim discussion, we will return to the research questions (1) - (5) raised in Chapter 5.1., and consider the findings obtained in the two perceptual experiments.

According to the results of the VPA analysis, within-speaker cross-language distances are lower than between-speaker distances in their mother tongue and in the foreign language. It was found that phonatory settings contribute to the within-speaker similarity across languages more than articulatory settings and are more robust to language change, even across different raters. Namely, almost half of the detected phonatory features are retained when the speakers switch from native Serbian to foreign English, whereas the retention of articulatory features is rater-dependant and varies between 14% and 50%.

The results obtained through naïve listener assessment reveal that while same-speaker voices were rated slightly more distinct in the cross-language context, different-speaker voices have a notably higher similarity score in the language-mismatching than in the language-matching condition. Such a result can be interpreted that naïve listeners ascribe the difference they hear to the language effect and thus try to compensate for it with a higher score. Involving a third condition, in which both samples would be in English, would shed more light on the issue and help interpret the results better.

When we observed the speaker pairs, the correlation statistics did not reveal any significant relationships between the similarity scores assigned by the naïve listeners and distance/similarity measures obtained through the experts' voice quality analysis. However,

when the results were converted to represent individual speakers, it was found that speakers whose voice quality was rated as very similar in Serbian and English on a VPA protocol tend to be easier to discriminate in the cross-language different-speaker setting. A similar benefit, but in the cross-language, same-speaker context, was detected for the speakers whose vocal profiles do not considerably diverge from the population.

In order to assess the discriminatory power of voice quality rated on a VPA protocol in a cross-language setting, we performed a simple speaker comparison varying the acceptance threshold. The discrimination was slightly improved relative to single-language speaker comparison, with an equal error rate between 30% and 37%. In this regard, the discrimination performance by naïve listeners appears more reliable than raw distance values of the vocal profiles, with error rates as low as 16% in the same-speaker and 18% in different-speaker pairs. While the naïve listeners' performance deteriorates in the mismatched conditions (same-language ER 16%/5%), it is still far more reliable than voice quality assessment through Euclidean distances and Cosine similarity. The reason for this most probably lies in the fact that naïve listeners use everything at their disposal to decide whether two samples originate from the same speaker or not, including, but not limited to, the pronunciation of individual segments, tempo, prosody, metrics, delivery and voice quality. On the other hand, the truncated version of the VPA protocol employed in this research primarily focuses on specific articulatory positions and phonation, therefore incorporating inherently less information than naïve listeners have at their disposal.

Such a result does not, however, imply that naïve listeners are to be considered more reliable than a phonetic instrument or that voice quality ought to be excluded from the speaker comparison procedure whatsoever. Namely, as the results have shown, naïve listeners tend to introduce a same-speaker bias in the cross-language comparisons, which is considered dangerous in the forensic context. Moreover, while Euclidean distances and Cosine similarity cannot capture the fine-grained differences in the vocal profiles of the speakers, expert auditory analysis can discriminate between two speakers on the basis of a single parameter (e.g. fronted or backed tongue body, creaky voice). Therefore, Vocal Profile Analysis can be considered a useful auditory tool to corroborate other evidence in FSC.

Finally, the study explored the relationship between voice quality and language proficiency. The results give weak statistical evidence that intermediate-level speakers exhibit lower within-speaker similarity across languages than upper-intermediate speakers. Such a finding is in stark contrast with the pilot research performed on the part of the corpus used here, which detected a relatively strong negative correlation between vocal profile similarity across

languages and English language proficiency estimated through both the test and an IELTS-based scoring (see Tomić & French, 2023), Namely, according to the pilot study, speakers with lower proficiency level tend to have a higher similarity of vocal profiles across the two languages. As pointed out above, the distinction in the results may stem from the fact that, for the present dataset, phonatory features play a more significant role in voice similarity across languages. At the same time, the pilot analysis relied on the assessment of only two experts (E1 and E2), one of whom produced remarkably detailed profiles in terms of articulatory settings. Future research should select a more balanced corpus of speakers with distinct proficiency levels and employ a greater number of voice quality experts to obtain more reliable results regarding this issue. A closer observation of articulatory and phonatory data in isolation may also provide insight into the dependency of cross-language voice quality on pronunciation. However, employing a more detailed, non-truncated protocol is strongly encouraged in this case, considering that the reduction in the number of features results in lower distances between speakers.

7. Part 2 – Acoustic Analysis and LR calculations

7.1. Acoustic Analysis

The acoustic analysis in the present study is performed to reflect both the articulatory and phonatory voice quality of the speakers. The articulatory parameters explored here are long-term frequencies of the first three formants (F1, F2 and F3), as well as covariance of the second and third formant. The fourth formant, even though it has shown low within-speaker variability in previous research (Tomić, 2020; Tomić & French, 2019), is not analysed because the current corpus is comprised of mobile phone recordings, thus, for a large number of speakers, it was impossible to extract its values correctly. The phonatory features evaluated in the present research primarily concern those that reflect whispery/creaky distinction of the voice, including H1*-H2* (difference between the amplitude of the first and the second harmonic), H2*-H4*, H1*-A1* (difference between the amplitude of the first harmonic and the harmonic nearest to F1), H1*-A2*, H1*-A3*, H4*-2K* (difference between the amplitude of the fourth harmonic and the harmonic nearest to 2000 Hz), HNR_{05} (harmonic-to-noise ratio between 0-500 Hz), HNR_{15} (between 0-1,500 Hz), HNR_{25} (between 0-2,500 Hz) and HNR_{35} (between 0-3,500 Hz), as well as CPP (cepstral peak prominence), a measure of voice perturbation Measures of harmonics close to 5000 Hz were not incorporated in the present study given the nature of the corpus.

The results are compared across language contexts using descriptive and inferential statistics, while the performance of the extracted parameters in a forensic speaker comparison system is evaluated within Bayesian likelihood ratio framework. The statistical analysis was performed in RStudio using packages “tidyverse” (Wickham et al., 2019), “dplyr” (Wickham et al., 2023) and “data.table” (Dowle & Srinivasan, 2020) for data organisation; “effectsize” (Ben-Shachar et al., 2020) and “pwr” (Champely, 2020) for estimation of strength of evidence and “ggpubr” (Kassambara, 2023) for generation of plots.

In the present study, we are also interested in comprehending whether the speakers’ proficiency in the foreign language affects the ability of the system to match or discriminate their voice samples across two languages. Such a perspective on the results is in accordance with recent shift of interest from mere evaluation of the robustness of a system to understanding of the performance of individual speakers within it (see Cardoso et al., 2019; Hughes et al., 2018; 2022a; 2022b; Lo, 2021).

7.1.1. Extraction of parameters

All of the parameters were measured on stressed and unstressed vowel segments throughout the utterance. The extraction of vowels was performed semi-automatically for both Serbian and English samples. The orthographic transcription of the utterances was performed automatically using an online transcription service (www.veed.io), after which the text strings were converted to TextGrid intervals using EasyAlign macro-segmentation tool (Goldman, 2012) with the manual correction of the text and boundaries. Next, a SpeCT automatic forced alignment tool (Lennes, 2022) was used to derive "word" and "phoneme" tier transcription, after which the highest tier was extracted and vowel boundaries were manually adjusted. Vowels were then extracted as separate sound files using a script available within Fast Tract Praat toolkit (Barreda, 2021). The number of the extracted vowels on the first pass was approximately 65,000, around 27,500 in English (average duration 0.12s) and 37,500 in Serbian (average duration 0.08s). The initial goal was to obtain 60 seconds of vowels for analysis (cf. Hughes et al., 2017, 2018, 2019), however, some of the speakers were not able to produce enough speech in English or some parts of the speech had to be excluded due to the quality of the signal. Therefore, in order to keep more speakers in the corpus, we opted for the 56-second-long vowel recordings[29]. The estimated number of the analysed vowels is approximately 57,800 (23,600 in English and 34,200 in Serbian).

Formant measurements were taken at 5ms throughout all vowel sounds longer than 35ms[30], using Fast Track, an LPC-based formant estimation toolkit for Praat (Barreda, 2021). The script performs multiple analyses following the adjustable settings and chooses the best track by modelling smooth formant contours across the vowel (Figure 7-1). One of the benefits of this toolkit is that it provides the images of analysed spectrograms, thus the researcher is able to manually check and discard poor analyses or correct formant paths where possible. Covariance of the second and third formant was performed using the base function in R, relying on Pearson's method, by co-varying adjacent 20 formant values (or 100ms length of vowels).

[29] For the formant analysis, due to poor spectrogram quality and removal of very short vowels, six (out of 100) recordings did not reach 56s in length, so the missing data was substituted by the average obtained through multiple imputation by chained equations from the existing measurements using predictive mean matching method in five iterations using the "mice" package in RStudio (van Buuren & Groothuis-Oudshoorn, 2011; Rubin & Schenker, 1986). The length of the compensatory material was between 0.12 seconds (0.21% of the missing data per recording) and 1.65 seconds (2.96% of the missing data per recording), average - 0.727s (1.3% missing data per recording, or 0.07% of the entire corpus). The same method was used for data imputation of VQ values for 3 recordings, parts of which were not successfully analysed by the program, whereby the imputed values comprise 0.14% of the data.

[30] Fast Track does not support the analysis of vowels shorter than 35ms (Barreda, 2021).

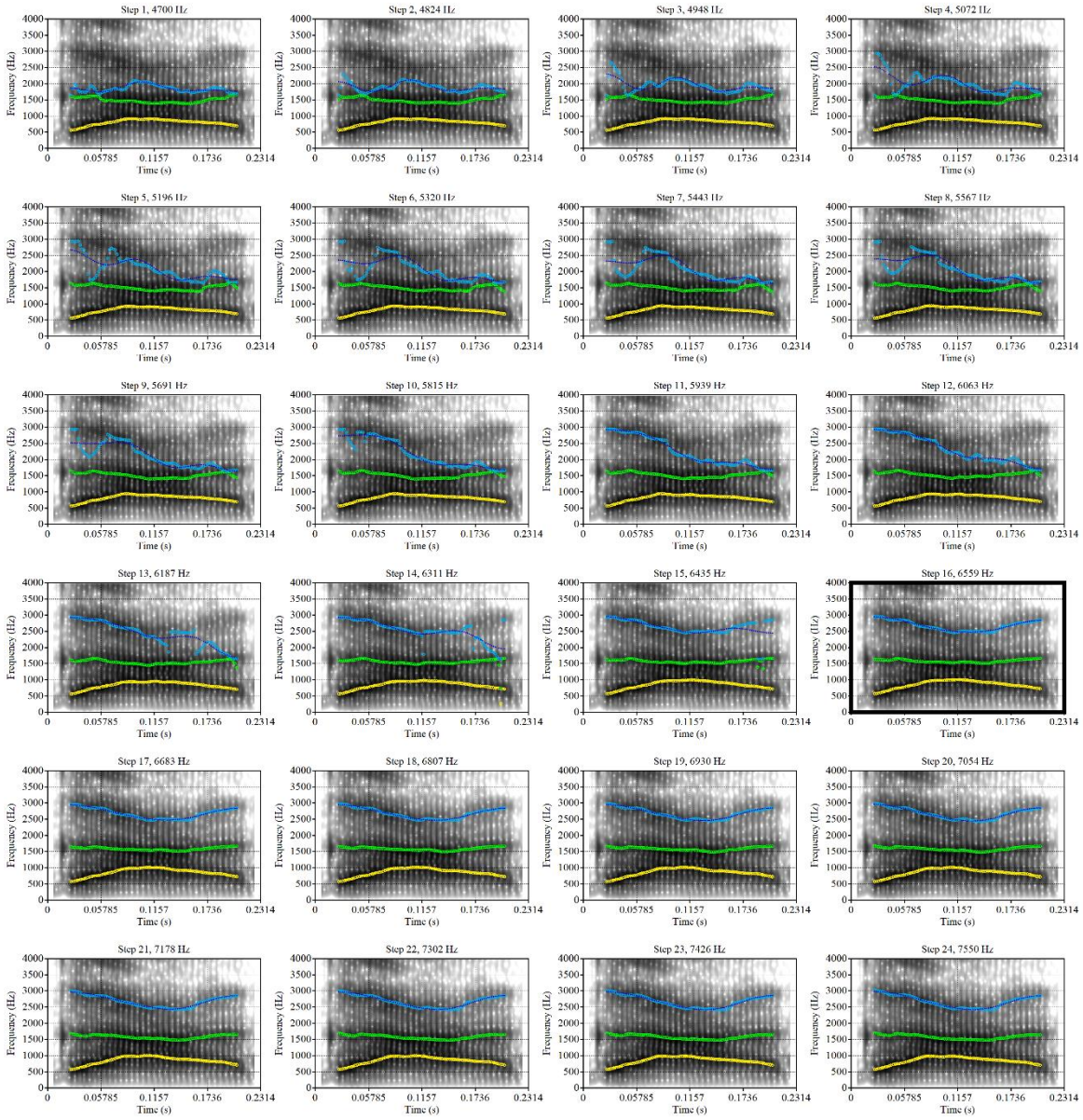


Figure 7-1
Fast Track Toolkit Analysis Illustration

Measurements related to harmonics were performed in VoiceSauce, a compiled MatLab script for automated voice analysis (Shue et al., 2011). The program was set to rely on Snack Sound Toolkit (Sjölander, 2004) for estimation of fundamental frequency and formants, necessary to find the location of harmonics; window length was set to 25ms and computation of harmonics and HNR was performed over five pitch periods at a 1-millisecond frame. Since formants boost the amplitude of any nearby harmonics, raw harmonic amplitudes reflect both the source and the filter, that is, depend on the vowel quality. Bearing in mind that in the present

study, vowels are observed cumulatively, all of the parameters were calculated with the corrected harmonic amplitude (Iseli et al., 2007). Estimation of CPP in VoiceSauce is performed according to the algorithm described in Hillenbrand et al. (1994), whereas HNR is calculated according to de Krom (1993).

7.1.2. Results – across languages

In this section, the acoustic analysis results are presented as raw data and through a series of statistical procedures, and primarily concern the comparison of values for speech samples in Serbian and English. First, we will observe the parameters derived from formant values; second, we will explore the acoustic correlates of phonatory features derived from harmonics and noise in the signal, and finally, we will explore the relationships between these parameters

Articulatory features - formant values

The mean values presented in the table below are first averaged across each speaker and the standard deviation (SD) is the measure of variation of averaged mean, which is why it is rather low compared to the SD of the particular parameter for a single speaker. Table 7-2 provides the results of the analysis of the averaged mean values.

Table 7-1
Summarised mean and SD of long-term formant values in Serbian and English

Parameter	***Serbian***		***English***	
	Mean (summarised)	*SD (mean)*	*Mean (summarised)*	*SD (mean)*
LTF1	587.87	35.69	590.11	30.96
LTF2	1616.05	72.45	1765.87	87.25
LTF3	2772.98	146.56	2783.53	128.24
cov(F2-F3)	14654.91	7405.26	14.791.33	7050.49

Table 7-2
Paired t-test of summarised formant values across Serbian and English

Parameter	***t-test***	***p-value***	***Cohen's d***	***d range***	***power***
LTF1	-0.73178	.4678	-0.1	-0.38, 0.17	.11
LTF2	-16.962	.0000	-2.4	-2.94, -1.85	1
LTF3	-1.0699	.2899	-0.15	-0.43, 0.13	.18
Cov*(*F2-F3)	-0.152.15	.8797	-0.02	-0.30, 0.26	.05

As seen in Table 7-2, the values of the first and third formant, as well as F2-F3 covariance, do not exhibit statistically significant difference across languages. However, the

power[31] of the statistical analysis is rather low, thus there is a high probability of Type II error. The results are unambiguous for the long-term F2, which has higher values in English and strong effect size. Such a result is in accordance with our observation in Chapter 4.2 that English vowels are pronounced as more fronted than Serbian. Figure 7-2 illustrates the results on boxplots.

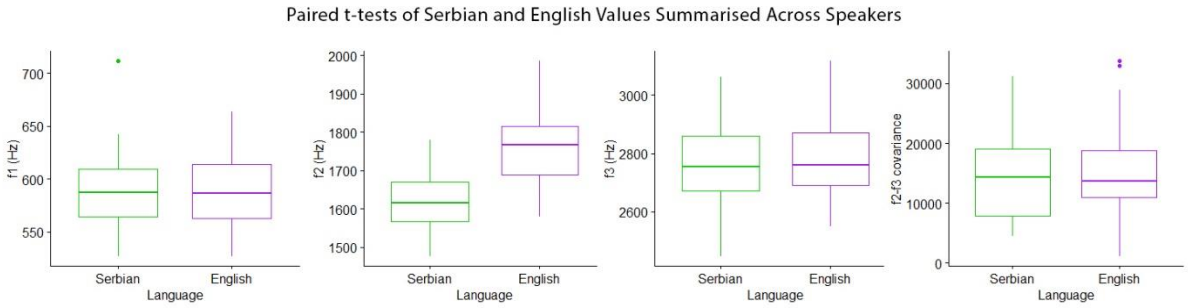


Figure 7-2
Comparison of formant values in Serbian and English, summarised across speakers

Appendix 7 lists the statistical analysis per speaker; however, the low effect size indicates that the test is too sensitive due to a huge sample size (n= 11,200) and the results are most likely indicating a Type I error. Therefore, the analysis was repeated with random sampling (200 measurements, corresponding to 1s of signal) relying on bootstrapping with 100 replications. Table 7-3 lists the average scores across all speakers while the full, per-speaker analysis is available in Appendix 8. The newly obtained t-values are notably lower than in the previous calculation and support the view that long-term formants do not differ significantly in Serbian and English for individual speakers. The exception is F2, which, much like in the summarised results above, exhibits strongest divergence across languages.

As seen in Table 7-3, the power measures now range from moderate to strong, therefore we can state more reliably that the results are representative of the entire dataset. The *t*-score and Cohen's *d* in Table 7-3 are the average values of t-scores and effect size across all speakers, SD is the measure of variation of these *t*-scores, whereas power is derived from the average *d*.

[31] For interpretation of power and effect size see Cohen (1988).

Table 7-3
Bootstrapped t-scores and effect size of cross-language formant value comparisons, averaged for all speakers

Parameter	*Averaged*	*Value*	*Derived*	*Value*
LTF1	t-score	0.1474	SD (t-test)	1.0060
	Cohen's d	0.0184	Power	.3983
LTF2	t-score	2.1168	SD (t-test)	1.0352
	Cohen's d	0.3097	Power	.8669
LTF3	t-score	0.2623	SD (t-test)	1.0075
	Cohen's d	0.0364	Power	.5193
Cov(F2-F3)	t-score	-0.0119	SD (t-test)	0.9163
	Cohen's d	0.0006	Power	.4047

In order to compare within-speaker and between-speaker variability, we performed one-way analysis of variance for data in each language respectively (Table 7-4). The results indicate higher between-speaker than within-speaker variability in the selected parameters for both languages, respectively, with effect size (Eta squared - η^2) ranging from low (for F2 and Covariance of F2-F3), across moderate (for F1) to high (for F3).

Table 7-4
One-way ANOVA of formant values across speakers

Parameter	*Language*	*df*	*denom df*	*F-test*	*p-value*	η^2
LTF1	Serbian	49	559950	682.39	.0000	.056
	English	49	559950	485.75	.0000	.041
LTF2	Serbian	49	559950	236.19	.0000	.02
	English	49	559950	326.38	.0000	.023
LTF3	Serbian	49	559950	3193.7	.0000	.218
	English	49	559950	1664.8	.0000	.127
Cov (F2-F3)	Serbian	49	27950	14398	.0000	.025
	English	49	27950	10276	.0000	.018

The measurements were repeated with reduced sample size (200 random formant values) in 100 replications, with the same effect size, thus confirming the results above. The summary of the bootstrapped ANOVA results is available in Table 7-5. The difference between the ANOVA over the entire data and bootstrapped ANOVA is notable for the covariance of the second and third formant. Namely, in the former analysis, this parameter exhibits the strongest difference between groups, while in the latter it is the weakest. The discrepancy most likely originates from the fact that in random sampling we do not control from which vowels the measurements are taken and the random 1-second sample could be phonetically unbalanced.

Table 7-5
Bootstrapped ANOVA and effect size of formant values, averaged across all speakers

Parameter	*Language*	*Av. F-test*	*SD of F-test*	*Av. p-value*	*Av. η^2*
LTF1	Serbian	13.14	1.02	.0000	.061
	English	9.66	0.78	.0000	.045
LTF2	Serbian	5.24	0.6	.0000	.025
	English	6.9	0.67	.0000	.032
LTF3	Serbian	58.22	2.57	.0000	.223
	English	30.65	1.74	.0000	.131
Cov (F2-F3)	Serbian	5.78	0.55	.0000	.028
	English	4.3	0.53	.0000	.021

To understand the dependence of formant values in the dataset on the language spoken and on the speaker who speaks, we performed two-factor analysis of variance. Table 7-6 contains the summary of the results while the complete ANOVA summary per parameter can be seen in Appendix 9. For each of the measured formants, it can be concluded that the language is a significant factor that affects their values. However, only with the second formant can we state that it is more responsible for formant values than the speaker. Covariance of F2 and F3 does not seem to be susceptible to the language effect, it is more dependent on the speaker. Nonetheless, the F score for covariance is rather low compared to the F scores for formant values.

Table 7-6
Two-factor ANOVA of formant values

Parameter	*Factor*	*F-score*	*p-value*
LTF1	Language	65.84	.0000
	Speaker	581.35	.0000
LTF2	Language	24642.5	.0000
	Speaker	282.4	.0000
LTF3	Language	335.5	.0000
	Speaker	2284.1	.0000
Cov (F2-F3)	Language	0.108	.743
	Speaker	12.092	.0000

Having considered the results above, it can be concluded that, in the present study, LTF2 is most dependant on the language spoken and is highly likely to result in most erroneous speaker comparisons. In order to assess whether the cross-language difference can be neutralised, we derived another parameter – a measure of Frontness, Frontness* (corrected), and Frontness** (double corrected), whereby Frontness is expressed as the difference between F2 and F1 (F2 - F1), while the corrected versions use the following formula for English values:

F2/k – F1; k being the constant that represents the mean ratio of English and Serbian F2 values. For Frontness*, k was derived from 100 replications on randomly sampled data (sample = 20,000, k = 1.21), whereas for Frontness**, k was derived from 100 replications on randomly sampled values summarised per speaker (sample = 10, k = 1.1). The statistical analyses results are summarised in Table 7-7.

Table 7-7
Cross-language comparison of Frontness

Statistics	***Frontness***		***Frontness****		***Frontness****	
mean and SD (English)	1175.75	85.96	869.28	71.98	1015.22	78.58
mean and SD (Serbian)	1028.18	71.87	1028.18	71.87	1028.18	71.87
t-test and p-value	-14.407	.0000	17.178	.0000	1.339	.1866
Cohen's d and power	-2.04	1	2.43	1	0.19	.26
One-way ANOVA (English)	252.46	.0000	244.55	.0000	248.41	.0000
One-way ANOVA (Serbian)	187.95	.0000	187.95	.0000	187.95	.0000
Two-factor ANOVA (Language)	19188.9	.0000	25942.7	.0000	160.5	.0852
Two-factor ANOVA (Speaker)	221.2	.0000	212.6	.0000	216.7	.0000

The results suggest that while the 10-percent fronting that exists for English vowels when spoken by Serbian speakers in the present dataset cannot be entirely removed by dividing the data with the derived constant, it is greatly neutralised. According to the two-factor analysis of variance, in Frontness**, language effect is notably lower than the speaker effect, while the between-speaker variability remains the same as in the dataset without correction. Considering that individual speakers differ by how fronted their English vowels are (2.5% to 16.5%), a certain trade-off between correction and accuracy is expected when performing speaker comparison through likelihood ratio calculations relying on corrected measures.

Phonatory features – spectral tilt, HNR and CPP

As with the articulatory measures, the mean values of phonatory measures presented in Table 7-8 and compared in Table 7-9 are first summarised for each speaker. The reported standard deviation (SD) and *t*-scores are the variation and comparison of summarised means. It should be pointed out, however, that the results presented in Table 7-8 do not give justice to the dataset when looked at face value. Namely, for each speaker, most of these parameters range between positive and negative values, often rendering the standard deviation value higher than the mean (see Figure 7-3).

Table 7-8
Summarised mean and SD of phonatory measures in Serbian and English

Parameter	***Serbian***		***English***	
	mean (summarised)	*SD (mean)*	*mean (summarised)*	*SD (mean)*
H1*-H2*	4.078	1.753	4.123	1.877
H2*-H4*	3.215	1.479	2.839	1.663
H1*-A1*	16.191	2.132	15.349	2.358
H1*-A2*	16.818	3.784	14.867	3.761
H1*-A3*	10.819	4.921	10.482	4.546
H4*-2K*	4.922	2.026	5.585	3.731
CPP	22.508	1.278	22.76	1.115
HNR_{05}	21.681	3.981	23.149	3.846
HNR_{15}	22.815	3.337	24.325	3.347
HNR_{25}	27.345	3.622	27.861	3.588
HNR_{35}	28.734	3.837	28.89	3.815

Observing the mean values of the measured phonatory parameters, we can note a trend that spectral tilt measures are generally higher in Serbian, whereas harmonic-to-noise ratio up to 2.5 kHz is higher in English. The difference can be interpreted to indicate that the speech in Serbian has breathier phonation but is at the same time hoarser, while the speech in English is creakier.

Table 7-9
Paired t-test of summarised phonatory measures across Serbian and English

Parameter	***t-test***	***p-value***	***Cohen's d***	***d range***	***power***
H1*-H2*	-0.3486	.7289	-0.05	-0.33, 0.23	.064
H2*-H4*	2.3183	.0247	0.33	0.04, 0.61	.628
H1*-A1*	4.8154	.0000	0.68	0.37, 0.99	.997
H1*-A2*	7.3863	.0000	1.04	0.7, 1.39	.999
H1*-A3*	1.1915	.2392	0.17	-0.11, 0.45	.218
H4*-2K*	-3.897	.0003	-0.55	-0.85, -0.25	.968
CPP	-4.0322	.0002	-0.57	-0.87, -0.27	.977
HNR_{05}	-7.3109	.0000	-1.03	-1.37, -0.69	.999
HNR_{15}	-6.7971	.0000	-0.96	-1.29, -0.62	.999
HNR_{25}	-2.1612	.0356	-0.31	-0.59, -0.02	.575
HNR_{35}	-0.7436	.4607	-0.11	-0.38, 0.17	.119

As observed in Table 7-9 above, the parameters that do not exhibit any difference across languages are the difference between the amplitude of the first and second harmonic (H1*-H2*), the difference between the amplitude of the first harmonic and the harmonic closest to the third formant (H1*-A3*) and harmonic-to-noise ratio when measured between 0 and 3,500 Hz (HNR_{35}). A weak distinction is detected for the difference between the amplitude of the second and fourth harmonic (H2*-H4*) and harmonic-to-noise ratio when measured

between 0 and 2,500 Hz (HNR_{25}), whereas the rest of the parameters exhibit significant difference across Serbian and English.

The effect size and power of the statistical test are satisfactory; therefore, as opposed to the articulatory parameters, it was not necessary to repeat the analysis with bootstrapping. The exception is H1*-H2*, for which, according to the power score, there is a strong likelihood of Type II error. Average bootstrapped t-score for H1*-H2* in Serbian and English is 0.115, average effect size is 0.006, while average power is now .365, indicating reduced likelihood of Type II error.

Considering that the distribution of phonatory measure values is rather dense and the values summarised across speakers do not properly represent the structure of the data, we provide density distribution of non-summarised values. Figure 7-3 corroborates the statistical analysis in Table 7-9.

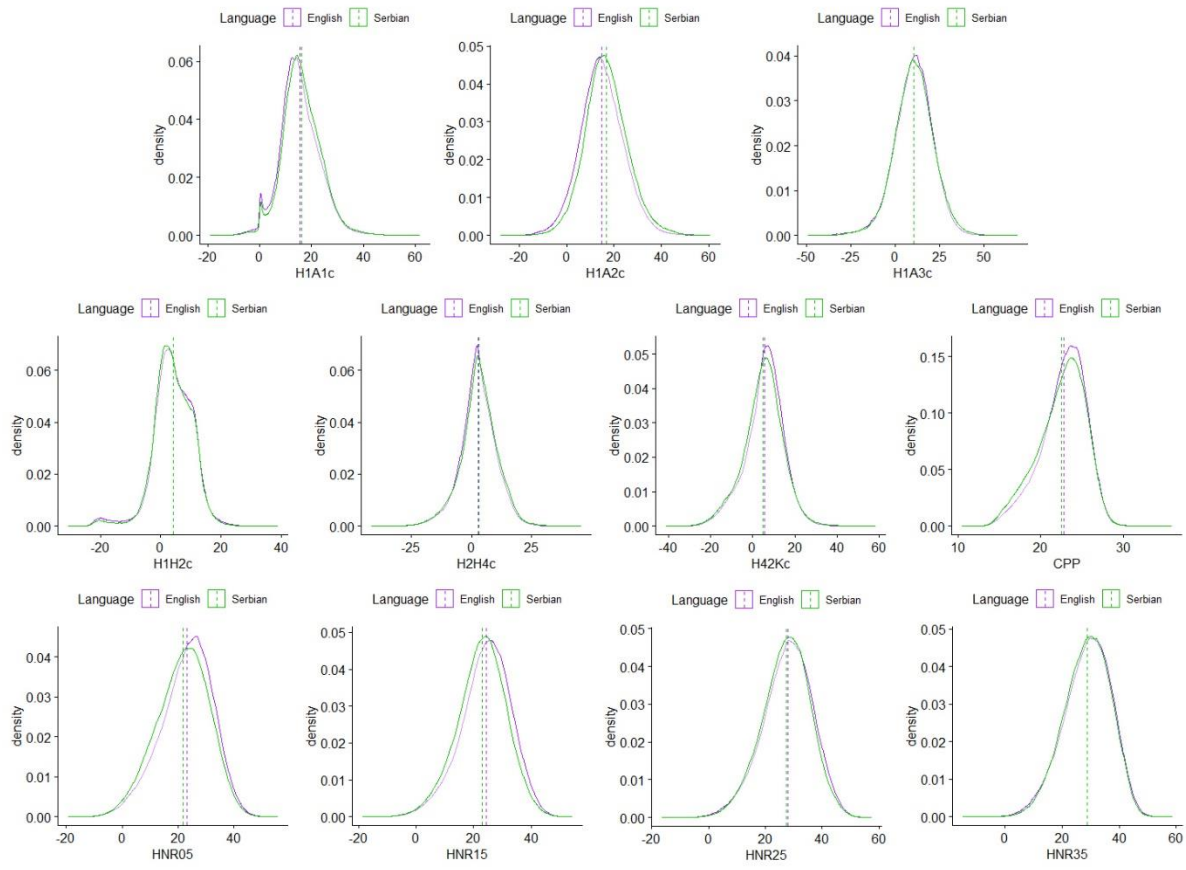


Figure 7-3

Density distribution of phonatory measures across all values

In Figure 7-3 we can observe that spectral tilt measure distributions have almost identical shapes in Serbian and English. In order to closer inspect some of these parameters on individual level, we exported density distribution plots for the first four speakers in the database

(Appendix 10). In the individual speaker plots, we can observe that despite the fact that some of these parameters exhibit no significant difference across languages on population level, the conclusion does not necessarily pertain to individual speakers. For instance, S7 exhibits notably higher HNR_{35} in English than in Serbian. Due to these individual differences, all of the parameters will be tested under the likelihood ratio framework on their own merit.

Next, in order to examine within-speaker and between-speaker variability, we performed one-way analysis of variance for the data in Serbian and English, respectively (Table 7-10). The F-test and p-value indicate that speakers exhibit significant between-group variability with moderate to very strong effect size for all of the measured parameters. The highest F score is observed for harmonic-to-noise ratio up to 3,500 Hz, followed by the difference between the amplitude of the first harmonic and harmonic closest to the third formant (H1*-A3*), cepstral peak prominence and other harmonicity parameters. The lowest F score is noted for the difference in amplitude of higher frequency harmonics (H2*-H4* and H4*-2K*). As with the t-test analysis above, due to strong effect size values, there was no need to repeat the analysis with bootstrapping.

Table 7-10
One-way ANOVA of phonatory measures across speakers

Parameter	*Language*	*df*	*denom df*	*F-test*	*p-value*	η^2
H1*-H2*	Serbian	49	1999950	3368.8	.0000	.08
	English	49	1999950	3453.5	.0000	.08
H2*-H4*	Serbian	49	1999950	1539	.0000	.04
	English	49	1999950	2132.4	.0000	.05
H1*-A1*	Serbian	49	1999950	3563.4	.0000	.08
	English	49	1999950	4312.1	.0000	.1
H1*-A2*	Serbian	49	1999950	8318.3	.0000	.17
	English	49	1999950	8074.7	.0000	.17
H1*-A3*	Serbian	49	1999950	10120	.0000	.20
	English	49	1999950	8678.3	.0000	.18
H4*-2K*	Serbian	49	1999950	1897.8	.0000	.04
	English	49	1999950	1498.8	.0000	.04
CPP	Serbian	49	1999950	9725.6	.0000	.19
	English	49	1999950	8177.7	.0000	.17
HNR_{05}	Serbian	49	1999950	8662.7	.0000	.18
	English	49	1999950	8245.2	.0000	.17
HNR_{15}	Serbian	49	1999950	7133.5	.0000	.15
	English	49	1999950	6950.8	.0000	.15
HNR_{25}	Serbian	49	1999950	8870.8	.0000	.18
	English	49	1999950	8027	.0000	.16
HNR_{35}	Serbian	49	1999950	11676	.0000	.22
	English	49	1999950	10678	.0000	.21

To understand the dependence of spectral tilt and harmonicity parameters on the language and speaker factor, we performed two-factor analysis of variance. In Table 7-11 we can see the summary of the results while the full analysis is available in Appendix 11. The statistical test has confirmed that the language is a significant factor for each of the examined parameters. However, there are a few measures which appear to be more speaker- than language-dependant, including H1*-H2*, H1*-A3*, HNR_{25} and HNR_{35}. Unsurprisingly, language appears to be the strongest factor for H1*-A2*, the parameter derived from the amplitude measure of the second formant.

Table 7-11
Two-way ANOVA of phonatory measures

Parameter	***Factor***	***F-score***	***p-value***
H1*-H2*	Language	52.58	.0000
	Speaker	3413.5	.0000
H2*-H4*	Language	2591	.0000
	Speaker	1822	.0000
H1*-A1*	Language	13845	.0000
	Speaker	3940	.0000
H1*-A2*	Language	54787	.0000
	Speaker	8195	.0000
H1*-A3*	Language	1192	.0000
	Speaker	9401	.0000
H4*-2K*	Language	5273	.0000
	Speaker	1706	.0000
CPP	Language	9908	.0000
	Speaker	8990	.0000
HNR_{05}	Language	29730	.0000
	Speaker	8456	.0000
HNR_{15}	Language	35909	.0000
	Speaker	7041	.0000
HNR_{25}	Language	4320	.0000
	Speaker	8432	.0000
HNR_{35}	Language	466.4	.0000
	Speaker	8432	.0000

Relationships between acoustic parameters

In order to understand relationships between measured parameters, Pearson correlation was performed between each two on data summarised per speaker.

Serbian		LTF1	LTF2	LTF3	Cov(f2-f3)	H1*-H2*	H2*-H4*	H1*-A1*	H1*-A2*	H1*-A3*	H4*-2K*	CPP	HNR05	HNR15	HNR25	HNR35
LTF1	r	-														
	p-value	-														
LTF2	r	0.263	-													
	p-value	0.065	-													
LTF3	r	0.255	0.628	-												
	p-value	0.073	0.000	-												
Cov(f2-f3)	r	0.094	0.219	0.118	-											
	p-value	0.514	0.126	0.413	-											
H1*-H2*	r	0.067	0.204	0.162	0.140	-										
	p-value	0.642	0.154	0.261	0.331	-										
H2*-H4*	r	-0.141	-0.101	-0.241	0.020	-0.239	-									
	p-value	0.327	0.485	0.092	0.890	0.094	-									
H1*-A1*	r	0.211	0.125	0.127	0.193	0.666	0.325	-								
	p-value	0.141	0.385	0.379	0.180	0.000	0.021	-								
H1*-A2*	r	-0.355	-0.163	0.030	-0.036	0.450	0.452	0.599	-							
	p-value	0.012	0.257	0.834	0.802	0.001	0.001	0.000	-							
H1*-A3*	r	-0.499	-0.357	-0.440	-0.122	0.387	0.362	0.428	0.779	-						
	p-value	0.000	0.011	0.001	0.399	0.005	0.009	0.002	0.000	-						
H4*-2K*	r	-0.502	-0.177	-0.197	-0.041	0.247	0.153	0.245	0.726	0.780	-					
	p-value	0.000	0.219	0.171	0.780	0.084	0.290	0.086	0.000	0.000	-					
CPP	r	0.352	-0.004	-0.221	0.095	-0.264	0.112	-0.070	-0.466	-0.357	-0.436	-				
	p-value	0.012	0.979	0.122	0.511	0.064	0.441	0.630	0.001	0.011	0.002	-				
HNR05	r	0.546	0.299	0.094	0.315	0.288	-0.302	0.133	-0.327	-0.340	-0.322	0.058	-			
	p-value	0.000	0.035	0.515	0.026	0.043	0.033	0.357	0.021	0.016	0.023	0.689	-			
HNR15	r	0.131	0.399	0.205	0.352	0.444	-0.165	0.284	0.036	-0.037	-0.022	-0.233	0.833	-		
	p-value	0.366	0.004	0.154	0.012	0.001	0.251	0.046	0.807	0.797	0.881	0.104	0.000	-		
HNR25	r	-0.016	0.284	0.325	0.241	0.533	-0.190	0.346	0.299	0.132	0.184	-0.458	0.649	0.920	-	
	p-value	0.914	0.045	0.021	0.092	0.000	0.186	0.014	0.035	0.360	0.201	0.001	0.000	0.000	-	
HNR35	r	-0.157	0.091	0.136	0.157	0.552	-0.148	0.361	0.427	0.376	0.355	-0.548	0.524	0.832	0.954	-
	p-value	0.277	0.529	0.348	0.277	0.000	0.304	0.010	0.002	0.007	0.011	0.000	0.000	0.000	0.000	-

Significant for p<0.00000001 | Significant for p < 0.01 | Significant for p < 0.05 | Significant for p < 0.1

English		LTF1	LTF2	LTF3	Cov(f2-f3)	H1*-H2*	H2*-H4*	H1*-A1*	H1*-A2*	H1*-A3*	H4*-2K*	CPP	HNR05	HNR15	HNR25	HNR35
LTF1	r	-														
	p-value	-														
LTF2	r	0.219	-													
	p-value	0.127	-													
LTF3	r	0.413	0.626	-												
	p-value	0.003	0.000	-												
Cov(f2-f3)	r	0.141	-0.018	0.034	-											
	p-value	0.329	0.901	0.815	-											
H1*-H2*	r	0.152	0.161	0.138	-0.009	-										
	p-value	0.292	0.264	0.338	0.949	-										
H2*-H4*	r	-0.327	0.006	-0.074	0.039	-0.115	-									
	p-value	0.021	0.969	0.608	0.790	0.427	-									
H1*-A1*	r	0.181	0.128	0.141	0.073	0.680	0.431	-								
	p-value	0.209	0.377	0.328	0.615	0.000	0.002	-								
H1*-A2*	r	-0.267	-0.146	0.027	0.012	0.510	0.491	0.650	-							
	p-value	0.061	0.312	0.852	0.932	0.000	0.000	0.000	-							
H1*-A3*	r	-0.473	-0.239	-0.441	-0.063	0.473	0.390	0.476	0.772	-						
	p-value	0.001	0.095	0.001	0.664	0.001	0.005	0.001	0.000	-						
H4*-2K*	r	-0.398	-0.073	-0.258	0.047	0.142	0.263	0.233	0.657	0.722	-					
	p-value	0.004	0.612	0.070	0.746	0.325	0.065	0.104	0.000	0.000	-					
CPP	r	0.233	-0.079	-0.306	0.153	-0.387	0.123	-0.106	-0.465	-0.353	-0.333	-				
	p-value	0.103	0.584	0.030	0.287	0.005	0.394	0.465	0.001	0.012	0.018	-				
HNR05	r	0.410	0.153	0.136	0.197	0.282	-0.312	0.043	-0.256	-0.295	-0.341	-0.048	-			
	p-value	0.003	0.290	0.345	0.170	0.047	0.027	0.768	0.073	0.038	0.015	0.741	-			
HNR15	r	0.025	0.280	0.171	0.148	0.409	0.000	0.255	0.096	0.040	-0.083	-0.286	0.838	-		
	p-value	0.865	0.049	0.236	0.306	0.003	0.999	0.073	0.508	0.785	0.568	0.044	0.000	-		
HNR25	r	-0.051	0.204	0.264	0.107	0.488	0.033	0.329	0.361	0.182	0.089	-0.478	0.689	0.931	-	
	p-value	0.724	0.156	0.064	0.462	0.000	0.819	0.020	0.010	0.205	0.541	0.001	0.000	0.000	-	
HNR35	r	-0.175	0.036	0.093	0.077	0.541	0.008	0.324	0.472	0.392	0.230	-0.571	0.584	0.854	0.959	-
	p-value	0.224	0.805	0.520	0.594	0.000	0.958	0.022	0.001	0.005	0.109	0.000	0.000	0.000	0.000	-

Significant for p<0.00000001 | Significant for p < 0.01 | Significant for p < 0.05 | Significant for p < 0.1

Figure 7-4

Pearson correlation between parameters (Serbian – top, English – bottom)

In Figure 7-4, for formant values in Serbian, we can notice a slight correlation between LTF1 to LTF2 and LTF3 and a rather strong correlation between LTF2 and LTF3. Such a relationship is expected as in front vowels, the second formant rises, pushing the third formant upward as well. LTF1 correlates with CPP and has negative association with several spectral tilt measures pertaining to formant amplitudes, while all three LTFs correlate negatively with H1*-A3*. All LTF and F2-F3 covariance values correlate with harmonicity in the 2.5 kHz frequency range. Covariance, however, appears to be independent of formant and spectral tilt measures.

As opposed to Serbian, English LTF1 and LTF2 do not correlate, whereas LTF3 correlates with both. Similarly as in the mother tongue, an inversely proportional relationship is detected between all LTFs and H1*-A3*, as well as between LTF3 and CPP. The correlation between the formant values and harmonicity exists in the frequency ranges where each formant is expected, respectively. In English speech. F2-F3 covariance is completely independent of other parameters – there is no correlation with formant, spectral tilt or harmonicity measures.

In both languages, spectral tilt and harmonicity measures exhibit strong correlation among themselves, which is not surprising considering that most of these parameters are derived from the amplitude of the fundamental frequency, which is at the same time the first harmonic. The most prominent is the relationship between H1*-A2* and H1*-A3*, and H1*-A3* and H4*-2K*, observed in both languages (Figure 7-4).

7.1.3. Results – individual speakers

Speaker space and speaker distances

Based on the articulatory and phonatory parameters measured above, we calculated Euclidean distances within speakers across languages and between speakers for each language respectively.

Table 7-12
Euclidean distances based on articulatory and phonatory parameters

	Mean	***SD***	***Min***	***Max***
Within-speaker	5340.71	4062.875	324.2815	18221.61
Between-speaker (English)	8289.718	3581.743	5620.202	20526.66
Between-speaker (Serbian)	9029.343	2594.397	6573.732	18006.96

Table 7-12 provides an overview of the Euclidean distances for the speakers in the present dataset. As the results suggest, within-speaker cross language distances range between 324 to over 18000, which indicates that certain speakers exhibit extremely high divergence in

the measured parameters when they speak English. On the other hand, the lowest within-speaker Euclidean distance is significantly below the lowest between-speaker distances in either language. Upon closer inspection of the speaker with extremely high within-speaker variability it was found that F2-F3 covariance was the only parameter that exhibited large difference across languages (around 14,000 vs 31,000) and removing it would result in a rather low within-speaker distance. Overall, the results indicate a fairly strong speaker-specificity of the cumulative effect of the measured parameters but suggest that likelihood ratio calculations will be prone to errors considering that some speakers exhibit extremely high variability across languages in some parameters.

The correlation statistics has confirmed that speakers with lower between-speaker distances in the mother tongue have lower between-speaker distances in the foreign language as well ($r = .315$, p-value = .026). In addition, speakers who exhibit higher between-speaker distances in the mother tongue also exhibit higher within-speaker distance across languages ($r = .466$, p-value < .001). On the other hand, higher within-speaker divergence across languages does not necessarily imply higher between-speaker divergence in the foreign language, much like it was found for auditory analysis of voice quality above. Figure 7-5 illustrates the correlation of these distances.

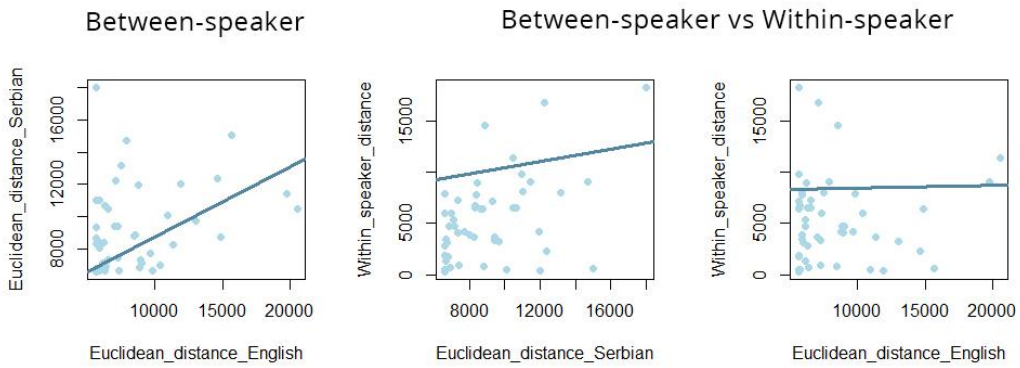


Figure 7-5
Correlation of between- and within-speaker Euclidean distances within and across languages

Next, we performed multidimensional scaling using z-score standardised values of articulatory and phonatory parameters. Figure 7-6 portrays the speaker space in Serbian and English, respectively.

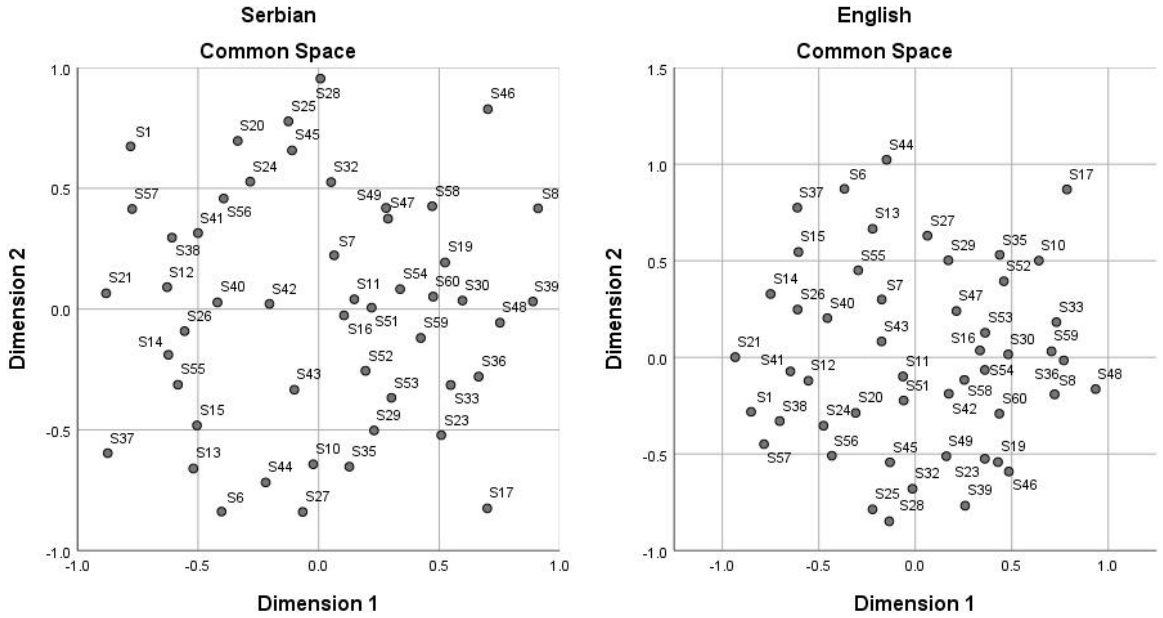


Figure 7-6

Multidimensional scaling of articulatory and phonatory parameters

Looking at Figure 7-6, we can notice that speakers who take peripheral positions in Serbian speaker space (e.g. S1, S17, S37, S44) tend to be on the periphery of English speaker space as well, whereas the ones around the central area of Serbian speaker space (e.g. S7, S11, S16, S51) also draw near the 0 value of English speaker space. Using the multidimensional scaling graph, we should be able to predict which speakers will perform well under likelihood ratio comparisons based on all parameters. Namely, the ones around the periphery of the graph are more distinctive and should therefore be easier to recognise compared to the ones near the central area of the graphs.

Relationships between acoustic and auditory parameters

In sections 7.1.2. and 7.1.3. it was already confirmed that the acoustic analysis corroborates some of the findings reached through expert auditory analysis. The frontness of English vowel space noted in the Vocal Profile Analysis above is confirmed by higher LTF2 values in English. Similarly, more instances of breathy voice for speech in Serbian and more instances of creaky voice for English marked on the VPA charts were corroborated by slightly higher spectral tilt values in Serbian. In the present section, we will focus on the narrow set of 20 participants whose speech was scored by expert listeners on VPA protocol and whose samples were used in the listening experiment with naïve listeners.

To understand the relationships between the auditory and acoustic results, we correlated distance scores obtained in the listening experiment with the experts, recognition

scores from the listening experiment with the naïve listeners, distance scores of acoustic analysis, as well as the values of each parameter. Table 7-13 summarises the most prominent correlations. Prefix EL implies that the parameter was derived from the expert listener experiment, NL – that it originates from the naïve listener experiment and AC means that it is the result of the acoustic analysis.

Table 7-13
Correlation between auditory and acoustic parameters

Parameter	***NL Cont. A correct***		***NL Cont. D correct***		***EL distance within***		***EL distance English***	
	r	*p*	*r*	*p*	*r*	*p*	*r*	*p*
AC distance Sr	-.597	.005			-.428	.06		
AC distance En	-.655	.002			-.379	.099		
AC dist. within			.454	.044				
Covariance Sr	-.470	.037						
HNR15 En							-.624	.003
HNR25 Sr	-.489	.029						
HNR25 En							-.637	.003
HNR35 En							-.637	.003
LTF1 Ee							.501	.024
CPP En							.485	.03

In the experiment with naïve listeners, speakers with the greatest percentage of correct recognition in the same-speaker-same-language setting (Context A) are the ones with lowest average between-speaker distances in both Serbian and English. This is rather counter-intuitive, as one would expect that speakers with greatest distances are more distinctive and therefore more easily recognised. Such a result may imply that the naïve listeners did not rely so much on the parameters that we measured in the acoustic experiment when making their decisions – other aspects of speech must have played a crucial role in naïve listener recognition. In addition, it was found that correct recognition in Context A correlates with low F2-F3 covariance and low harmonicity in the 2.5 kHz domain. This can be interpreted that the speakers with harsher voice were more easily recognised than the ones with a more neutral voice. Furthermore, in different-speaker-different-language context (Context D), it was found that the speakers who are correctly recognised as different most of the time are the ones with highest cross-language distance. As it was already established that the speakers who diverge most from themselves also diverge from others (see 6.1.2), it is not surprising that their voices were easily distinguished in different-language-different-speaker pairs.

According to the within-speaker cross language distance scores based on the expert listening, the speakers with lowest distances across languages are the ones with highest between-speaker distances derived from the acoustic analysis in Serbian and English,

respectively. Furthermore, it was noted that higher between-speaker distance scores for English are associated with higher LTF1 and CPP in this language and lower HNR. The negative correlation between distance scores and harmonicity can be understood to imply that speakers with harsher voice were often rated as more distinct than others.

In accordance with the previous observations, we can conclude that expert listeners appear to have been guided by acoustic aspects of speech when scoring the speakers on the VPA chart, as opposed to naïve listeners, who seem to have observed the voices holistically. As a result, no significant relationships were detected between correct identifications of individual speakers and their acoustic values.

In order to understand whether the measured acoustic parameters contribute to relative judgements by naïve listeners, we calculated the distances between all speaker pairs presented in the Experiment 2 in Context B (different speaker, same language), Context C (same speaker, different language) and Context D (different speaker, different language). Three distance measures were derived for each pair: Euclidean distance based on all of the articulatory and phonatory parameters (D1), Euclidean distance based on long-term formant values, excluding F2-F3 covariance (D2), and Euclidean distance based on phonatory values, but containing only HNR_{35} as the harmonicity measure. (D3). Speaker similarity scores attributed by the naïve listeners and the percentage of "correct", "false" and "not sure" recognitions were correlated with the derived distances. Figures 7-7 provides the summary statistics, while Figure 7-8, 7-9 and 7-10 illustrate these relationships.

Descriptive Statistics (Context B)

	Mean	Std. Deviation	N
Similarity	3.611458699	1.156506785	20
Sim_SD	2.373944975	.3469804222	20
Correct_ans	90.89104812	10.78779832	20
False_ans	5.011348139	8.270530910	20
Not_sure	4.097603741	4.273985268	20
D1	7170.876634	5847.510356	20
D2	235.8820309	122.0411022	20
D3	12.71275855	4.946547125	20

Descriptive Statistics (Context C)

	Mean	Std. Deviation	N
Similarity	8.453757062	.9727748118	20
Sim_SD	1.858445807	.6033065080	20
Correct_ans	75.14689266	18.49382057	20
False_ans	14.00988701	13.03661717	20
Not_sure	10.84322034	6.809135769	20
D1	5217.161464	2377.995527	20
D2	226.2912129	75.57435870	20
D3	4.989414587	2.695650309	20

Descriptive Statistics (Context D)

	Mean	Std. Deviation	N
Similarity	4.429679038	1.559431311	20
Sim_SD	2.419356991	.3645546428	20
Correct_ans	72.78847652	20.30063036	20
False_ans	14.97058251	15.30334276	20
Not_sure	12.24094097	6.944710981	20
D1	8354.233142	5494.083572	20
D2	309.2784474	107.9971612	20
D3	12.70626854	4.847572157	20

Figure 7-7
Summary of distance measures and naive listener scores

Descriptive statistics in Figure 7-7 reveals that, for the same pairs of different speakers, distance based on formant values (D2) notably increases in the cross-language context, whereas distance based on phonatory features (D3) remains the same. On the other hand, for the same set of speakers, in same-speaker-different-language pairs, articulatory-based distance remains the same as for the different-speaker-same-language pairs, whereas phonatory-based distance is evidently lower.

With regard to correlation statistics, in Context B, it was found that the higher the overall distance between pairs, the lower the standard deviation of similarity score is for that pair ($r = -.446$, $p = .049$), implying that listeners generally agreed about the similarity score for those pairs of speakers that were acoustically distinct. Distances, however, did not affect the correct identification percentage. Negative association was also detected for similarity scores and distance calculated from formant values ($r = -.511$, $p = .021$). Put differently, the lower the distance (D2) between speakers, the higher the similarity score. However, the more similar the speaker pair was, the less agreement there was between the listeners regarding their similarity ($r = -.588$, $p = .006$). In addition, there is a weak association between formant-based distances and correct identifications ($r = .406$, $p = .076$), that is, false acceptances ($r = -.443$, $p = .05$). Unambiguously, it can be confirmed that the more distanced two speakers are in speaker space based on their articulatory features, the higher the chance that they will be correctly identified as different speakers by naïve listeners. Distances derived from phonatory measures were not found to correlate with similarity scores or identification percentage in Context B.

Correlations (Context B)

		Similarity	Sim_SD	Correct_ans	False_ans	Not_sure	D1	D2	D3
Similarity	Pearson Correlation	1	.668**	-.960**	.880**	.720**	-.180	-.511*	-.161
	Sig. (2-tailed)		.001	.000	.000	.000	.447	.021	.497
	N	20	20	20	20	20	20	20	20
Sim_SD	Pearson Correlation	.668**	1	-.599**	.553*	.442	-.446*	-.588**	-.003
	Sig. (2-tailed)	.001		.005	.012	.051	.049	.006	.990
	N	20	20	20	20	20	20	20	20
Correct_ans	Pearson Correlation	-.960**	-.599**	1	-.933**	-.718**	.094	.406	.163
	Sig. (2-tailed)	.000	.005		.000	.000	.694	.076	.492
	N	20	20	20	20	20	20	20	20
False_ans	Pearson Correlation	.880**	.553*	-.933**	1	.420	-.077	-.443	-.059
	Sig. (2-tailed)	.000	.012	.000		.065	.748	.050	.804
	N	20	20	20	20	20	20	20	20
Not_sure	Pearson Correlation	.720**	.442	-.718**	.420	1	-.088	-.168	-.297
	Sig. (2-tailed)	.000	.051	.000	.065		.711	.480	.204
	N	20	20	20	20	20	20	20	20
D1	Pearson Correlation	-.180	-.446*	.094	-.077	-.088	1	.170	-.224
	Sig. (2-tailed)	.447	.049	.694	.748	.711		.473	.343
	N	20	20	20	20	20	20	20	20
D2	Pearson Correlation	-.511*	-.588**	.406	-.443	-.168	.170	1	-.111
	Sig. (2-tailed)	.021	.006	.076	.050	.480	.473		.642
	N	20	20	20	20	20	20	20	20
D3	Pearson Correlation	-.161	-.003	.163	-.059	-.297	-.224	-.111	1
	Sig. (2-tailed)	.497	.990	.492	.804	.204	.343	.642	
	N	20	20	20	20	20	20	20	20

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Figure 7-8

Correlation of speaker distances and naïve listener scores, Context B

Formant-based distances appear to have the greatest association with similarity scores and correct identification in Context C as well. Namely, the lower the distance between two speakers, the higher the similarity score ($r = -.572$, $p = .008$), higher general agreement – lower SD ($r = .394$, $p = .085$), higher correct identification percentage ($r = -.602$, $p = .005$), and lower percentage of "not sure" responses ($r = .650$, $p = .002$).

Correlations (Context C)

		Similarity	Sim_SD	Correct_ans	False_ans	Not_sure	D1	D2	D3
Similarity	Pearson Correlation	1	-.873**	.979**	-.965**	-.813**	.344	-.572**	-.087
	Sig. (2-tailed)		.000	.000	.000	.000	.137	.008	.715
	N	20	20	20	20	20	20	20	20
Sim_SD	Pearson Correlation	-.873**	1	-.824**	.789**	.728**	-.383	.394	-.137
	Sig. (2-tailed)	.000		.000	.000	.000	.095	.085	.565
	N	20	20	20	20	20	20	20	20
Correct_ans	Pearson Correlation	.979**	-.824**	1	-.966**	-.867**	.321	-.602**	-.138
	Sig. (2-tailed)	.000	.000		.000	.000	.167	.005	.560
	N	20	20	20	20	20	20	20	20
False_ans	Pearson Correlation	-.965**	.789**	-.966**	1	.708**	-.342	.514*	.122
	Sig. (2-tailed)	.000	.000	.000		.000	.139	.020	.609
	N	20	20	20	20	20	20	20	20
Not_sure	Pearson Correlation	-.813**	.728**	-.867**	.708**	1	-.217	.650**	.143
	Sig. (2-tailed)	.000	.000	.000	.000		.359	.002	.548
	N	20	20	20	20	20	20	20	20
D1	Pearson Correlation	.344	-.383	.321	-.342	-.217	1	-.131	.379
	Sig. (2-tailed)	.137	.095	.167	.139	.359		.581	.100
	N	20	20	20	20	20	20	20	20
D2	Pearson Correlation	-.572**	.394	-.602**	.514*	.650**	-.131	1	.159
	Sig. (2-tailed)	.008	.085	.005	.020	.002	.581		.502
	N	20	20	20	20	20	20	20	20
D3	Pearson Correlation	-.087	-.137	-.138	.122	.143	.379	.159	1
	Sig. (2-tailed)	.715	.565	.560	.609	.548	.100	.502	
	N	20	20	20	20	20	20	20	20

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Figure 7-9

Correlation of speaker distances and naïve listener scores, Context C

On the other hand, in Context D, phonatory-based distances appear to be the ones to affect similarity scores and percentage of correct identifications. The higher the phonatory-based distance of two speakers, the lower the similarity score ($r = -.503$, $p = .024$) and standard deviation ($r = -.525$, $p = .018$). The number of correct identifications also increases with the distance ($r = .516$, $p = .020$), whereas, conversely, the number of false acceptances and "not sure" answers decreases ($r = -.471$, $p = .036$; $r = -.470$, $p = .037$).

Correlations (Context D)

		Similarity	Sim_SD	Correct_ans	False_ans	Not_sure	D1	D2	D3
Similarity	Pearson Correlation	1	.640**	-.976**	.927**	.810**	-.340	-.119	-.503*
	Sig. (2-tailed)		.002	.000	.000	.000	.142	.618	.024
	N	20	20	20	20	20	20	20	20
Sim_SD	Pearson Correlation	.640**	1	-.510*	.423	.558*	-.407	-.109	-.525*
	Sig. (2-tailed)	.002		.022	.063	.011	.075	.646	.018
	N	20	20	20	20	20	20	20	20
Correct_ans	Pearson Correlation	-.976**	-.510*	1	-.963**	-.802**	.226	.097	.516*
	Sig. (2-tailed)	.000	.022		.000	.000	.338	.683	.020
	N	20	20	20	20	20	20	20	20
False_ans	Pearson Correlation	.927**	.423	-.963**	1	.610**	-.175	-.097	-.471*
	Sig. (2-tailed)	.000	.063	.000		.004	.461	.684	.036
	N	20	20	20	20	20	20	20	20
Not_sure	Pearson Correlation	.810**	.558*	-.802**	.610**	1	-.276	-.070	-.470*
	Sig. (2-tailed)	.000	.011	.000	.004		.238	.768	.037
	N	20	20	20	20	20	20	20	20
D1	Pearson Correlation	-.340	-.407	.226	-.175	-.276	1	.285	.068
	Sig. (2-tailed)	.142	.075	.338	.461	.238		.223	.775
	N	20	20	20	20	20	20	20	20
D2	Pearson Correlation	-.119	-.109	.097	-.097	-.070	.285	1	-.206
	Sig. (2-tailed)	.618	.646	.683	.684	.768	.223		.382
	N	20	20	20	20	20	20	20	20
D3	Pearson Correlation	-.503*	-.525*	.516*	-.471*	-.470*	.068	-.206	1
	Sig. (2-tailed)	.024	.018	.020	.036	.037	.775	.382	
	N	20	20	20	20	20	20	20	20

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Figure 7-10

Correlation of speaker distances and naïve listener scores, Context D

Relationships between acoustic parameters and language proficiency

To study the relationship between foreign language proficiency/fluency and acoustic aspects of pronunciation, we correlated the relevant proficiency scores with the derived distance measures as well as with the values of the acoustic parameters.

It was found that lower fluency is in negative correlation with between-speaker distances derived from all parameters in both Serbian and English, respectively ($r = -.308$, $p = .029$; $r = -.274$, $p = .054$). Fluency was also found to be reversely associated with English between-speaker distances derived from phonation parameters ($r = -.311$, $p = .028$). The results indicate that, at least in the present dataset, the less fluent the speakers are the more distinctive they are. Therefore, we can predict that, among the selected speakers, the ones with lower proficiency (fluency in particular) will potentially perform better under the likelihood ratio framework. No correlations were detected between within-speaker distances and language

proficiency parameters. Also, there does not seem to be any relationship between the acoustic values of the measured parameters in English and relevant fluency/pronunciation scores.

To understand the importance of fluency for a speaker's difference across languages, for each parameter we derived a value that represents a difference between its value in Serbian and its value in English ($x_{dif} = x_{sr} - x_{en}$). The detected relationships are presented in Table 7-14:

Table 7-14
Correlation between fluency and cross-language variability of acoustic parameters

		LTF1_dif	*LTF3_dif*	*HNR15_dif*	*HNR25_dif*	*HNR35_dif*
F/C	r	-.303			.261	.249
	p	.032			.068	.081
PRON	r	-.369	.280	.316	.387	.358
	p	.008	.049	.025	.006	.011
Band	r	-.345		.248	.336	.309
	p	.014		.083	.017	.029

Higher fluency and pronunciation scores, as well as the higher final score are associated with lower difference in LTF1. Since we subtracted the English value from the Serbian, we can conclude that the speakers who increase LTF1 when speaking English are also the ones scored better for fluency and pronunciation. On the other hand, better pronunciation is associated with higher difference in LTF3, meaning that, in the present dataset, the speakers who exhibit lower LTF3 in English are often scored higher for pronunciation. With regard to the phonatory parameters, higher fluency and pronunciation scores seem to be associated with an increased difference in HNR parameters, or, lower harmonicity in English.

7.1.4. Discussion

In the present study, we performed the acoustic analysis of the parameters associated with both articulation and phonation. Considering that formant values are seen as acoustic correlates of articulatory settings, we measured long-term F1, long term F2, long-term F3 and covariance of F2-F3. The statistical analysis has confirmed that speakers exhibit higher F2 in foreign English than in native Serbian, which indicates greater degree of fronting in the foreign language. The fronting in English was also detected by expert listeners on the VPA protocol and reflected in the advanced tongue tip/blade setting (compare Table 6-7 and Table 6-8). That fronting is more prominent in English than in Serbian for the analysed speakers is confirmed by another measure obtained through subtraction of the first and second formant. These two measures (LTF2 and F2-F1) are, at the same time, the parameters for which the

language effect is the strongest. Based on the present dataset, statistical analyses have revealed no language effect for F2-F3 covariance. On the other hand, even though LTF1 and LTF3 also seem to depend on the language spoken, for these parameters, between-speaker variability is significantly more prominent than within-speaker variability across languages, which is in accordance with the results of Hereen et al. (2015) and Tomić (2020).

The measures of phonation examined in the present study include spectral tilt measures (H1*H2*, H2*-H4*, H1*-A1*, H1*-A2*, H1*-A3*, H4*-2K*), cepstral peak prominence (CPP) and harmonicity measures (HNR_{05}, HNR_{15}, HNR_{25}, HNR_{35}). Even though there is not a large discrepancy of the Serbian and English values of the measured parameters, the results indicate that the recorded speakers have higher spectral tilt but lower harmonicity in Serbian. This can be interpreted that the recorded speakers in general have breathier but hoarser phonation in their mother tongue than in the foreign language, which is corroborated by the auditory analysis results where there were more instances of creaky voice and fewer instances of breathy voice in English than in Serbian (compare Table 6-7 and Table 6-8). The phonation parameters that emerge as more speaker dependant than language dependant include H1*-H2*, H1*-A3*, HNR_{25} and HNR_{35}, whereas those for which the language effect is the strongest are H1*-A2*, HNR_{05} and HNR_{15}.

Correlation between the measured parameters in English mostly reflects the relationships that exist in the mother tongue as well. In Serbian, however, the three LTFs are mutually correlated, while in the foreign language, the relationship is detected only for the higher formants. Covariance of F2-F3 emerges as a rather independent parameter, whereas spectral tilt and harmonicity measures are inter-related to different degrees. In calculation of the overall likelihood ratio, HNR of up to 3.5 kHz appears to be the most appropriate parameter to corroborate formant-based analysis as it does not correlate with formant values in either language.

Calculation of distances between speakers in speaker space based on all of the measured parameters has revealed that within-speaker cross-language distances are notably lower than between-speaker distances in either language. In addition, speakers who take up the periphery of the designated speaker space in one language are likely to be there in the other as well. Covariance appears to be the least predictable parameter as with certain speakers it exhibits extremely large variance across languages, which is why it was excluded from distance calculations based solely on articulatory measures.

Upon examining the relationships between the results obtained through auditory experiments and acoustic analysis, we could observe that expert listeners and naïve listeners

rely on acoustic cues to a different extent. Namely, naïve listeners do not necessarily rely on acoustic cues for successful identification of a particular speakers, however, when the samples are presented in a pair, acoustic cues become relevant for recognition. It was found, in all contexts, that the more acoustically distinct the voices are, the stronger is the agreement among listeners about the degree of similarity of the voices. Conversely, for acoustically close voices, the listeners exhibit higher variability in their similarity scores. Furthermore, in the different-speaker same-language context and in the same-speaker different language context, the percentage of correctly performed recognitions is associated with the increased articulatory-based distances between the presented voices. On the other hand, in the different-speaker different-language context, successful rejection is associated with the higher phonatory-based distances. Such results raise some questions concerning neurological voice processing. There might be several explanations why in the contexts with one differing factor (either speaker or language), the articulatory cues play a significant role whereas in the context with two differing factors (both speaker and the language), the relationship is detected for phonatory features. It is already known that the "decisions" about language and speaker identity do not happen sequentially but occur in parallel, because the acoustic cues underlying the perception of linguistic and indexical information are the same (Foulkes, 2010; Geers et al., 2013; Redford & Baese-Berk, 2023). However, there are no studies that examine which cues take priority for the "decision" and whether these vary across different contexts. It is already known that familiar and unfamiliar voice processing occurs in separate brain regions (Maguinness et al., 2018; Stevenage, 2018) and by analogy, one of the possible explanations is that different centres in the brain are activated depending on the type of mismatch. Another possibility is that the brain enters a sort of sequential decision-making process, whereby (1) the brain realises the samples are in a different language, (2) it attempts to access articulatory information first and if they are similar (as with same speakers), it makes a decision, however, (3) if the process fails because they are incomparable, the brain proceeds to the phonatory information

Finally, we examined the relationship between the acoustic parameters and language proficiency, that is, fluency. The analysis has revealed that, in the present dataset, low fluency scores are attributed to those speakers with higher between-speaker distances in both Serbian and English, respectively. The results should be observed to reflect the structure of our, quite homogenous, dataset, rather than a general trend. In addition, it was found that the speakers who increase LTF1 when speaking English are also the ones scored better for fluency and pronunciation. Although this relationship is not necessarily causal, it is logical, considering that that speakers who produce their vowels as more open when they speak English are closer

to the native pronunciation of English, which has a wider vowel space and a greater number of open vowels than Serbian (see Figure 4-1 and Figure 4-2). Similarly, we have shown that the speakers who exhibit lower LTF3 in English are scored higher for pronunciation. As the third formant is often associated with lip rounding, we may conclude that the speakers who exhibit higher lip-rounding are perceived as more proficient in English. With regard to the phonatory parameters, higher fluency and pronunciation scores seem to be associated with an increased difference in HNR parameters, or, lower harmonicity in English. This relationship is difficult to interpret as there are no studies that comparatively examine harmonicity of voice in native English and native Serbian. One of the possible explanations is that, by default, the English language is characterised by lower periodicity than Serbian and that is why these speakers are scored as more proficient in pronunciation. Another explanation could be that there is a direct relationship between fluency and harmonicity.

7.2. Calculation of Likelihood Ratio

7.2.1. Likelihood ratio measurements

For estimation of system performance of the measured parameters and their combination, we employed two models of likelihood ratio estimation, the Gaussian Mixture Model-Universal Background Model (GMM-UBM) (Reynolds et al., 2000) and multi-variate kernel density (MVKD) likelihood ratio (Aitken & Lucy, 2004). Both models have their advantages and have been employed in forensic phonetic research (see Gold, 2012; 2014; Gold et al., 2013; Holmes, 2023; Hughes et al., 2017b; Kinoshita, 2001; 2014; Kinoshita & Ishihara, 2014; Lo, 2021; Holmes, 2024; Rose, 2013; Rose & Wang, 2016; Tomić, 2014; Tomić & French, 2019), while the MVKD model has found its way into forensic casework as well (see Rose, 2022). Likelihood ratio calculations were performed with "fvclrr" package (Lo, 2022) with modifications to enable calculations using MVKD formula. The package relies on "mclust" (Scrucca et al., 2023) for density estimation.

Following Lo's (2021) implementation of GMM-UBM likelihood ratio, the dataset was randomly split into three groups (test speakers, training speakers and background speakers) and log-likelihood ratios were calculated for each same-speaker (SS) and different-speaker (DS) pair in the test data and training data. The scores obtained through these calculations were then calibrated using calibration-fusion, in a similar way it is used for combining LR scores of multiple features. Cross-language speaker comparison was performed under three conditions: (a) reference data in Serbian, (b) reference data in English, (c) reference data comprising both Serbian and English measurements. The GMM-UBM model is advantageous over the MVKD model because it can accept large sets of raw data without any prior averaging, and employment of the training set ensures well-calibrated scores. The disadvantage, however, is that in order to obtain representative LR scores, datasets with a large number of speakers need to be fed into the formula. In the present study, GMM-UBM LR measurements yielded 16 same-speaker and 240 different-speaker comparisons, which is a rather small set for drawing general conclusions. A solution that scientists usually employ to neutralise the sampling effect is replication with repeated drawing from the same data, however, regardless of the number of replications, both the test and background set remain small (16-17 speakers), which affects the LR scores across all replications.

Due to the reasons above, in the present study, the focus is on the likelihood ratio calculations using the multivariate kernel density model. Considering that the formula falters and outputs C_{llr} scores much higher than 1 when data with a lot of points is used, it is necessary

to average the obtained measurements over fixed amount of time prior to comparison (see Gold, 2014; Tomić & French, 2019). Therefore, we performed two sets of calculations, averaging the measurements across (1) one second of speech and (2) across two seconds of speech. For different databases this implied a different number of data points (100 or 200 formant measurements and 1,000 or 2,000 phonation measurements). The comparisons were performed using leave-one-out cross-validation method, whereby, for each comparison, the background population was comprised of all the speakers in the dataset apart from the ones who are being compared, yielding 50 SS and 2,450 DS comparisons. Data for the reference (background) speakers in same-language comparisons was drawn from the measurements for that language, whereas, like with the GMM-UBM model, cross-language speaker comparison was performed under three conditions: (a) reference data in Serbian, (b) reference data in English, (c) reference data comprising both Serbian and English measurements. Overall or true likelihood ratio scores for combination of features were produced through calibration-fusion (Morrison, 2013).

Selection of appropriate reference (background) population always poses a challenge in speaker comparison, especially when the samples are recorded under mismatched condition. Watt et al. (2020) assessed the effects of accent-mismatched reference population on the performance of an ASR system. When using good-quality, contemporaneous samples, the ASR system is able to successfully separate same- and different-speaker pairs irrespective of the reference data used to assess typicality. Accent mismatch between the questioned and reference samples, however, produced scores that were more poorly calibrated than those where the accent was closely matched. Furthermore, accent mismatch produced much stronger same-speaker evidence. Bearing in mind the structure of the LR system, it is expected that, in the present study, the calculations with the background data comprised of measurements in both languages would yield best-calibrated LLR scores; however, as explained above, in forensic reality, such a condition would be difficult to reproduce since there might not be bilingual databases for the languages that are being compared.

Finally, we were interested in individual speaker performance within the system and the relationship of their performance in cross-language comparisons to foreign language proficiency. For same-speaker comparisons, the measure of individual performance was the log-likelihood ratio (LLR) score (the higher the score, the better the performance), whereas for different-speaker measurements we relied on average LLR and error rates (ER). Average LLR was derived from all DS comparisons for the particular parameter (the lower the average LLR, the better the performance), while the ER was calculated as a percentage of false positive identifications of that particular speaker for the parameter in question.

The results obtained through two LR models are compared below. Individual speaker performance was assessed only with the MVKD results averaged over 2 s of speech.

7.2.2. Results – EER and C_{llr} overview

Complete results for both LR models for all the measured and combined parameters across all tested conditions are available in Appendix 12. Table 7-15 below provides the summary of EER and C_{llr} scores.

Table 7-15
Range of EER and Cllr scores across conditions and LR models for 23 parameters

Model/ Parameter		***GMM-UBM***		***MVKD_1***		***MVKD_2***	
		EER	*C_{llr}*	*EER*	*C_{llr}*	*EER*	*C_{llr}*
Serbian	min	6.25%	0.26	1.63%	0.05	1.67%	0.05
	max	30%	0.84	34%	1	34%	0.97
English	min	7.5%	0.35	2.26%	0.09	2.22%	0.09
	max	31%	0.81	40%	1.08	40%	1.05
Cross-language (mixed bckg)	min	12.5%	0.5	8.2%	0.3	8%	0.3
	max	43%	1.09	50%	3.36	50%	2.85
Cross-language (Serbian bckg)	min	17.9%	0.54	3.88%	0.14	3.88%	0.14
	max	37.7%	1.08	42%	3.37	42%	2.95
Cross-language (English bckg)	min	13%	0.53	3.9%	0.15	3.9%	0.15
	max	37.5%	1.1	40%	2.77	42%	2.43

Across all conditions, both the equal error rates and C_{llr} scores of multivariate kernel-density LR model appear to have a wider range than those of the GMM-UBM model. Whereas the error rates do not change depending on the length of period for which the data was averaged in MVKD model (Figure 7-11), averaging the values across 2 seconds of speech seems to reduce the range of C_{llr} scores.

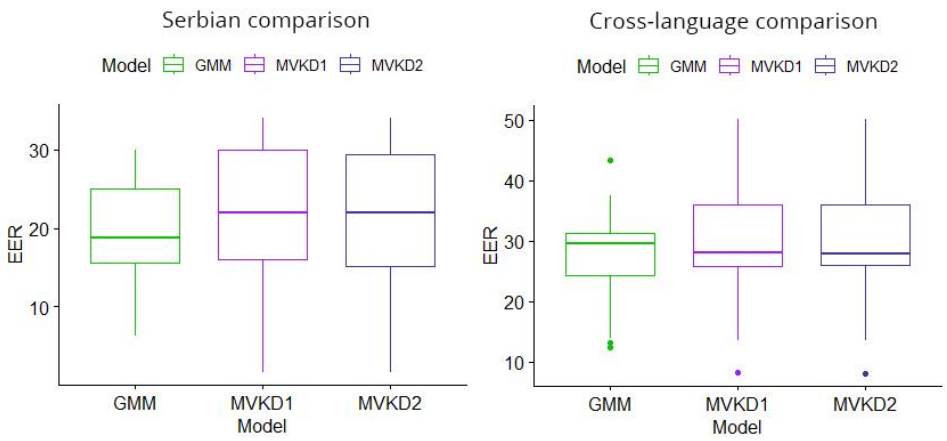


Figure 7-11
EER scores of 23 parameters across LR models

In cross language comparison, irrespective of the background population used, MVKD LR provides lower error rates; however, non-calibrated C_{llr} scores are much higher than for the GMM-UBM model. Extremely high cross-language comparison C_{llr} scores in MVKD likelihood ratio can be brought below 1 with an additional step of linear regression calibration (Figure 7- 12), as will be demonstrated in Section 7.2.5 for individual parameters.

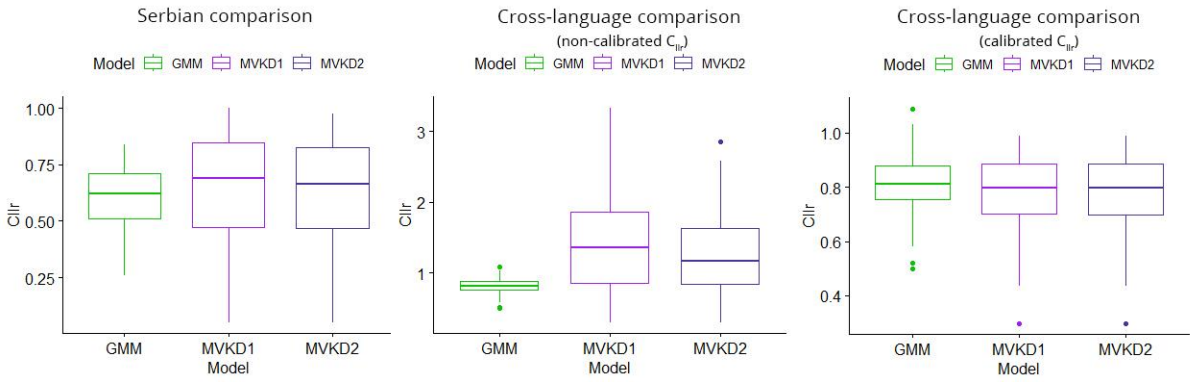


Figure 7-12
Cllr scores of 23 parameters across LR models

Looking at the overall picture, it is difficult to determine which model provides a better option for data assessment, inasmuch as the error rates for individual parameters differ across models (see Appendix 12). For instance, LTF1 and LTF2 appear to be much stronger discriminants when assessed with the GMM-UBM than with the MVKD model. Conversely, LTF3 performs better under the latter. Similarly, most of the spectral tilt parameters have better performance in same-language comparisons under the GMM-UBM model, whereas harmonicity parameters have lower EER under the MVKD formula. Notwithstanding, as explained in Section 7.2.1., in the present study, considering a small number of speakers assigned to each dataset (test, training and background), the focus will be on the results obtained through the cross-validated MVKD likelihood ratio formula. A more in-depth analysis of the performance of individual parameters across conditions will be given in the following sections. Considering that slightly better C_{llr} scores are obtained when the data is averaged across 2 seconds of speech, only MVKD2 results will be explored further.

7.2.3. Results – single-language comparison

Table 7-16 lists the equal error rate and C_{llr} scores of the examined parameters in single-language speaker comparisons – when both the questioned and the known sample are either in Serbian or English. While at a first glance it may seem that most of the parameters

perform better when both samples are in the mother tongue than when they are in the foreign language, the statistical analysis has not confirmed the observation (t = -0.654, p = .5165).

In the Serbian language comparisons, error rates for the selected parameters range between 18% and 34%, whereby among the best performing parameters are phonatory measures such as H1*-A2*, H1*-A3*, CPP and the articulatory measure: long-term F3, whereas the highest error rates are observed for F2-related measures (LTF2, Covariance, Frontness) and for the phonatory measure H4*-2K*. In the foreign language, the error rates for F2-related parameters notably increase. Such results are not surprising for the present data, considering that LTF2 has proven to exhibit the greatest difference across languages (see Section 7.1.2.). Harmonic-to-noise ratio in the range of 3.5 kHz and cepstral peak prominence emerge as the best performing parameters for comparisons in English.

Table 7-16
LR performance of individual parameters in single language comparisons (EER and C_{llr})

Parameter	***Serbian***		***English***	
	EER (%)	C_{llr}	*EER (%)*	C_{llr}
LTF1	30.18	0.85	31.98	1.05
LTF2	32.37	0.8	38	0.9
LTF3	20.18	0.67	24.49	0.71
Covariance	32.33	0.97	31.94	0.9
Frontness	33.98	0.84	40.08	0.96
Frontness*	33.98	0.84	39.98	0.96
H1*-H2	25.98	0.67	26	0.79
H2*-H4	28.43	0.87	30.06	0.74
H1*-A1*	21.98	0.61	20.18	0.62
H1*-A2*	17.98	0.51	20.98	0.61
H1*-A3*	18.08	0.51	21.94	0.7
H4*2K*	33.98	0.79	32.31	0.88
CPP	18.06	0.56	19.9	0.57
HNR_{05}	19.96	0.62	23.73	0.72
HNR_{15}	23.76	0.77	24.04	0.68
HNR_{25}	22.18	0.87	21.86	0.66
HNR_{35}	22.02	0.82	18.45	0.61

To obtain the overall likelihood ratio, we combined a selection of features through calibration-fusion, which resulted in reduced error rates and C_{llr} scores for both the comparisons in the mother tongue and the foreign language (Table 7-17). The combined power of the first three formants yielded an equal error rate of around 12% for the comparisons in Serbian and 18% for comparisons in English. The inclusion of F2-F3 covariance along with the formant measures did not improve the performance of the system for comparisons in the mother tongue,

whereas for comparisons in the foreign language the error rate decreased for more than 3% with addition of this parameter.

Table 7-17
LR performance of the combination of parameters in single language comparisons (EER and C_{llr})

Parameter	***Serbian***		***English***	
	EER (%)	*C_{llr}*	*EER (%)*	*C_{llr}*
Formants	12.02	0.38	18	0.53
Formants + Cov	12.02	0.37	14.73	0.48
Phonation - all	1.67	0.05	2.22	0.09
Cov, H1*-H2*, HNR_{35}	12.45	0.43	13.98	0.39
Cov, H1*-A3*, HNR_{35}	9.98	0.31	10.41	0.33
F3, H1*-A3*, HNR_{35}	4.1	0.15	5.76	0.2
Articulation + Phonation	0	0.0000	0.16	0.03

An impressive error rate of 1.67% and C_{llr} of 0.05 was achieved for combination of all recorded phonatory features. The fact that most of these features were found to correlate with one another should not pose a problem for the LR calculation through the multivariate kernel density formula as it is devised to account for correlations that may exist between parameters. Notwithstanding, to be on the safe side, we selected three parameters (one articulatory, one related to spectral tilt and one to harmonicity) which do not exhibit any correlation and performed calibration-fusion. The best score was noted for LTF3, H1*-A3*, and HNR_{35}, which in combination yielded an EER of around 4% and C_{llr} of 0.15 for the comparisons in Serbian and a slightly higher EER of 5.76% and C_{llr} of 0.2% for comparisons in English. Even though the performance of the selected features in combination is rather good for the samples in the foreign language, on average, error rates appear to be slightly higher than in the mother tongue.

7.2.4. Results – cross-language comparison

Table 7-18 lists the EER and C_{llr} scores for cross-language comparisons in three conditions – background population comprised of values in Serbian and English, background population in Serbian and background population in English. According to our results, the error rates between the comparisons with the background population matching either the questioned or the suspect sample do not differ ($t = -0.09$, $p = .93$), whereas the error rates obtained with the background population comprised of bilingual data are slightly higher than both ($t = 1.885$, $p = .066$; $t = 1.76$, $p = .085$). The potential reasons for this observation will be discussed in Section 7.2.7.

For most of the parameters in cross-language comparisons, C_{llr} scores are higher than 1, indicating a poorly calibrated system. The performance of the system can be improved by implementing logistic regression calibration on a single parameter without affecting the error rates (compare Table 7-18 and Table 7-20).

Table 7-18
LR performance of individual parameters in cross-language comparisons with different background populations (EER and C_{llr})

Parameter	***Serbian + English***		***Serbian***		***English***	
	EER (%)	C_{llr}	*EER (%)*	C_{llr}	*EER (%)*	C_{llr}
LTF1	36.02	1.27	31.2	1.14	30.04	1.17
LTF2	49.96	2.58	38.04	2.65	39.84	2.43
LTF3	27.63	1.99	22	2.23	22.1	1.41
Covariance	40.16	0.97	39.49	1.11	39.8	1.06
Frontness	50.27	2.38	38	2.18	40.37	2.15
Frontness*	40.02	1.01	42.02	1.04	42.08	1.05
H1*-H2	26.06	0.99	24.12	0.97	24.71	0.89
H2*-H4	38.31	1.16	34.06	1.46	33.98	1.29
H1*-A1*	32.55	1.8	28.04	1.44	28.04	1.51
H1*-A2*	36.02	2.85	25.63	2.95	26.04	2.32
H1*-A3*	28	1.19	23.51	1.13	23.53	1.05
H4*2K*	32.16	1.29	32.88	1.13	34.35	1.06
CPP	25.86	1.15	20.08	1.22	20.08	1.31
HNR_{05}	27.94	1.17	27.53	1.49	27.59	1.59
HNR_{15}	33.67	1.78	26.31	2	26.27	2.03
HNR_{25}	30	1.5	21.96	1.38	22.08	1.41
HNR_{35}	28	1.39	20.45	1.14	21.53	1.16

Performing calibration-fusion on LLR scores obtained in cross-language comparisons significantly reduces the C_{llr} scores, rendering the system well-calibrated. Error rates for combinations of features are significantly lower if the background population is comprised of monolingual data (see Table 7-19). However, performing logistic-regression calibration on individual parameters prior to fusing them to obtain the overall likelihood ratio does not additionally improve the performance of the system (compare Table 7-19 and Table 7-21).

Table 7-19
LR performance of the combination of parameters in cross-language comparisons with different background populations (EER and C_{llr})

Parameter	***Serbian + English***		***Serbian***		***English***	
	EER (%)	C_{llr}	*EER (%)*	C_{llr}	*EER (%)*	C_{llr}
Formants	26.02	0.72	12.51	0.46	13.47	0.47
Formants + Cov	20.02	0.68	12	0.42	12.1	0.43
Phonation - all	8.06	0.3	3.88	0.14	3.92	0.15
Cov, H1*-H2*, HNR_{35}	19.94	0.55	11.84	0.37	11.8	0.37
Cov, H1*-A3*, HNR_{35}	17.96	0.5	11.49	0.33	10.27	0.33
*f*3, H1*-A3*, HNR_{35}	13.57	0.43	6.1	0.24	6.14	0.24

According to the calibrated scores in Table 7-20, the best performing articulation-related parameter across all conditions is LTF3 with an EER of 22% with monolingual background population (27% with combined population) and a C_{llr} of 0.66 (or 0.8). Similar to single-language comparisons, the highest error rates are observed for all of the F2-related parameters, whereby the performance of the system is almost equal to chance. Among the best performing parameters are phonatory measures such as CPP, HNR_{35}, HNR_{25}, H1*-A3*, and H1*H2*, with EER ranging between 20%-24% in comparisons with monolingual background data and 26%-30% in comparisons with bilingual background data.

Table 7-20
Calibrated C_{llr} scores of individual parameters in cross-language comparisons with different background populations

Parameter	***Serbian + English***		***Serbian***		***English***	
	EER (%)	C_{llr}	*EER (%)*	C_{llr}	*EER (%)*	C_{llr}
LTF1	36.02	0.88	31.2	0.78	30.04	0.78
LTF2	49.96	0.98	38.04	0.87	39.84	0.88
LTF3	27.63	0.8	22	0.66	22.1	0.66
Covariance	40.16	0.91	39.49	0.88	39.8	0.88
Frontness	50.27	0.99	38	0.88	40.37	0.9
Frontness*	40.02	0.95	42.02	0.89	42.08	0.89
H1*-H2	26.06	0.8	24.12	0.69	24.71	0.69
H2*-H4	38.31	0.89	34.06	0.82	33.98	0.82
H1*-A1*	32.55	0.87	28.04	0.74	28.04	0.74
H1*-A2*	36.02	0.86	25.63	0.74	26.04	0.74
H1*-A3*	28	0.72	23.51	0.61	23.53	0.61
H4*2K*	32.16	0.94	32.88	0.83	34.35	0.83
CPP	25.86	0.67	20.08	0.59	20.08	0.58
HNR_{05}	27.94	0.72	27.53	0.65	27.59	0.65
HNR_{15}	33.67	0.82	26.31	0.72	26.27	0.72
HNR_{25}	30	0.78	21.96	0.65	22.08	0.65
HNR_{35}	28	0.72	20.45	0.58	21.53	0.58

Table 7-21
Calibrated C_{llr} scores of the combined parameters in cross-language comparisons with different background populations

Parameter	***Serbian + English***		***Serbian***		***English***	
	EER (%)	C_{llr}	*EER (%)*	C_{llr}	*EER (%)*	C_{llr}
Formants	26.02	0.72	12.51	0.46	13.47	0.47
Formants + Cov	20.02	0.68	12	0.42	12.1	0.43
Phonation - all	8.06	0.3	3.88	0.14	3.92	0.15
Cov, H1*-H2*, HNR_{35}	19.94	0.55	11.84	0.37	11.8	0.37
Cov, H1*-A3*, HNR_{35}	17.96	0.5	11.49	0.33	10.27	0.33
F3, H1*-A3*, HNR_{35}	13.57	0.43	6.1	0.24	6.14	0.24

The combined power of the first three formants yielded an equal error rate of around 13% for the comparisons with reference population in either Serbian or English and twice as high EER (26%) for comparisons with reference population comprised of bilingual data. The inclusion of F2-F3 covariance along with the formant measures only slightly improved the performance of the system for the latter condition. Not unlike single-language comparisons, the best equal error rate (around 4%) and C_{llr} scores (0.14) are achieved for combination of all recorded phonatory features. In addition, LTF3, H1*-A3* and HNR_{35} in combination yielded an EER of around 6% and C_{llr} of 0.24 for the comparisons with background data in Serbian or English and twice as high EER of 13.57% and C_{llr} of 0.43% for comparisons with combined background data. Even though the performance of the selected features in combination is rather good when the combined background data is used, error rates and C_{llr} scores are significantly better when monolingual background population is used as reference.

To understand how many parameters are enough for speaker characterisation in the cross-language speaker comparison, we performed post hoc combinations of parameters for the calibrated scores with the background population derived of Serbian data. The results are available in Table 7-22.

Table 7-22
Post-hoc combination of parameters in cross-language comparisons with background population in Serbian (EER and C_{llr})

Parameters	***EER (%)***	***C_{llr}***
LTF1 + LTF3	17.76	0.52
LTF3 + CPP + HNR_{35}	8	0.24
LTF3 + CPP + H1*-A3* + HNR_{35}	4	0.17
LTF3 + CPP + H1*H2* + H1*-A3* + HNR_{35}	3.55	0.13
LTF1 + LTF2 + LTF3 + CPP + H1*H2* + H1*-A3* + HNR_{35}	2	<0.1
Articulation + Phonation	1.59%	0.06

The post-hoc combination of parameters reveals that despite its poor individual performance, LTF2 is crucial in speaker characterisation and its combination with LTF1 and LTF3 contributes to improved EER and C_{llr} in cross-language speaker comparisons. On the other hand, combination of all phonatory features extracted in the present study may be redundant as similar scores are achieved with combination of CPP, H1*-A3* and HNR_{35} alone. Finally, combination of the three long-term formants with the above-mentioned phonation measures, brings the system performance to an impressive EER of 2% and C_{llr} of 0.999.

7.2.5. Estimation of language-proficiency effect

To understand the relationship between individual speaker performance within the LR system we derived three individual-speaker measures for each parameter respectively: LLR_{SS} (same-speaker log-likelihood ratio), LLR_{DS} (average different-speaker log-likelihood ratio) and ER_{DS} (percentage of false positive identifications). The obtained measures were then correlated using Pearson correlation with speakers' proficiency scores to determine if there is any relationship between them. The condition for which proficiency relationships were examined is cross-language speaker comparison under multivariate kernel density formula for samples averaged over 2 seconds of speech with background population data derived from the measures in Serbian or English.

Parameter		Fluency	Lexical	Grammar	Pron	IELTS_band	Test
f3_LLR(SS)	r	-0.294	-0.355	-0.245		-0.298	
	p-value	0.038	0.011	0.87		0.036	
cov_LLR(DS)	r	0.328		0.258	0.276	0.299	
	p-value	0.02		0.071	0.052	0.035	
frontcor_ER(DS)	r	0.341					
	p-value	0.015					
Formants_LLR(SS)	r	-0.307	-0.315			-0.292	
	p-value	0.03	0.026			0.04	
Formants_cov_LLR(SS)	r	-0.274	-0.32			-0.275	
	p-value	0.054	0.023			0.053	
Formants_cov_ER(DS)	r	0.264					
	p-value	0.064					
H1*A1*_LLR(DS)	r		-0.259	-0.3			
	p-value		0.07	0.034			
H1*-A2*_ER(DS)	r	0.243	0.296	0.27			
	p-value	0.09	0.037	0.058			
H1*-A3*_ER(DS)	r	0.299	0.408			0.299	
	p-value	0.035	0.003			0.035	
HNR05_ER(DS)	r			-0.281			-0.322
	p-value			0.048			0.023
f3_H1*-A3*_HNR35_LLR(SS)	r	-0.264	-0.33		-0.282	-0.279	
	p-value	0.063	0.019		0.047	0.05	
f3_H1*-A3*_HNR35_LLR(DS)	r	0.375	0.345		0.253	0.309	
	p-value	0.007	0.014		0.076	0.03	
cov_H1*-H2*_HNR35_LLR(DS)	r	0.297					
	p-value	0.036					
cov_H1*-A3*_HNR35_LLR(SS)	r		-0.247				
	p-value		0.084				
cov_H1*-A3*_HNR35_LLR(DS)	r	0.383	0.359			0.286	
	p-value	0.006	0.011			0.044	
Significant for p < 0.01	Significant for p < 0.05		Significant for p < 0.1				

Figure 7-13
Pearson correlation between individual performance measures and foreign language proficiency (Serbian reference population)

Figure 7-13 depicts the relationships observed between various proficiency parameters and derived measures of individual performance significant for $p < 0.1$. Parameters that emerge to be related to individual speaker performance under the LR system are LTF3, Covariance (F2-F3), H1*-A3* and to a lower extent Frontness* and H1*-A2*. Consequently, any combinations of parameters that contain any of the above were also found to correlate with the proficiency scores. For these parameters it was found that higher same-speaker log-likelihood ratio and, conversely, lower different-speaker log-likelihood ratio are expected for speakers with lower foreign language proficiency as well as that speakers with higher proficiency exhibit higher error rates in different-speaker comparisons. Parameters for which higher proficiency correlates to better performance in cross-language speaker comparisons under the MVKD LR system are H1*-A1* and HNR_{05}.

Next, we performed group-wise comparison according to the three CEFR levels and two proficiency categories of the IELTS exam. Table 7-23 lists all of the individual performance parameters for which differences were detected between groups. Some of the results from the table are presented as boxplots further below.

Table 7-23
Relationship between individual performance measures and different proficiency groups (Serbian reference population)

Parameter	***B1 – B2 – C1***		***Independent - Proficient***	
	F-test	*p-value*	*t-test*	*p-value*
LTF1 LLR_{SS}	0.325	.068	-	-
LTF3 LLR_{SS}	1.527	.026	-	-
LTF3 LLR_{DS}	0.831	.023	-	-
LTF3 ER_{DS}	-	-	-1.726	.093
Covariance F2-F3 LLR_{DS}	-	-	-1.784	.085
H1*H2* ER_{DS}	-	-	-1.998	.052
H1*-A1* LLR_{DS}	0.373	.002	-	-
H1*-A3* LLR_{DS}	5.688	.089	-2.385	.021
H1*-A3* ER_{DS}	5.417	.016	-2.327	.025
H4*-2K* LLR_{DS}	-	-	-2.067	.044
HNR_{05} LLR_{SS}	-	-	-2.063	.045
HNR_{05} LLR_{DS}	-	-	1.836	.083
HNR_{05} ER_{DS}	-	-	2.001	.057
HNR_{25} LLR_{DS}	2.233	.064	-	-
HNR_{25} ER_{DS}	3.238	.001	1.799	.088
HNR_{35} LLR_{DS}	0.824	.065	-	-
HNR_{35} ER_{DS}	0.095	.024	-	-

Group-wise comparison has confirmed that LTF3, Covariance of F2-F3 and H1*-A3* are potentially more useful for speaker characterisation with lower-proficiency speakers, whereas H1*A1*, and HNR parameters perform better for speakers with higher proficiency.

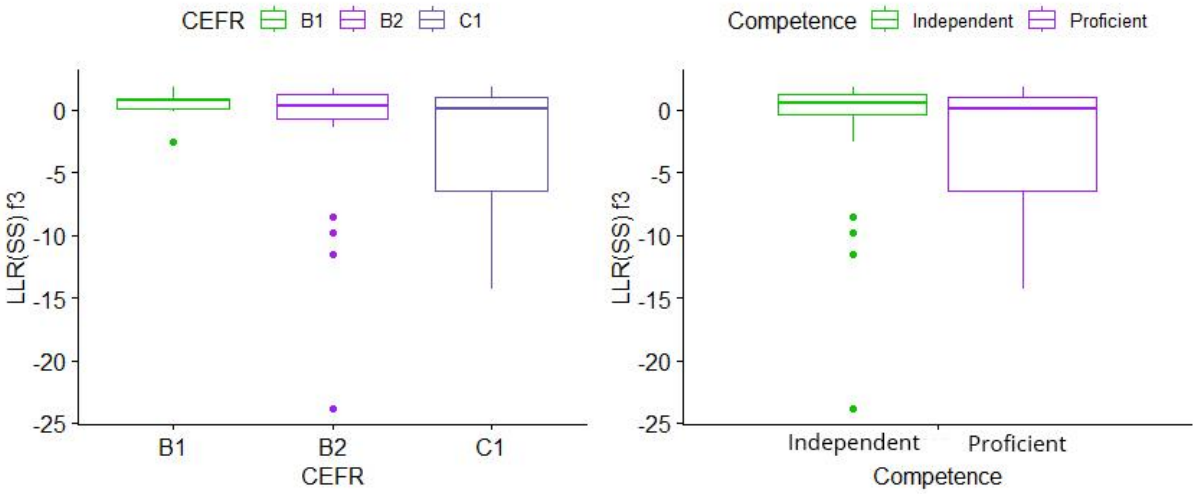


Figure 7-14
Dependence of LTF3 LLR in SS cross-language comparisons on language proficiency (Serbian reference population)

Looking at Figure 7-14 we can note that more proficient speakers exhibit a greater number of negative LLRs for same-speaker comparisons in relation to LTF3. Such a relationship can be observed for F2-F3 covariance as well, although to a lower extent (Figure 7-15). The most prominent phonation parameter that is in correlation with foreign language proficiency is the difference in amplitude between the first harmonic and the harmonic closest to the third formant (H1*-A3*). In Figure 7-16 we can observe that the increase in error rates is proportional to increase in proficiency.

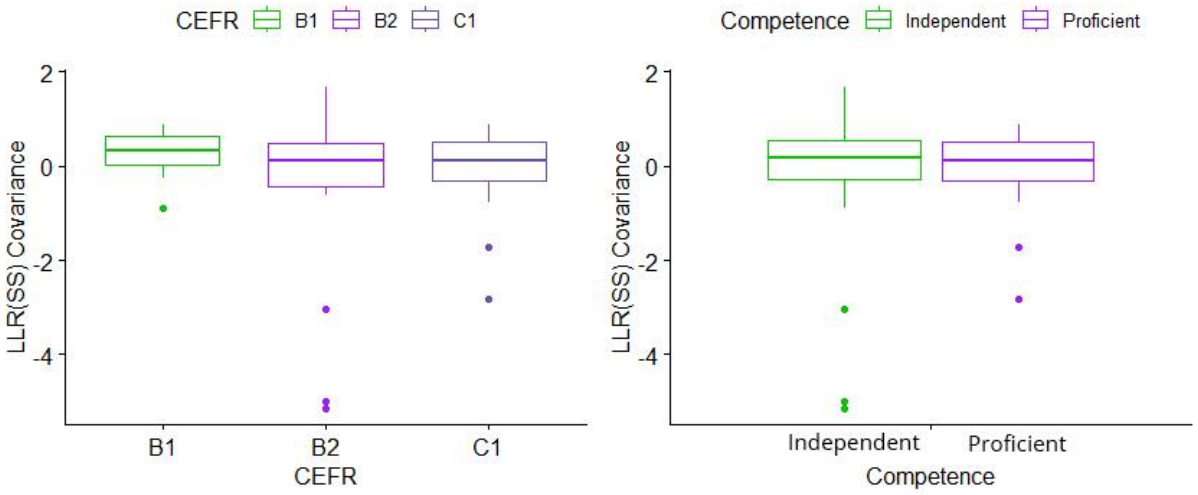


Figure 7-15
Dependence of F2-F3 covariance LLR in SS cross-language comparisons on language proficiency (Serbian reference population)

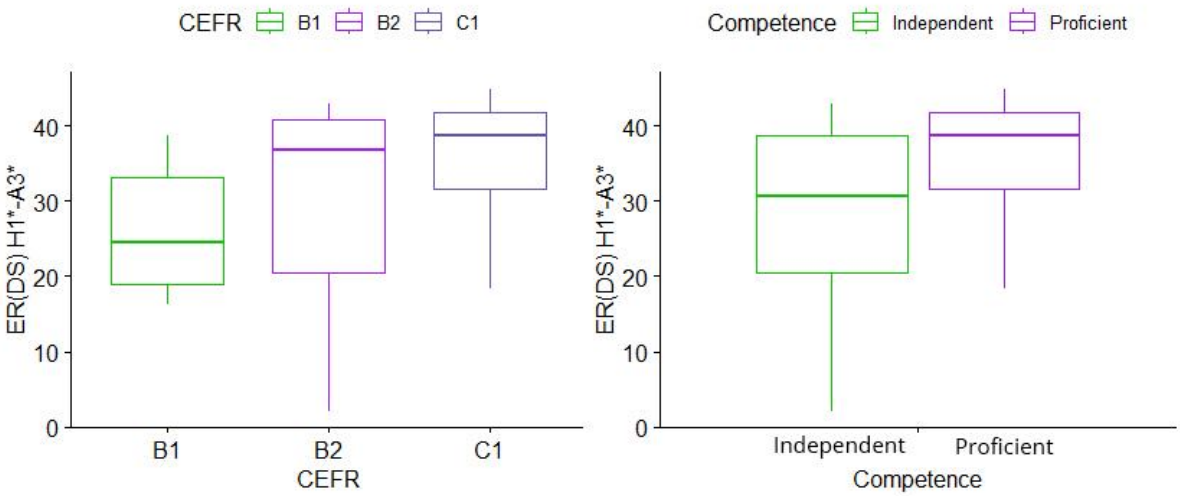


Figure 7-16
Dependence of H1-A3* ER in DS cross-language comparisons on language proficiency (Serbian reference population)*

Very similar relationships between individual speaker performance and proficiency can be observed for cross-language comparisons where background population is derived from data in English (Figure 7-18). In this condition, as well as in the previous one, parameters that perform better for speakers with lower foreign language proficiency are LTF3, F2-F3 covariance, Frontness*, H1*-A3*, H1*A2* and any fused combinations containing these parameters. The parameters that emerge to perform better for speakers with higher foreign language proficiency include CPP, H1*-A1* and HNR_{05}. Group-wise comparisons can be observed in Table 7-24, whereas Figure 7-17 depicts the relationship between HNR_{05} and proficiency.

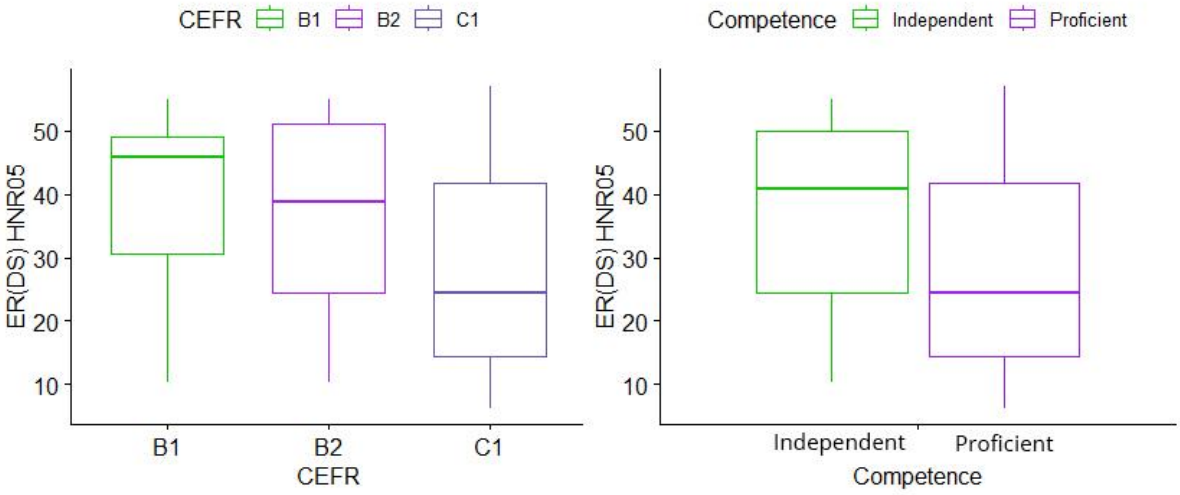


Figure 7-17
Dependence of HNR_{05} ER in DS cross-language comparisons on language proficiency (Serbian reference population)

As observed in Figure 7-17 above, proficient speakers exhibit lower error rates in different-speaker cross language comparisons than independent speakers for harmonic-to-noise ratio in the lower frequency range. A possible reason for this phenomenon could be the fact that less proficient participants tend to speak more quietly, with more hesitation and effort, which renders their speech hoarser than the speech in their mother tongue, which in turn results in the change of values of otherwise language-independent parameter.

Parameter		Fluency	Lexical	Grammar	Pron	IELTS_band	Test
f3_LLR(SS)	r	-0.292	-0.347	-0.246		-0.298	
	p-value	0.039	0.014	0.085		0.036	
cov_LLR(DS)	r	0.315		0.244	0.267	0.286	
	p-value	0.026		0.088	0.061	0.044	
frontcor_ER(DS)	r	0.353					
	p-value	0.012					
Formants_LLR(SS)	r	-0.272	-0.291			-0.256	
	p-value	0.056	0.04			0.073	
Formants_cov_LLR(SS)	r	-0.255	-0.308			-0.255	
	p-value	0.073	0.03			0.074	
Formants_cov_ER(DS)	r	0.278				0.238	
	p-value	0.051				0.097	
H1*A1*_LLR(DS)	r		-0.257	-0.296		-0.259	
	p-value		0.072	0.037		0.069	
H1*-A2*_ER(DS)	r		0.282	0.266		0.247	
	p-value		0.048	0.062		0.084	
H1*-A3*_ER(DS)	r	0.301	0.403			0.299	
	p-value	0.033	0.004			0.035	
H4*-2K*_LLR(DS)	r		0.281				
	p-value		0.048				
CPP_LLR(DS)	r				-0.26		
	p-value				0.069		
HNR05_ER(DS)	r			-0.264			-0.306
	p-value			0.064			0.031
f3_H1*-A3*_HNR35_LLR(SS)	r	-0.266	-0.329		-0.288	-0.28	
	p-value	0.062	0.02		0.043	0.048	
f3_H1*-A3*_HNR35_LLR(DS)	r	0.378	0.348		0.257	0.313	
	p-value	0.007	0.013		0.072	0.027	
cov_H1*-H2*_HNR35_LLR(DS)	r	0.299	0.269				
	p-value	0.035	0.059				
cov_H1*-A3*_HNR35_LLR(SS)	r		-0.258				
	p-value		0.07				
cov_H1*-A3*_HNR35_LLR(DS)	r	0.382	0.358			0.286	
	p-value	0.006	0.011			0.044	
Significant for p < 0.01	Significant for p < 0.05		Significant for p < 0.1				

Figure 7-18

Pearson correlation between individual performance measures and foreign language proficiency (English reference population)

Table 7-24
Relationship between individual performance measures and different proficiency groups (English reference population)

Parameter	***B1 – B2 – C1***		***Independent - Proficient***	
	F-test	*p-value*	*t-test*	*p-value*
LTF1 LLR_{SS}	0.292	.074	-	-
LTF3 LLR_{SS}	1.5	.034	-	-
LTF3 LLR_{DS}	0.786	.023	-	-
LTF3 ER_{DS}	-	-	-1.712	.095
Covariance F2-F3 LLR_{DS}	-	-	-1.734	.093
H1*H2* ER_{DS}	-	-	-2.005	.051
H1*-A1* LLR_{DS}	0.37	.002	-	-
H1*-A3* LLR_{DS}	5.676	.09	-2.382	.021
H1*-A3* ER_{DS}	5.239	.016	-2.289	.028
H4*-2K* LLR_{DS}	-	-	-2.092	.042
HNR_{05} LLR_{DS}	-	-	1.821	.085
HNR_{05} ER_{DS}	-	-	1.875	.073
HNR_{15} LLR_{SS}	-	-	-2.104	.041
HNR_{25} LLR_{DS}	2.188	.065	-	-
HNR_{25} ER_{DS}	2.904	.002	-	-
HNR_{35} LLR_{DS}	0.812	.065	-	-
HNR_{35} ER_{DS}	0.08	.018	-	-

7.2.6. Discussion

In the present research, we explored two methodologies of likelihood ratio calculation, multivariate kernel density formula with leave-one-out cross-validation technique (Aitken & Lucy, 2004) and GMM-UBM (Reynolds et al., 2000) likelihood ratio with three groups of speakers: test set, training set and background set. While GMM-UBM LR yields a narrower range of EER and C_{llr} scores, MVKD model was considered better for assessing our dataset as it allows for a larger number of background speakers to be employed for each comparison. Furthermore, the MVKD system can produce comparable C_{llr} scores to the ones obtained with GMM-UBM calculation provided that additional calibration of individual parameters is performed. However, since MVKD model yields poor cost log likelihood ratio if a large dataset with many points is used, it is first necessary to summarise the raw values across predetermined number of entries (cf. Gold, 2014; Tomić & French, 2019), which was tested by summarising the data over 1s and over 2s of extracted vowels. Slightly better cost log likelihood ratio was achieved when the values were averaged across 2 seconds of vowels, which is in accordance with Gold's (2014) findings, where she obtained the best C_{llr} score for "package" length between 2.5s and 5s (p. 171). It is important to point out that due to different statistical models, error rates obtained through GMM-UBM and MVKD calculations may differ for

individual parameters. Namely, parameters that are good discriminants under one model do not necessarily perform so well under the other.

In single-language comparison, it was found that, in terms of formant values, LTF3 performs best with an EER of 20%, whereas the second and the third formant exhibit an EER of around 30%. Comparable results were reported by Gold et al., (2013) and Gold (2014), who employed a similar methodology to estimate the performance of long-term formant frequencies, obtaining an EER of 17% for LTF3, 28% for LTF1 and 32% for LTF2, and Lo (2021), who utilised GMM-UBM likelihood ratio and reached EER between 18.8% and 27.2% for individual formants. However, as a result of utilisation of a different statistical model, Lo's (2021) findings suggest that LTF1 is the best discriminant among the long-term formants. Such findings were confirmed in the present study when GMM-UBM calculation was employed (see Appendix 12). With combination of the first three formants, Tomić and French (2019) reached an EER of around 8%, Lo (2021) obtained an EER of around 9%, and Gold et al. (2013) and Gold (2014) of 11.47%, matched by 12% in the present study. Considering that cited studies were performed on studio recordings and the present study on mobile phone recordings, we can conclude that the results are comparable. The C_{llr} scores of the present study indicate a very well calibrated system, which has proven to be even more reliable with the combination of parameters. The performance of the system, however, notably deteriorates when both the known and the questioned sample are in the foreign language. One of the possible reasons for that could be the unequal language competence of the participants, which introduces inconsistencies in pronunciation and delivery in the second language as will be discussed later. Whether the observation can be replicated for any pair of languages is difficult to determine since previous studies have not dealt with bilingual speaker comparison in the non-dominant language under the likelihood ratio framework.

In the present research, an even better performance of the LR system was detected for phonation parameters, with EER ranging between 18% and 34%. For the majority of examined parameters, including H1*-A1*, H1*-A2*, H1*-A3*, CPP, HNR_{05}, HNR_{25}, HNR_{35}, the EER in single-language comparisons does not exceed 22%, matching the performance of LTF3. Combining all of the examined phonatory features yielded an EER of below 2% and a C_{llr} of below 0.05 for comparisons in Serbian. The results confirm Vaňková and Skarnitzl's (2014) findings that H1*-H2*, H1*-A1* and H1*-A2* outperform formant values in forensic speaker comparison (p. 1081). For comparisons in the foreign language, most of the parameters exhibit slightly higher EER and C_{llr}, but the performance of the system can be described as comparable to its performance for the comparisons in the mother tongue. Combining all of the

phonation parameters in foreign-language comparisons resulted in an EER of 2.22% and a C_{llr} lower than 0.09, which can be described as an overall excellent performance. The results of the present system outperform even Cardoso et al.'s (2019) GMM-UBM system, where, across multiple replications, an average EER of 9.6% was reached for all phonation features in high-quality studio recordings and 13.4% in mobile phone recordings. The results are in contrast with Holmes (2023), who found that inclusion of HNR with formant values, f_0 and intensity results in deteriorated performance of the system. The reason for the discrepancy most likely stems from the difference in the applied LR methodologies.

Considering that extraction of such a large number of parameters as performed in the present study can be time-consuming, we examined which combination of articulatory and phonatory features would potentially characterise speakers well enough. An impressive EER of 4% and C_{llr} of 0.5 were achieved by combining only three of all of the examined articulatory and phonatory measures: LTF3, H1*-A3* and HNR_{35}. Cardoso et al.'s (2019) conclude that spectral tilt measures are almost unaffected by the transmission channel, which renders them reliable parameters for speaker characterisation in forensic casework. However, more research is needed to confirm these findings and place the acoustic analysis of phonation on the speaker comparison map.

With regard to cross-language speaker comparison under the likelihood ratio framework, the analysis was performed in three conditions: (1) reference (background) population derived from both the values in English and Serbian, (2) reference population derived from the values in Serbian and (3) reference population derived from the values in English. The results obtained for cross-language comparisons with background data in Serbian are very close to the results of single-language comparisons. However, calibration of log-likelihood ratio scores for individual parameters is essential to reduce C_{llr} below 1, what is more, it does not affect error rates or fusion. The EER and calibrated C_{llr} scores obtained in condition (2) and (3) are almost identical.

LTF3 was found to be the best performing formant with an EER of 22%, followed by LTF1 (31%) and LTF2 (38%). The combination of the three parameters through fusion yielded an EER of around 12.5%, which could further be improved by taking into account the F2-F3 covariance. The combination of the LTF1 and LTF3 yields an EER of almost 18%, indicating that, even though it does not perform so well on its own, LTF2 is crucial in speaker characterization across languages. In previous research, Tomić and French (2019) reported an EER of around 18% for cross-language comparison for the combination of the first three formants, an EER of 25% for LTF3, 34.5% for LTF1 and 40% for LTF2. Similarly, Lo (2021)

found LTF1 and LTF3 to yield an EER of around 25% and LTF2 around 30% in cross-language comparisons (p. 229). Likewise, in Persian-English comparisons, Asadi et al. (2022) note that F2 exhibits the poorest performance compared to other formants. The results of the present study performed on mobile phone recordings appear to be comparable to the previous research performed on studio quality recordings. The better performance of LTF1 in Lo's (2021) research is possibly related to the employment of GMM-UBM LR model.

As with the single-language comparisons, phonatory features were found to perform better within the MVKD LR system than the articulatory features, with CPP, HNR_{35}, HNR_{25}, and H1*-A3* exhibiting the EER between 20% and 23.5% and calibrated C_{llr} between 0.59 and 0.65. All phonation measures combined with fusion achieved an EER lower than 4% and C_{llr} of 0.14. The performance of the cross-language speaker comparison under the present system outperforms the systems employed in the previous studies for single language-comparisons on either studio or mobile phone recordings for the same parameters (cf. Cardoso et al., 2019). The results may be interpreted to indicate two things: (1) multi-variate kernel density formula might be more appropriate for speaker comparison based on acoustic phonatory features than the GMM-UBM-based one and (2) inclusion of a greater number of speakers in the background population yields more reliable speaker comparison results. The two hypotheses need to be researched further for the sake of obtaining a more definite explanation of the observations noted above.

Similar to the speaker comparison without language mismatch, in cross-language speaker comparison, it was found that the combination of articulatory and phonatory features results in a better speaker characterisation. With the combination of only three parameters (LTF3, H1*-A3*, HNR_{35}), the system reached an EER of around 6% and C_{llr} of 0.24. An EER of 2.5% and C_{llr} of 0.0999 were achieved with the fusion of all three long-term formants, cepstral peak prominence, H1*H2*, H1*-A3* and harmonic-to-noise ratio in the domain of up to 3.5 kHz.

Finally, we observed the relationships between individual speaker performance within the LR system and their language proficiency. The scores reached through an IELTS-based fluency rating appear to be the most relevant to individual speaker performance. Perhaps pronunciation was expected to correlate with LR scores more, however, our results have not detected a strong relationship for this parameter. Certain correlations are noted for the ratings of lexical and grammatical competence, but these should not be observed as being in a direct causal relationship with LR scores and error rates. Namely, when a learner has a better fluency they are more likely to have a larger repository of vocabulary and more comprehensive

knowledge of grammar as well. According to the present study, voice quality parameters that are more useful for speaker characterisation when proficiency is lower include the third formant, F2-F3 covariance, H1*-A2*, H1*-A3* and their combinations, whereas those that seem to perform better when proficiency is higher are H1*-A1* and HNR_{05}. The relationships between the acoustic values of these parameters and language proficiency were also observed in section 7.1.3. and will be further discussed in the following chapter.

7.3. Acoustic Analysis and LR Calculation Discussion

In the present chapter, we explained the procedures employed in the acoustic analysis of voice quality correlates and displayed the results of likelihood ratio calculations for the purposes of forensic speaker comparison is language-matched and language-mismatched conditions. In this interim discussion, we will return to the research questions (6) - (11) raised in Chapter 5.1., and consider the findings obtained via the acoustic analysis.

The acoustic analysis of articulatory features in native Serbian and foreign English has revealed that LTF1, LTF3 and F2-F3 covariance do not differ significantly in Serbian and English for individual speakers, whereas LTF2 exhibits notable divergence across languages. The reason for such a distinction most likely lies in the difference between the phonemic systems of the two languages and degree of phonemic acquisition by individual speakers. From population studies (e.g. Bjelaković, 2018; Hillenbrand et al., 1995; Tomić & Milenković, 2019), we already know that English has a wider vowel space than Serbian and could be described as more fronted; however, studies exploring acquisition of English by the native speakers of Serbian reveal that most of the learners create some kind of "compromise" between the vowel categories existing in their mother tongue and target vowel categories in the foreign language and their formant values, more often than not, do not reach the target values (see Bjelaković, 2018; Marković, 2009a; Marković, 2009b; Marković & Jakovljević, 2016; Paunović, 2011). Therefore, both factors must interplay to influence the formant values in the target language.

On the other hand, most of the examined phonatory features were found to exhibit significant difference across languages. The only parameters that do not exhibit any difference across languages are the difference between the amplitude of the first and second harmonic (H1*-H2*), the difference between the amplitude of the first harmonic and the harmonic closest to the third formant (H1*-A3*) and harmonic-to-noise ratio when measured between 0 and 3,500 Hz (HNR_{35}). Considering that there are not any population studies that list the values of phonatory features in native Serbian and native English, it is difficult to determine whether the

observed distinction in the two languages is a consequence of degree of acquisition or a result of fluency in general.

The articulatory and phonatory parameters that were found to be steady across languages could be considered speaker-specific rather than language-specific and are thus seen as useable in cross-language forensic speaker comparison. However, to understand the performance of the parameters in the condition of language-mismatch, we first explored their contribution to FSC in the single-language context. For speaker comparison under the likelihood ratio in Serbian, the third formant emerged as a most reliable discriminant (EER = 20.18%, C_{llr} = 0.67), whereas the second formant exhibited highest EER (32.37%) and much higher C_{llr} (0.8). However, despite its poor performance on its own, when combined with other formants through fusion, LTF2 significantly contributes to speaker characterisation and improved EER and C_{llr} scores (EER = 12.02%, C_{llr} = 0.38). The same is true of phonatory features which, in most cases, perform equivalently to LTF3 but in combination yield significantly more accurate results. The phonatory parameters with lowest EER and C_{llr} in speaker comparison in Serbian are H1*-A2* (EER = 17.98%, C_{llr} = 0.51), H1*-A3 (EER = 18.08%, C_{llr} = 0.51), and CPP (EER = 18.06%, C_{llr} = 0.56).

The ERR and C_{llr} scores of cross-language comparisons reflect the scores obtained for speaker comparison without language mismatch. The best performing articulatory parameter is LTF3 (EER = 22%, C_{llr} = 0.66), whereas the worst is LTF2 (EER = 38.04%, C_{llr} = 0.87). The first three formants in combination, however, yield the results equivalent to speaker comparison with both samples in Serbian (EER = 12.51%, C_{llr} = 0.46). The phonatory parameters with lowest EER and C_{llr} in speaker comparison under language mismatch are CPP (EER = 20.08%, C_{llr} = 0.59), HNR_{25} (EER = 21.96%, C_{llr} = 0.65), and HNR_{35} (EER = 20.45%, C_{llr} = 0.58), whereas H1*-A2* (EER = 25.63%, C_{llr} = 0.74) and H1*-A3 (EER = 23.51%, C_{llr} = 0.61) perform slightly worse than in the previous context. The language effect previously noted by Tomić & French (2019) and Lo (2021) was detected in the present study as well, as the performance of most parameters slightly deteriorates in the language-mismatched condition. Notwithstanding, given the obtained EER and C_{llr} scores, the examined parameters could be described as favourable for cross-language speaker comparison, especially when considered in combination.

With regard to cross-language FSC, phonatory measures appear to be more robust than formant values, but the contribution of the articulatory features to the overall system performance cannot be disregarded. Therefore, according to the results of the present study, just

like with single-language comparisons, it is best that the phonatory and articulatory features be observed in combination.

In forensic speaker comparison with a mismatch in conditions there is always a question of which reference (background) population to use for feature assessment. For instance, if the offender (questioned) sample is in English and background (reference) population in Serbian, the offender sample will be less typical of the background population, and, therefore, any similarities with the known sample will be overstated. On the contrary, if the offender sample is matched in condition with the background population (both in English), the sample will be more typical of the population in question and any similarities to the known sample may be understated. In the judicial context, the latter bias appears to be less harmful than the former one. The best, non-biased, option appears to be to include the mismatched conditions in the background population as well, whereby the data of the background population would be comprised of both the language of the questioned (offender) and known sample. However, in the present research, the condition with the background population comprised of data in both languages resulted in higher EER and C_{llr} rates than either of the conditions where the background sample was comprised of measurements in a single language. One of the possible reasons for poorer performance of this condition could be methodological, considering that the two samples in the background population were treated as originating from different speakers. The results presented in this chapter and used for more in-depth analysis originate from the condition where the background data is comprised of values in Serbian. Such a decision was made on two bases: first, the EER and C_{llr} scores obtained in this condition indicate the best system performance and second, in forensic reality, the experts are far more likely to have access to the background population data for the speaker's native language than for the foreign language. The conclusion that we have reached in the present study are in accordance with recommendations by Alexander and Drygajlo (2004), González-Rodríguez et al. (2006), and Morrison et al. (2012), who suggest that the reference population should match the conditions of the known sample.

Our final question refers to the individual speaker performance within the LR system and its relationship to the foreign language proficiency. These relationships were examined for each parameter respectively and, as the results indicate, they are not quite straightforward. Namely, for particular parameters (e.g. LTF3, F2-F3 covariance, H1*-A2*, H1*-A3*) speakers tend to exhibit better performance within the LR system provided that their proficiency in the foreign language is lower. On the other hand, for parameters such as H1*-A1* and HNR_{05}, individual performance within the system seems to improve with higher

proficiency. If we remember that, in the present study, LTF3, H1*-A2*, and H1*-A3* were the best-performing parameters in single-language speaker comparisons, we can conclude that it is not a coincidence that the less proficient someone is in the foreign language, the more they will "sound like themselves" across languages. On the other hand, since the performance of H1*-A1* and HNR_{05} in the same-language comparisons is not lacking behind the former three by much, and the proficiency effect is quite opposite for them, we may conclude that these parameters are rather competence-dependant than language-dependant, and as such, their reliability deteriorates with the decrease in fluency. We assessed our hypothesis with a two-factor ANOVA performed on the summarised and non-summarised values of these two parameters, labelling all L1 values as proficient. A strong language competence effect was confirmed for HNR_{05} in both datasets but for H1*-A1* it was confirmed only when the non-summarised values were analysed (Table 7-25). Therefore, the noted relationship between DS LLR scores and language proficiency for H1*-A1* may also stem from other, less obvious factors.

Table 7-25
Two-factor ANOVA for Language and Proficiency effect

Parameter	***Summarised***	***F-score***	***p-value***	***All values***	***F-score***	***p-value***
HNR_{05}	Language	3.59	.0611	Language	24759	.000
	Competence	3.08	.0824	Competence	21236	.000
H1*-A1*	Language	3.485	.0649	Language	12629	.000
	Competence	0.197	.6585	Competence	712.2	.000

8. Final Remarks

The final chapter first provides a brief overview of the research goals and initial hypothesis of the thesis; then, it discusses how the research has answered the broad theoretical and practical questions raised in the very beginning (Section 8.1.). Next, we explore some of the limitations of the present study and suggest possible directions for future research (Section 8.2.). Finally, we discuss the significance of the present research and conclude the thesis with the outlook for the field in general (Section 8.3.).

8.1. Research goals revisited

The underlying hypothesis of the present research is that biological factors outweigh the sociolinguistic ones in characterisation of voice quality. Put differently, the present study examined the view that individual, speaker-specific features of the anatomy of the vocal tract are more responsible for voice quality than the language spoken. Voice quality is considered in the broad sense, encompassing both the laryngeal and supralaryngeal features of speech (Laver, 1983). The motivation behind the research is to improve the process of speaker comparison under language mismatch for forensic purposes. Therefore, the principles and theoretical framework under which the study was performed, including the likelihood ratio framework, are drawn from the domain of forensic sciences.

To understand how voice quality differs when we speak a foreign language, we set out to answer three general questions:

How similar are the voices of the same/different speakers when speaking Serbian (L1) and English (L2)?

The first and most comprehensive research question was explored through two perceptual experiments and the acoustic analysis of articulatory and phonatory voice quality parameters. The first listening experiment, presented in Chapter 6.1., involved expert listeners scoring the voice quality features on the Vocal Profile Analysis Protocol (Laver et al., 1981). In contrast, in the second experiment (Chapter 6.2.), naïve listeners were engaged to assess the similarity of presented voice pairs and perform speaker discrimination. The analysed acoustic parameters include articulatory measures such as long-term formant values (LTF1, LTF2, LTF3), F2-F3 covariance, frontness (F2-F1), and phonatory measures: H1*-H2* (difference between the amplitude of the first and the second harmonic), H2*-H4*, H1*-A1* (difference between the amplitude of the first harmonic and the harmonic nearest to F1), H1*-A2*, H1*-A3*, H4*-2K* (difference between the amplitude of the fourth harmonic the harmonic nearest

to 2,000 Hz), HNR_{05} (harmonic-to-noise ratio between 0-500 Hz), HNR_{15} (between 0-1,500 Hz), HNR_{25} (between 0-2,500 Hz), HNR_{35} (between 0-3,500 Hz), and CPP (cepstral peak prominence), a measure of voice perturbation (Chapter 7.1.).

Expert listeners' results have shown that the vocal profiles of individual speakers across languages are closer than the vocal profiles of different speakers in the same language, either Serbian or English. In addition, it was found that phonatory settings contribute to the within-speaker similarity across languages more than articulatory settings do. Such results are expected, considering that the articulatory base of the two compared languages is different and, irrespective of how successful they are in it, non-native speaker will be bound to modify their native articulatory base in order to communicate in the foreign language. Moreover, phonatory settings are more robust across experts. The results obtained through naïve listener assessment revealed that, while same-speaker voices were rated slightly more distinct in the cross-language context, different-speaker voices have a notably higher similarity score in the language-mismatching than in the language-matching condition. One of the possible interpretations of this observation is that naïve listeners ascribe the difference they hear to the language effect and thus try to compensate for it with a higher score.

With regard to the acoustic values of the selected parameters, it was confirmed that LTF2 is the only articulatory feature that exhibits notable divergence across native Serbian and foreign English. In addition, it is the only parameter for which the language effect is stronger than the speaker effect. Nonetheless, for LTF2, as with the rest of the articulatory parameters, between-speaker variability within the language is higher than within-speaker variability across languages. Such findings indicate that acoustic values of LTF2 for individual speakers most likely depend on both the level of acquisition and the difference between the phonemic systems of the native and target language. Conversely, the majority of the examined phonatory parameters exhibit significant differences across languages. The only parameters for which the speakers retain their values in Serbian and English are the difference between the amplitude of the first and second harmonic (H1*-H2*), the difference between the amplitude of the first harmonic and the harmonic closest to the third formant (H1*-A3*) and harmonic-to-noise ratio when measured between 0 and 3,500 Hz (HNR_{35}). These, along with HNR_{25}, are simultaneously the only phonatory parameters for which the language effect is lower than the speaker effect. Notwithstanding, all phonatory parameters have higher within-speaker variability across languages than between-speaker variability within a single language.

Based on the observations derived from perceptual and acoustic experiments, it can be concluded that while both articulatory and phonatory features of voice quality may be

affected by the language spoken, they remain characteristic of the individual, even across languages. The conclusion is corroborated by the results presented as the answer to the second research question below.

What is the effect of language mismatch on forensic speaker comparison using the acoustic correlates of voice quality with Serbian (L1) and English (L2) samples?

To explore the reliability of voice quality parameters in cross-language FSC, we assessed each same-speaker and different-speaker pair through the multi-variate kernel density likelihood ratio formula (Aitken & Lucy, 2004) and estimated overall system performance by calculating equal error rate (ERR) and cost-log likelihood ratio (C_{llr}) for each parameter as well as for their combinations (Chapter 7.2.). The optimal results were obtained when the language of the reference population matched the language of the known sample (Serbian); therefore, the presented results correspond to this condition.

The ERR and C_{llr} scores of cross-language comparisons reflect those obtained for speaker comparison without language mismatch. The best performing articulatory parameter is LTF3 (EER = 22%, C_{llr} = 0.66), followed by LTF1 (EER = 31.2%, C_{llr} = 0.78), and LTF2 (EER = 38.04%, C_{llr} = 0.87). In combination, the three formants perform almost equally well in the cross-language context as in the single-language context (EER = 12.51%, C_{llr} = 0.46, as opposed to EER = 12.02%, C_{llr} = 0.38). The EER for phonatory parameters in speaker comparison under language mismatch ranges between 20.08% and 34.06%, whereas C_{llr} is between 0.58 and 0.83 – the results rather similar to single-language comparisons (EER: 17.98%-33.98%; C_{llr}: 0.51-0.87). The performance of all phonatory parameters in combination yielded an EER of 3.88% and C_{llr} of 0.14, only slightly inferior to the result obtained for comparisons in Serbian (EER = 1.67%, C_{llr} = 0.05).

Concerning our question, the experiments have confirmed that the language effect is reflected in the deteriorated performance of the likelihood ratio formula under language mismatch; however, the system performance in this condition can still be described as quite satisfactory. The combination of parameters through calibration-fusion can only further improve the system's performance, and even the parameters that are not very good discriminants on their own (such as LTF2) can significantly contribute to speaker characterisation across languages in combination with other features. Therefore, if both the articulatory and phonatory aspects of voice quality are considered in combination, we can achieve a reliable speaker comparison system.

How does foreign language proficiency/fluency affect voice perception and cross-language forensic speaker comparison?

The instrument for assessing the recorded participants' foreign language proficiency was modelled after the IELTS speaking task (IELTS, 2023b). The speakers were rated according to four criteria: fluency and coherence, lexical resource, grammar and accuracy, and pronunciation, all of which were included in the derivation of the final band.

According to the results obtained through the expert listening experiment, there is some evidence that the vocal profiles of less proficient speakers exhibit lower similarity across languages than the vocal profiles of more proficient ones. In the experiment with naïve listeners, we did not detect any relationships between speakers' proficiency and perception results. In contrast, in FSC under the likelihood ratio framework, for particular parameters (e.g. LTF3, F2-F3 covariance, H1*-A2*, H1*-A3*), speakers tend to exhibit better performance within the LR system provided that their proficiency in the foreign language is not high. For other parameters (such as H1*-A1* and HNR_{05}), individual performance within the system seems to improve with higher proficiency.

Intuitively, it might be expected that the less proficient someone is, the more they will "sound like themselves", and this hypothesis may be true of the pronunciation of individual segments or rhythm. However, with lower proficiency, speakers are often less fluent and less confident, which results in speech with lower intensity and lower variability in pitch (cf. Čubrović, 2020; Kainada & Lengeris, 2015; Marković, 2011; Paunović, 2013; 2015; 2019), ultimately contributing to speakers "sounding different" from when they speak in their mother tongue. Considering that we have proven that phonatory features are more responsible for vocal profile similarity across languages than articulatory features, it becomes logical that the lower similarity of the vocal profiles may, among other factors, be a reflection of increased disfluency and insecurity, that is, the lack of proficiency. Similar reasoning may apply to the relationship between individual performance under the likelihood ratio system and proficiency. Concerning the naïve listeners, it has already been shown that they rely on a holistic rather than an analytical approach when assessing the similarity between speakers; therefore, their conclusions are most likely based on both the features that vary with high and the ones that vary with low fluency.

In conclusion, while the relationship between fluency/proficiency and voice quality across languages is undeniable, it is intricate and more complex than initially hypothesised. The degree to which a certain influence is caused by the difference in the linguistic structures in the two languages and to which it is caused by the degree of acquisition of these structures is quite

difficult to determine and requires more sophisticated instruments and analyses than employed in the present research.

8.2. Limitations and Future research

One of the main limitations of the present study is that it has set out to explore rather general questions on a limited dataset, that is, on the example of one native and one foreign language alone. Therefore, our conclusions should be regarded to reflect only the relationship between Serbian and English. The research would need to be replicated in other languages of different origins and structures in order for us to understand the universality of the findings reached here.

Namely, while they are not part of the same language group, both Serbian and English belong to the Indo-European language family and as such they share many features. Examining bilinguals across languages of different origin, such as Afro-Asiatic family (e.g. Arabic) or Sino-Tibetan family (e.g. Chinese) would be important for understanding the dependence of voice quality on the language spoken. Furthermore, it would be of benefit to compare language of different morpho-syntactic and phonological typology. For instance, Serbian is a highly inflective language and has a higher speech rate than English, yielding more phonemes per second for analysis. Exploring this contrast further can shed light on how the number of analysed phonemes reflects individual difference/similarities in voice quality. It would be of even greater value to observe the languages of different phonological typology, such as different syllable structure, word-prosodic systems, distinctive features, vowel harmony, presence or absence of nasalized vowels or glottalised consonants etc. (see Hyman & Planks, 2018).

With regard to the experiment on the VPA protocol, one of the main challenges we encountered was the homogeneity of the speaker corpus, which resulted in relatively low distance scores. This leads us to conclude that while the VPA protocol is helpful for speaker characterisation in general, the instrument falters with similar voices and requires additional methodologies to corroborate its results. In addition, applying the non-truncated protocol with more scalar degrees (see Laver et al., 1981) is likely to result in higher distance scores. Future research exploring vocal profiles might also benefit from selecting a more balanced corpus of speakers with distinct proficiency levels and employing a greater number of voice quality experts to obtain more reliable and less ambiguous results. A closer observation of articulatory and phonatory data in isolation may also provide insight into the dependency of cross-language voice quality on pronunciation.

Several limitations have been noted for the experiment involving naïve listeners as well. Namely, due to an already comprehensive task the listeners faced, we decided to exclude the condition with both samples in the foreign language. Including such a condition in the experiment would provide insight into how listeners process the identity of speakers in a foreign language. In addition, according to Orena et al. (2019), the language experience of the listeners is of great importance when identifying bilingual talkers; therefore, researchers may consider controlling for this parameter when performing experiments with naïve listeners in the future. Furthermore, while using the same pairs of voices across conditions (as in the present study) might be better for direct comparability of the results, it increases the chances of listeners memorising the voices they have already heard and approaching the next discrimination task as listening to the familiar rather than an unfamiliar voice. Bearing that familiar and unfamiliar voices are processed in different regions of our brain (Maguinness et al., 2018; Stevenage, 2018), repetition of the same voices across conditions might have affected the results. As an alternative to the described approach, employing pre-tests prior to the listening experiment might help detect particularly memorable voices and exclude them from the experiment (see Tomić, 2020). Finally, the results obtained in the present study have raised some questions concerning language and identity processing in the brain. Namely, it was found that in the different-speaker same-language context and the same-speaker different-language context, the percentage of correctly performed recognition is associated with the increased distances between samples derived from articulatory settings. On the other hand, in the different-speaker different-language context, successful rejection is associated with higher phonatory-based distances. Since neuro-linguistic research requires access to specialised equipment, studies exploring voice processing from a neurological perspective are scarce compared to psychoacoustic studies. Experiments employing EEG or fMRI could help us shed light on voice processing depending on the language spoken and provide insight into mechanisms employed in speaker discrimination in the cross-language context as well as into interdependence of voice quality and speaker recognition.

Let us now return to the main focus of the present research – the usability of voice quality parameters for cross-language forensic speaker comparison. Namely, when justifying the selection of long-term articulatory and phonatory features as acoustic parameters for speaker comparison under language mismatch, in the very beginning (Chapter 1.1.), we explained that voice quality is an extralinguistic feature in most languages; it is an index of someone's speaking habit and the nature of their vocal apparatus rather than a bearer of communicative information (Laver, 1994: p. 22-23). Nonetheless, in Chapter 3.3.1. we explored a variety of

languages in which phonatory voice quality is used to signal linguistic information. Studying some of the mentioned languages in the present context would be very informative both from the perspective of forensic speaker comparison and voice quality theory. Some of the research questions that arise from this premise are, for instance, whether the parameters explored here could still be employed for speaker characterisation in languages that employ phonatory voice quality for linguistic purposes and to what extent. In addition, from the perspective of cross-language FSC, it would be valuable to know how phonation features vary if a native speaker of, for instance, Burmese or Hindi switches to a foreign language that does not employ phonation as a distinctive feature and vice-versa.

Finally, one of the most important questions raised in the present study is the matter of the relationship between linguistic systems and degree of foreign language proficiency, that is, to what extent the language effect that has been detected is a matter of difference between the mother tongue and foreign language linguistic system and to what the degree of acquisition of that system. As demonstrated in the present research, the relationship between the realisation of different voice quality parameters in a foreign language and proficiency is complex and requires a much more sophisticated instrument to be understood precisely. For instance, instead of assessing foreign language proficiency (including grammar and lexical resources), it might be more informative to engage native listeners to perform pronunciation quality assessment. Similarly, while the IELTS speaking task provides guidelines for scoring pronunciation, the description of what needs to be achieved for a particular band is rather vague and leaves much room for subjective interpretation. Instead, constructing a custom, fine-grained pronunciation scale with a detailed explanation of which features need to be acquired for a particular grade/rank might contribute to better evaluating speakers' pronunciation.

Furthermore, the present study was performed on the corpus of foreign language learners, not simultaneous bilinguals. Replicating the research with second or third-generation immigrants in the area where the language of interest is spoken would help broaden the knowledge about the voice quality of bilingual speakers and, at the same time, be more forensically relevant. Similarly, in countries such as Serbia, where a dialect switch commonly occurs depending on the social context, voice quality across accents can be studied without catering for the proficiency effect in the target language.

Last but not least, there have already been some attempts to employ neural networks to automatically cluster voices based on phonation (e.g. Chanclu et al., 2021), or estimate the quality of articulation of individual phonemes in speech pathology (e.g. Bilibajkić et al., 2014;

Furundžić, 2018). With the recent developments in artificial intelligence and neural networks, forensic speaker comparison could benefit from future research that applies these technologies.

8.3. Conclusion

The present research, with its unique focus on the relationship of language fluency and voice quality, holds both theoretical and applied significance. Theoretically, the study aimed to delve into the impact of language on the voice quality of individual speakers. From an applied perspective, it sought to explore the potential of acoustical measures of voice quality in cross-language forensic speaker comparison.

The idea that people can be recognised based on their voice is not a novel notion. However, no finite set of features has been established that distinguish one voice from another. Moreover, not all voices show a discrepancy in the same regard, and between two very distinct voices, there are multiple ones that differ just enough - in the manner of a grayscale image with many shades between the two extremes (black and white). Not unlike language varieties that form a dialectological continuum with gradual changes across regions until they finally become different languages (which may still have many overlaps), human voices form a continuum in speaker space, each different from the other but with numerous overlapping characteristics so that even our delicate ears could often be deceived. That is why employing multiple features increases the probability of correct speaker characterisation, regardless of the strength of the discriminatory power of individual parameters. The above, of course, does not imply that speaker comparison in casework should be performed with parameters that have not been repeatedly tested and proven to perform above the chance level in isolation.

Regarding cross-language forensic speaker comparison, multiple studies have shown that the parameters that perform well in single-language comparisons have inferior performance in mismatched conditions due to the language effect. However, the phenomenon of language effect should not prevent the performance of cross-language speaker comparison in casework as long as it can be proven that, for a particular parameter, within-speaker variability across languages is lower than between-speaker variability within a single language. If we want to explore the language effect on a particular parameter, however, we need to conduct the analysis with the same participants on the same corpus in single and cross-language conditions and then estimate the language effect on the final scores. As noted before, every system performance depends "on the makeup of the development, test, and reference sets" (Cardoso et al., 2019). Therefore, using different conditions will likely result in different

performances within the system, which would have little to do with the language effect we are exploring.

So far, we have demonstrated that, regarding voice quality, the language effect, at least partly, depends on the speaker's fluency. Nonetheless, neither the fluency nor the difference in the linguistic systems outweigh the speaker-specific nature of the assessed parameters. The biological makeup of the shape and size of the vocal tract of an individual and idiosyncratic, habitual adjustments in the vocal apparatus persist across languages, at least as proven for native Serbian and foreign English. With more research and testing, we can replicate the results in enough pairs of languages to understand the universality of speaker-specificity of voice quality. Until then, as advised in the best practice guidelines of the International Association of Forensic Phonetics and Acoustics, we still need to approach cross-language forensic speaker comparison with caution.

References

Abercrombie, D. (1967). *Elements of General Phonetics.* Edinburgh: Edinburgh University Press.

Ackermann, H., & Ziegler, W. (2010). Brain Mechanisms Underlying Speech Motor Control. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), *The Handbook of Phonetic Sciences* (Second ed., pp. 202-250). Chichester: Wiley-Blackwell.

Aglieri, V., Watson, R., Pernet, C., Latinus, M., Garrido, L., & Belin , P. (2017). The Glasgow Voice Memory Test: Assessing the ability to memorize and recognize unfamiliar voices. *Behavior Research Methods, 49,* 97–110. https://doi.org/10.3758/s13428-015-0689-6

Agnitio. (2009). Batvox 3.0 Basic User Manual. Madrid.

Aitken, C. G., & Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *Applied Statistics, 53*(4), 109-122. https://doi.org/10.1046/j.0035-9254.2003.05271.x

Aitken, C. G., & Taroni, F. (2004). *Statistics and the Evaluation of Evidence for Forensic Scientists* (Second ed.). Chichester: John Wiley & Sons Ltd.

Aitken, C., Taroni, F., & Bozza, S. (2021). *Statistics and the Evaluation of Evidence for Forensic Scientists.* (Third, Ed.) Chichester: John Wiley & Sons Ltd.

Alamri, S. S. (2015). Text-independent, Automatic Speaker Recognition System Evaluation With Males Speaking Both Arabic and English. *Master thesis, Faculty of the Graduate School of the University of Colorado*.

Alexander, A., & Drygajlo, A. (2004). Scoring and direct methods for the interpretation of evidence in forensic speaker recognition. *Proceedings of Interspeech 2004 (ICSLP),* (pp. 2397-2400).

Alexander, A., Dessimoz, D., Botti, F., & Drygajlo, A. (2005). Aural and automatic forensic speaker recognition in mismatched conditions. *The International Journal of Speech, Language and the Law, 12*(2), 214-234. https://doi.org/10.1558/sll.2005.12.2.214

Amino, K., & Osanai, T. (2015). Cross-language differences of articulation rate and its transfer into Japanese as a second language. *Forensic Science International, 249,* 116-122. http://dx.doi.org/10.1016/j.forsciint.2015.01.029

Andruski, J. E., & Ratliff, M. (2000). Phonation types in production of phonological tone: the case of Green Mong. *Journal of the International Phonetic Association, 30*(1-2), 37-61. https://doi.org/10.1017/S0025100300006654

Armbrecht, J. (2015). Hesitation Rate as a Speaker-Specific Cue in Bilingual Individuals. *Master thesis, Department of Communication Sciences and Disorders College of Behavioral and Community Sciences University of South Florida*.

Asadi, H., & Asiaee, M. (2022). Acoustic variation within and between bilingual speakers. *A paper presented at the 1st Interdisciplinary Conference on Voice Identity (VoiceID): Perception, Production, and Computational Approaches, Zurich, Switzerland.*

Asadi, H., & Dellwo, V. (2019). Analyzing F0 and vowel formants of Persian based on long-term features. *Proceedings of the 28th Annual Conference of the Internatonal Assocaton for Forensc Phonetcs and Acoustcs.* Istanbul, July 14th-17th, 2019.

Asadi, H., Asiaee, M., & Alinezhad, B. (2022). Acoustic analysis of parameters affecting the between-speaker variability in Persian-English bilinguals. *ZABANPAZHUHI (Journal of Language Research)*. https://doi.org/10.22051/jlr.2022.40224.2174

Asiaee, M., & Asadi, H. (2022). Bilingual acoustic variation: The case of Sorani Kurdish-Persian speakers. *Acta Universitatis Carolinae: Philologica, 1*, 23-33. https://doi.org/10.14712/24646830.2022.26

Asiaee, M., Nourbakhsh, M., & Skarnitzl, R. (2019). Can LTF discriminate bilingual speakers? *Proceedings of the 28th Annual Conference of the Internatonal Assocaton for Forensc Phonetcs and Acoustcs.* Istanbul, July 14th-17th, 2019.

Askar, R., Wang, D., Bie, F., Wang, J., & Zheng, F. (2015). Cross-lingual speaker verification based on linear transform. *Proceedings of the Conference: 2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP).* https://doi.org/10.1109/ChinaSIP.2015.7230457

Aston University. (2022). *Forensic Linguistics MA.* Retrieved October 28, 2021, from Aston University, Birmingham UK: https://www.aston.ac.uk/study/courses/forensic-linguistics-ma/september-2022

Atkinson, M., & McHanwell, S. (2018). *Basic Medical Science for Speech and Language Therapy Students* (Second ed.). Emsworth: J&R Press Ltd.

Awan, S. N. (2011). The effect of smoking on the dysphonia severity index in females. *Folia phoniatrica et Logopaedica, 63*(2), 65-71. https://doi.org/10.1159/000316142

Bahmanbiglu, S. A., Mojiri, F., & Abnavi, F. (P249.E9-249.E12). The Impact of Language on Voice: An LTAS Study. *Journal of Voice, 31*(2), 2017.

Baken, R. J., & Orlikoff, R. F. (2000). *Clinical measurement of speech and voice* (Second ed.). San Diego: Singular Thomson Learning.

Baldwin, J. (1979). Phonetics and Speaker Identification. *Medicine, Science and the Law, 19*(4), 231-232. https://www.doi.org/10.1177/002580247901900405

Ball, M. J. (1996). Describing voice quality: Transcription and instrumentation. *Logopedics Phoniatrics Vocology, 21*(1), 59-63. https://doi.org/10.3109/14015439609099204

Ball, M. J., Esling, J. H., & Dickson, B. C. (1995). The VoQS system for the transcription of voice quality. *Journal of the International Phonetic Association, 25*(2), 71-80. https://doi.org/10.1017/S0025100300005181

Ball, M. J., Esling, J. H., & Dickson, B. C. (2018). Revisions to the VoQS system for the transcription of voice quality. *Journal of the International Phonetic Association, 48*(2), 165-171. https://doi.org/10.1017/S0025100317000159

Ball, M. J., Howard, S. J., H, E. J., & Dickson, B. C. (2016). Revisions to the extIPA and VoQS symbol sets. *Poster presented at the 16th ICPLA conference, Halifax, Nova Scotia.*

Barreda, S. (2021). Fast Track: fast (nearly) automatic formant-tracking using Praat. *Linguistics Vanguard, 7*(1), 1-12. https://doi.org/10.1515/lingvan-2020-0051

Bartle, A., & Dellwo, V. (2015). Auditory speaker discrimination by forensic phoneticians and naive listeners in voiced and whispered speech. *The International Journal of Speech, Language and the Law, 22*(2), 229–248. https://doi.org/10.1558/ijsll.v22i2.23101

Belotel-Grenié, A., & Grenié, M. (2004). The creaky voice phonation and the organisation of Chinese discourse. *Proceedings of the International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages. Beijing, China.*

Ben-Shachar, M. S., Lüdecke, D., & Makowski, D. (2020). effectsize: Estimation of Effect Size Indices and Standardized Parameters. *Journal of Open Source Software, 5*(56), 2815. https://doi.org/10.21105/joss.02815

Bhattacharjee, U., & Sarmah, K. (2012). GMM-UBM Based Speaker Verification in Multilingual Environment. *International Journal of Computer Science Issues, 9*(6), 373-380.

Bilibajkić, R., Furundžić, D., & Subotić, M. (2014). Application of neural networks for the detection of pathological articulation for Serbian phonemes. *Proceedings of the 22nd Telecommunications Forum (TELFOR), Belgrade, Serbia, 2014*, (pp. 873-876). https://doi.org/10.1109/TELFOR.2014.7034544

Bjelaković, A. (2018). Vokali savremenog standardnog britanskog izgovora i njihovo usvajanje kod izvornih govornika srpskog jezika. Beograd: Doktorska disertacija.

Blake, T. (2018). Reframing Foreign Language Learning as Bilingual Education: Epistemological Changes towards the Emergent Bilingual. *International Journal of Bilingual Education and Bilingualism, 21*(8), 1041-1048.

Bóna, J. (2014). Temporal characteristics of speech: The effect of age and speech style. *The Journal of the Acoustical Society of America, 136*(2), EL116. https://doi.org/10.1121/1.4885482

Bone, D., Kim, S., Lee, S., & Narayanan, S. S. (2010). A Study of Intra-Speaker and Inter-Speaker Affective Variability using Electroglottograph and Inverse Filtered Glottal Waveforms . *Proceedings of InterSpeech 2010, Makuhari, Chiba, Japan*, (pp. 913-916).

Borsky, M., Mehta, D. D., Stan, V., H, J., & Gudnason, J. (2017). Modal and Nonmodal Voice Quality Classification Using Acoustic and Electroglottographic Features. *IEEE/ACM Transactions on Audio, Speech and Language Processing, 25*(12), 2281-2291. https://doi.org/10.1109/taslp.2017.2759002

Braun, A., & Künzel, H. J. (1998). Is forensic speaker identification unethical - or can it be unethical not to do it? *Forensic Linguistics, 5*(1), 10-21. https://doi.org/10.1558/ijsll.v5i1.10

British Council. (2023). *Online English level test*. Retrieved June 2023, from Learn English by BritishCouncil: https://learnenglish.britishcouncil.org/english-levels/online-english-level-test

Broeders, A. (1996). Earwitness Identification: Common Ground, Disputed Territory and Uncharted Areas. *Forensic Linguistics, 3*, 1-13.

Broeders, A. P. (1999). Some observations on the use of probabilityscales in forensic identification. *Forensic Linguistics, 6*(2), 228-241. https://doi.org/10.1558/sll.1999.6.2.228

Broeders, A. P. (2001). Forensic Speech and Audio Analysis Forensic Linguistics 1998 to 2001: A Review. *Proceedings of the '13th INTERPOL Forensic Science Symposium'*, (pp. 1-23). Lyon, France.

Broeders, A., & van Amelsvoort, A. (1999). Lineup Construction for Forensic Earwitness Identification: A Practial Approach. *14th International Congress of Phonetic Sciences, San Francisco, CA*, (pp. 1373-1376).

Browman, C. P., & Goldstein, L. (1992). Articulatory Phonology: An Overview. *Phonetica, 49*(3-4), 23-42. https://doi.org/10.1159/000261913

Brown, P., & Levinson, S. C. (1987). *Politeness - Some Universals in Language Usage.* Cambridge: Cambridge University Press.

Brümmer, N., & du Preez, J. (2006). Application-Independent Evaluation of Speaker Detection . *Computer Speech and Language*.

Brunelle, M., Nguyên, D. D., & Nguyên, K. H. (2010). A Laryngographic and Laryngoscopic Study of Northern Vietnamese Tones. *Phonetica, 67*(3), 147-169. https://doi.org/10.1159/000321053

Bruyninckx, M., Harmegnies, B., Llisterri, J., & Poch-Oiivé, D. (1994). Language-induced voice quality variability in bilinguals. *Journal of Phonetics, 22*(1), 19-31. https://doi.org/10.1016/S0095-4470(19)30265-7

Burin, L. (2018a). Investigating voice quality with an electroglottograph (EGG). *A paper presented at the ALOES conference April, 2018, Paris, France.*

Burin, L. (2018b). Electroglottography as a method to analyse paralinguistic adaptation. *A paper presented at the Workshop on Accommodation in Speech Communication, 13 December 2018, University of Zurich, Switzerland.*

Byrne, C., & Foulkes, P. (2004). The 'Mobile Phone Effect' on vowel formants. *The International Journal of Speech Language and the Law, 11*(1), 83-102. https://doi.org/10.1558/ijsll.v11i1.83

California University. (2022). *FORENSIC LINGUISTICS DEGREE*. Retrieved October 28, 2021, from California University of Pennsylvania: https://www.calu.edu/academics/graduate/masters/forensic-linguistics/index.aspx

Calvache, C., Solaque, L., Velasco, A., & Peñuela, L. (in press). Biomechanical Models to Represent Vocal Physiology: A Systematic Review. *Journal of Voice, Available online 5 March 2021*. https://doi.org/10.1016/j.jvoice.2021.02.014

Camargo, Z. A., Madureira, S., Pessoa, A. N., & Rusilo, L. C. (2012). Voice Quality and Gender: Some Insights on Correlations between Perceptual and Acoustic Dimensions. *Proceedings of the Speech Prosody 2012, May 22-25, Shanghai, China*, (pp. 115-118).

Camargo, Z., & Canton, P. d. (2019). Qualidade vocal de crianças com alteração de frênulo na língua. *Journal of Speech Sciences, 8*(1), 15-26. https://doi.org/10.20396/joss.v8i1.14990

Cambier-Langeveld, T. (2016). Language analysis in the asylum procedure: a specification of the task in practice. *The International Journal of Speech, Language and the Law, 23*(1), 25-41. https://doi.org/10.1558/ijsll.v23i1.17539

Canonge, E. D. (1957). Voiceless vowels in Comanche. *International Journal of American Linguistics, 23*(2), 63-67. https://doi.org/10.1086/464394

Cardiff University. (2022). *Forensic Linguistics (MA)*. Retrieved October 28, 2021, from Cardiff University: https://www.cardiff.ac.uk/study/postgraduate/taught/courses/course/forensic-linguistics-ma

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

Carne, M. J. (2015). A Likelihood ratio-based forensic voice comparison in microphone vs. mobile mismatched conditions using Japanese /ai/. *Proceedings of Interspeech 2015, Dresden, Germany*, (pp. 3471-3475).

Catford, J. C. (1964). Phonation Types: The Classification of Some Laryngeal Components of Speech Production. In *In honour of Daniel Jones: Papers contributed on the occasion of his eightieth birthday, 12 September 1961* (pp. 26-37). London: Longman.

Cháfer, J. V. (2019). Voice line-ups: Testing aural-perceptual recognition on native speakers of a foreign language. *PhD Thesis. University of Barcelona*.

Champely, S. (2020). pwr: Basic Functions for Power Analysis. R package version 1.3-0. Retrieved May 2023, from https://CRAN.R-project.org/package=pwr

Champod, C., & Evett, I. W. (2000). Commentary on A. P. A. Broeders (1999) 'Some observations on the use of probability scales in forensic identification', Forensic Linguistics 6(2): 228–41. *International Journal of Speech Language and the Law, 7*(2), 238-243. 10.1558/ijsll.v7i2.239

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

Chanclu, A., Ben, A. I., Gendrot, C., Ferragne, E., & Bonastre, J.-F. (2021). Automatic classification of phonation types in spontaneous speech: towards a new workflow for the characterization of speakers' voice quality. *Proceedings of Interspeech 2021*, (pp. 1015-1018). https://doi.org/10.21437/Interspeech.2021-1765

Cheung, W. H., & Wee, L.-H. (2008). Viability of VOT as a Parameter for Speaker Identification: Evidence from Hong Kong. *Proceedings of the Conference: International Congress of Linguists (CIL), Seoul, Korea*, (pp. 1-15).

Chin, S. B., & Pisoni, D. B. (1997). *Alcohol and Speech.* San Diego: Academic Press.

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, Tokyo, Japan*, (pp. 1-6).

Chomsky, N. (1968). *Language and Mind.* New York: Harper and Row.

Clary, R. A., Pengilly, A., Bailey, M., Jones, N., Albert, D., Comins, J., & Appleton, J. (1996). Analysis of Voice Outcomes in Pediatric Patients Following Surgical Procedures for Laryngotracheal Stenosis. *Archives of OtorhinoLaryngology- Head & Neck Surgery, 122*(11), 1189-1194. https://doi.org/10.1001/archotol.1996.01890230035008

Cleveland, R. A. (1991). Acoustic and Perceptual Correlates of Breathy Vocal Quality. *Master thesis, Department of Speech Pathology and Audiology, Western Michigan University*.

Clifford, B. R., Rathborn, H., & Bull, R. (1981). The effects of delay on voice recognition accuracy. *Law and Human Behavior, 5*(2-3), 201–208. https://doi.org/10.1007/BF01044763

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (Second ed.). LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS.

Coulthard, M., & Johnson, A. (2007). *An Introduction to Forensic Linguistics: Language in Evidence.* Abingdon, Oxon: Routledge.

Cruttenden, A. (2014). *Gimson's Pronunciation of English* (Eigth ed.). London & New York: Routledge.

Crystal, D. (1975). *The English Tone of Voice. Essays in intonation, prosody and paralanguage.* London: Edward Arnold.

Čubrović, B. (2020). Acquisition of English Pitch Contours in Serbian Speakers of English. *Belgrade English Language and Literature Studies, 12*, 77-94. https://doi.org/10.18485/bells.2020.12.4

Cullen, P., French, A., & Jakeman, V. (2014). *The Official Cambridge Guide to IELTS for Academic and General Training.* Cambridge : Cambridge University Press.

Das, A., Zhao, G., Levis, J., Chukharev-Hudilainen, E., & Gutierrez-Osuna, R. (2020). Understanding the Effect of Voice Quality and Accent on Talker Similarity. *INTERSPEECH 2020*, (pp. 1763-7). 10.21437/Interspeech.2020-2910

de Boer, M. M., & Heeren, W. F. (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. https://doi.org/10.1121/10.0002871

de Boer, M., & Heeren, W. F. (2020). Binnensprekervariatie in de uitspraak van /m/ in verschillende talen: abstract. *Amsterdam: Nederlandse Vereniging voor Fonetische Wetenschappen.*

de Jong, G. (1998). Earwitness Characteristics and Speaker Identification Accuracy. *PhD Thesis*. University of Florida, Gainesville, Fl.

de Jong-Lendle, G., Nolan, F., McDougall, K., & Hudson, T. (2015). Voice Lineups: A Practical Guide. *18th International Congress of Phonetic Sciences. Glasgow, Scotland,* (pp. 10-14).

de Krom, G. (1993). A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals. *Journal of speech and hearing research, 36*(2), 254–266. https://doi.org/10.1044/jshr.3602.254

de Lima Silva, M. F., Madureira, S., Rusilo, L. C., & Camargo, Z. (2017). Vocal quality assessment: methodological approach for a perceptive data analysis. *Revista CEFAC, 19*(6). https://doi.org/10.1590/1982-021620171961417

Dejonckere, P. H., Remacle, M., Fresnel-Elbaz, E., Woisard, V., Crevier-Buchman, L., & Millet, B. (1996). Differentiated perceptual evaluation of pathological voice quality: reliability and correlations with acoustic measurements. *Revue de laryngologie - otologie - rhinologie, 117*(3), 219–224.

Derlatka-Kochel, M., Majos, M., Ludwisiak, K., & Majos, A. (2021). The value of 1.5T MRI in the evaluation of vocal fold mobility in patients with goiter. *European Journal of Radiology Open, 8,* 100368. https://doi.org/10.1016/j.ejro.2021.100368

Diesner, C., & Ishihara, S. (2016). Adapted Gaussian Mixture Model in Likelihood Ratio based Forensic Voice Comparison using Long Term Fundamental Frequency. *Proceedings of the Sixteenth Australasian International Conference on Speech Science and Technology*, (pp. 141-144).

Dong, L., & Kong, J. (2021). Electroglottographic Analysis of the Voice in Young Male Role of Kunqu Opera. *Applied Sciences, 11*(9), 3930. https://doi.org/10.3390/app11093930

Đorđević, J. P., Kašić, Z., & Jovičić T, S. (2011). Forensic - Phonetic Speech Analysis. (S. T. Jovičić, & M. Subotić, Eds.) *Verbal Communication Quality Interdisciplinary Research I*, pp. 79-94.

Đorđević, M., & Rajković, M. (2004). Kvantifikovanje značajnosti akustičkih obeležja u inter i intra spikerskoj distinkciji emotivnog govora. *Kvantifikovanje značajnosti akustičkih obeležja u inter i intra spikerskoj distinkciji emotivnog govora*, *2*, pp. 351-354.

Đorđević, M., Jovičić, S., Rajković, M., & Kašić, Z. (2004). Analiza značajnosti akustičkih obeležja u distinkciji primarnih emocija u emotivnom govoru. *Zbornik radova DOGS 2004*, pp. 53-56.

Dorreen, K. (2017). Fundamental frequency distributions of bilingual speakers in forensic speaker comparison. *Master thesis, The University of Canterbury*.

Dowle, M., & Srinivasan, A. (2020). data.table: Extension of `data.frame`. R package version 1.13.2. https://CRAN.R-project.org/package=data.table

Drygajlo, A., Jessen, M., Gfroerer, S., Wagner, I., Vermeulen, J., & Niemi, T. (2015). *Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition including Guidance on the Conduct of Proficiency Testing and Collaborative Exercises.* ENFSI. Retrieved January 2023, from https://enfsi.eu/wp-content/uploads/2016/09/guidelines_fasr_and_fsasr_0.pdf

Drygajlo, A., Meuwly, D., & Alexander, A. (2003). Statistical methods and Bayesian interpretation of evidence in forensic automatic speaker recognition. *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech 2003*, (pp. 689-692). https://doi.org/10.21437/Eurospeech.2003-297

Durou, G. (1999). Multilingual Text-independent Speaker Identification. *Proceedings of Multi-lingual Interoperability in Speech Technology (MIST)* (pp. 115-118). Leusden, The Netherlands: PN.

Earnshaw, K. (2016). Assessing the Discriminatory Power of /t/ and /k/ for Forensic Speaker Comparison using a Likelihood Ratio Approach. *A paper presented atThe 25th annual conference of the International Association for Forensic Phonetics and Acoustics* , (pp. 92-95).

Edmond, G., Martire, K., & San Roque, M. (2011). 'Mere guesswork': Cross-Lingual Voice Comparisons and the Jury. *Sidney Law Review, 33*(3), 395-425.

Edmondson, J. A., & Esling, J. H. (2006). The valves of the throat and their functioning in tone, vocal register, and stress: Laryngoscopic case studies. *Phonology, 23*(2), 157–191. https://doi.org/10.1017/S095267570600087X

Edmondson, J. A., Ziwo, L., Esling, J. H., Harris, J. G., & Shaoni, L. (2001). The aryepiglottic folds and voice quality in the Yi and Bai languages: Laryngoscopic case studies. *Mon-Khmer Studies, 31,* 83-100.

ENFSI. (2009). Best practice guidelines for ENF analysis in forensic authentication of digital evidence FSAAWG-BPM-ENF-001 (1.0).

ENFSI. (2021). Best Practice Manual for the Methodology of Forensic Speaker Comparison. *ENFSI-BPM-FSC-01 Version 01*. Retrieved January 2023, from https://enfsi.eu/wp-content/uploads/2021/07/2021-07-07-final-draft-BPM-SPEAKER-COMPARISON.pdf

ENFSI. (2022). Best Practice Manual for Digital Audio Authenticity Analysis ENFSI-FSA-BPM-002 (1.0).

Enzinger, E. (2014). A first attempt at compensating for effects due to recording-condition mismatch in formant-trajectory-based forensic voice comparison. *Proceedings of the 15th Australasian International Conference on Speech Science and Technology*, (pp. 133-136).

Enzinger, E. (2016). Likelihood ratio calculation in acoustic-phonetic forensic voice comparison: Comparison of three statistical modelling approaches. *Proceedings of Interspeech, 2016*, (pp. 535-539). http://dx.doi.org/10.21437/Interspeech.2016-1611

Enzinger, E., Zhang, C., & Morrison, G. S. (2012). Voice source features for forensic voice comparison-An evaluation of the GLOTTEX software package. *Proceedings of Odyssey 2012-The Speaker and Language Recognition Workshop.*

Eriksson, A. (2010). The disguised voice: Imitating accents or speech styles and impersonating individuals. In C. Llamas, & D. Watt (Eds.), *Language and Identities* (pp. 86-98). Edinburgh: Edinburgh University Press.

Esling, J. (2005). There are no back vowels: The laryngeal articulator model. *Canadian Journal of Linguistics, 50*, 13-44. https://doi.org/10.1353/cjl.2007.0007

Esling, J. (2013). Voice and Phonation. In M. J. Jones, & R.-A. Knight (Eds.), *The Bloomsbury Companion to Phonetics* (pp. 110-125). London, New Delhi, New York, Sidney: Bloomsbury.

Esling, J. (2017). The laryngeal articulator's influence on voice quality and vowel quality. (C. Bertini, C. Celata, G. Lenoci, C. Meluzzi, & I. Ricc, Eds.) *Fattori sociali e biologici nella variazione fonetica [Social and biological factors in speech variation],Studi AISV 3*, pp. 13-26. https://doi.org/10.17469/O2103AISV000001

Esling, J. H. (1978). Voice quality in Edinburgh: A sociolinguistic and phonetic study. *Doctoral dissertation, University of Edinburgh, Scotland*.

Esling, J. H. (1996). Pharyngeal consonants and the aryepiglottic sphincter. *Journal of the International Phonetic Association, 26*(2), 65-88. https://doi.org/10.1017/S0025100300006125

Esling, J. H. (1999). Voice Quality Settings of the Pharynx. *Proceedings of the 14th International Congress of Phonetic Sciences, San Francisco, 3*, pp. 2449-2452.

Esling, J. H. (2000). Crosslinguistic aspects of voice quality. In R. D. Kent, & M. J. Ball (Eds.), *Voice quality measurement* (pp. 25-36). San Diego: Singular Publishing Group.

Esling, J. H., & Clayards, J. A. (1999). Laryngoscopic Analysis of Pharyngeal Articulatons and Larynx-height Voice Quality Settings. In A. Braun (Ed.), *Advances in Phonetics: Proceedings of the International Phonetic Sciences Conference (IPS), Bellingham, WA, June 27-30, 1998* (pp. 22-33). Stuttgart: Franz Steiner Verlag.

Esling, J. H., & Harris, J. G. (2005). States of the Glottis: An Articulatory Phonetic Model Based on Laryngoscopic Observations. In W. J. Hardcastle, & J. Mackenzie Beck (Eds.), *A Figure of Speech: A Festschrift for John Laver* (pp. 347-383). Mahwah, New Jersey: Lawrence Erlbaum Associates Publishers.

Esling, J. H., & Moisik, S. R. (2022). Voice Quality. In R.-A. Knight, & J. Setter (Eds.), *The Cambridge Handbook of Phonetics* (pp. 237-257). Cambridge: Cambridge University Press.

Esling, J. H., & Wong, R. F. (1983). Voice Quality Settings and the Teaching of Pronunciation. *TESOL Quarterly, 17*(1), 89-95. https://doi.org/10.2307/3586426

Esling, J. H., Heap, L. M., Snell, R. C., & Dickson, B. C. (1994). Analysis of pitch-dependence of pharyngeal, faucal, and larynx-height voice quality settings. *Proceedings of the 3rd International Conference on Spoken Language Processing (ICSLP 1994), Yokohama, Japan*, (pp. 1475–1478). https://doi.org/10.21437/ICSLP.1994-383

Esling, J. H., Moisik, S. R., Benner, A., & Crevier-Buchman, L. (2019). *Voice Quality: The Laryngeal Articulator Model.* Cambridge: Cambridge University Press.

Esposito, C. M. (2005). An acoustic and electroglottographic study of phonation in Santa Ana del Valle Zapotec. *Poster presented at the 79th meeting of the Linguistic Society of America.*

Esposito, C. M. (2012). An acoustic and electroglottographic study of White Hmong tone. *Journal of Phonetics, 40*(3), 466-476. https://doi.org/10.1016/j.wocn.2012.02.007

Esposito, C. M., & Khan, S. (2020). The cross-linguistic patterns of phonation types. *Language and Linguistics Compass, 14*(12). https://doi.org/10.1111/lnc3.12392

Evett, I. W. (1991). Interpretation: A Personal Odyssey. In C. G. Aitken, & D. A. Stoney, *The Use of Statistics in Forensic Science* (pp. 9-22). Chichester: Ellis Horwood Ltd.

Fant, G. (1956). On the predictability of formant levels and spectrum envelopes from formant frequencies. In M. Halle, H. Lunt, & H. McLean (Eds.), *For Roman Jacobson* (pp. 109-120). The Hague: Mouton.

Fant, G. (1960). *Acoustic theory of speech produclroz.* The Hague: Mouton.

Farrús, M., Eriksson, E., Sullivan, K. P., & Hernando, J. (2006). Dialect imitations in speaker recognition. *Proceedings of the 2nd European IAFL Conference on Forensic Linguistics, Language and the Law*, (pp. 347-353). Barcelona.

Faundez-Zanuy, M., & Satué-Villar, A. (2006). Speaker Recognition Experiments on a Bilingual Database. *Proceedings of IV Jornadas en Tecnologias del Habla (4JTH)*, (pp. 261-264). Zaragoza, Spain.

Fecher, N., & Watt, D. (2011). Speaking under cover: The effect of face-concealing garments on spectral properties of fricatives. *Proceedings of the XVIIth ICPhS*, (pp. 663-666). Hong Kong.

Ferreira Engelbert, A. P. (2014). Cross-Linguistic Effects on Voice Quality: A Study on Brazilians' Production of Portuguese and English. *Proceedings of the International Symposium on the Acquisition of Second Language Speech Concordia Working Papers in Applied Linguistics*, *5*, pp. 157-170.

Ferreira Engelbert, A. P., Kluge, D. C., Silva, P., & Hercilia, A. (2016). Línguas diferentes, vozes distintas: evidências da fala. *Ilha do Desterro, 69*(1), 33-48. https://doi.org/10.5007/2175-8026.2016v69n1p33

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378–382. https://doi.org/10.1037/h0031619

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed.). Hoboken, NJ: Wiley.

Fleming, D., Giordano, B. L., Caldera, R., & Belin, P. (2014). A language-familiarity effect for speaker discrimination without comprehension. *Psychological and Cognitive Sciences, 111*(38), 13795–13798. https://doi.org/10.1073/pnas.1401383111

Flynn, S., Foley, C., & Vinnitskaya, I. (2005). New Paradigm for the Study of Simultaneous v. Sequential Bilingualis. *Proceedings of the 4th International Symposium on Bilingualism* (pp. 768-774). Somerville, MA: Cascadilla Press.

Foulkes, P. (2010). Exploring social-indexical knowledge: A long past but a short history. *Laboratory Phonology, 1*(1), 5-39. https://doi.org/10.1515/labphon.2010.003

Foulkes, P., & French, P. (2001). Forensic Phonetics and Sociolinguistics. In R. Mesthrie (Ed.), *Concise Encyclopedia of Sociolinguistics* (pp. 329-332). Amsterdam: Elsevier.

Franco-Perdoso, J., & González-Rodríguez, J. (2016). Feature-based likelihood ratios for speaker recognition from linguistically-constrained formant-based i-vectors. *Proceedings of Odyssey 2016, Spain*, (pp. 237-244). https://doi.org/10.21437/Odyssey.2016-34

Fraser, J., Pengilly, A., & Mok, Q. (1998). Long-term ventilator-dependent children: a vocal profile analysis. *Pediatric Rehabilitation, 2*(2), 71-75. https://doi.org/10.3109/17518429809068158

French, P. (1994). An overview of forensic phonetics with particular reference to speaker identification. *Forensic Linguistics, 1*(2), 169-181. https://doi.org/10.1558/ijsll.v1i2.169

French, P. (1998). Mr. Akbar's nearest ear versus the Lombard reflex: a case study in forensic phonetics. *Forensic Linguistics, 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. In *English Phonetics* (pp. 271-286). Retrieved October 5th, 2018, from http://eprints.whiterose.ac.uk/117763/

French, P., & Harrison, P. (2007). Position Statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases, with a foreword by Peter French & Philip Harrison. *The International Journal of Speech, Language and the Law, 14*(1), 137-144. https://doi.org/10.1558/ijsll.v14i1.137

French, P., & Stevens, L. (2013). Forensic Speech Science. In M. M. Jones, & K. Rachael-Anne (Eds.), *The Bloomsbury Companion to Phonetics* (pp. 183-197). London/New York: Bloomsbury Publishing Plc.

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. *Proceedings of the 18th International Conference of Phonetic Sciences (ICPhS), Glasgow, Scotland.*

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. *The International Journal of Speech, Language and the Law, 17*(1), 143-152. https://doi.org/10.1558/ijsll.v17i1.143

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 Australasian Language Technology Association Workshop*, (pp. 39-47).

Furundžić, D. V. (2018). Ocena kvaliteta artikulacije glasova srpskog jezika primenom neuronskih mreža. *Doctoral dissertation, School of Electrical Engineering, University of Belgrade*.

Garellek, M. (2020). Acoustic Discriminability of the Complex Phonation System in !Xóõ. *Phonetica, 77*(2), 131-160. https://doi.org/10.1159/000494301

Garvin, P. L., & Ladefoged, P. (1963). Speaker Identification and Message Identification in Speech Recognition. *9*(4), 193-199. https://doi.org/10.1159/000258404

Gawell, M. J. (1981). The Effects of Various Drugs on Speech. *British Journal of Disorders of Communication, 16*(1), 51-57. https://doi.org/10.3109/13682828109011386

Geers, A., Davidson, L., Uchanski, R., & Nicholas, J. (2013). Interdependence of Linguistic and Indexical Speech Perception Skills in School-Aged Children with Early Cochlear Implantation. *Ear and Hearing, 34*(5), 562-574. https://www.doi.org/10.1097/AUD.0b013e31828d2bd6

Gfroerer, S. (2003). Auditory-Instrumental Forensic Speaker Recognition. *Eurospeech*, (pp. 705-708).

Gfroerer, S., & Baldauf, C. (2000). Sprechererkennung, Tonträgerauswertung und Autorenerkennung. In N. Beleke (Ed.), *Kriminalistische Kompetenz. Kriminalwissenschaften, kommentiertes Recht und Kriminaltaktik für Studium und Praxis* (pp. 3-16). Lübeck, Berlin, Essen: Schmidt-Römhild.

Gibbons, J., & Turell, T. M. (Eds.). (2008). *Dimensions of Forensic Linguistics.* Amsterdam/Philadelphia: John Benjamins.

Gobl, C., & Ni Chasiade, A. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication, 40*(1-2), 189-212. https://doi.org/10.1016/S0167-6393(02)00082-1

Gobl, C., & Ni Chasiade, A. (2010). Voice Source Variation and Its Communicative Functions. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), *The Handbook of Phonetic Sciences* (pp. 378-423). Blackwell Publishing Ltd. https://doi.org/10.1002/9781444317251.ch11

Goggin, J. P., Thompson, C. P., Strube, G., & Simental, L. R. (1991). The role of language familiarity in voice identification. *Memory & Cognition, 19*(5), 448–458. https://doi.org/10.3758/BF03199567

Gold, E. (2012). Articulation rate as a discriminant in forensic speaker comparisons. *UNSW Forensic Speech Science Conference 2012.* Sidney, Australia.

Gold, E. A. (2014). Calculating likelihood ratios for forensic speaker comparisons using phonetic and linguistic parameters. *PhD Thesis*. The Unversity of York, Department of Language and Linguistic Science.

Gold, E., & French, P. (2011). International Practices in Forensic Speaker Comparison. *The 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. *Journal of Speech, Language and the Law, 26*(1), 1-20. 10.1558/ijsll.38028

Gold, E., French, P., & Harrison, P. (2013). Examining long-term formant distributions as a discriminant in forensic speaker comparisons under a likelihood ratio framework. *Proceedings of Meetings on Acoustics, 19*(1), 060041. https://doi.org/10.1121/1.4800285

Goldman, J.-P. (2012). EasyAlign: an automatic phonetic alignment tool under Praat. *Proceedings of InterSpeech, September 2011 , Firenze, Italy*.

González Hautamäki, R., Kanervisto, A., Hautamäki, V., & Kinnunen, T. (2018). Perceptual evaluation of the effectiveness of voice disguise by age modification. *Speaker Odyssey 2018: The Speaker and Language Recognition Workshop.* https://doi.org/10.48550/arXiv.1804.08910

González Hautamäki, R., Sahidullah, M., Hautamäki, V., & Kinnunen, T. (2017). Acoustical and perceptual study of voice disguise by age modification in speaker verification. *Speech Communication, 95*, 1-15. https://doi.org/10.1016/j.specom.2017.10.002

Gonzalez, J., & Carpi, A. (2004). Early effects of smoking on the voice: a multidimensional study. *Medical science monitor : international medical journal of experimental and clinical research, 10*(12), CR649–CR656.

González-Rodríguez, J., Drygajlo, A., Ramos-Castro, D., Garcia-Gomar, M., & Ortega-Garcia, J. (2006). Robust Estimation, Interpretation and Assessment of Likelihood Ratios in Forensic Speaker Recognition. *Computer Speech and Language, 20*(2-3), 331 – 355. 10.1016/j.csl.2005.08.005

Gonzalez-Rodriguez, J., Fiérrez-Aguilar, J., & Ortega-Garcia, J. (2003). Forensic identification reporting using automatic speaker recognition systems. *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)*, (pp. II-94-II-96). https://doi.org/10.1109/ICASSP.2003.1202302

Gonzalez-Rodriguez, J., Fiérrez-Aguilar, J., Ortega-Garcia, J., & Lucena-Molina, J. J. (2002). Biometric Identification in Forensic Cases According to the Bayesian Approach. In M.

Tistarelli, J. Bigun, & A. K. Jain (Eds.), *Biometric Authentication. BioAW 2002. Lecture Notes in Computer Science* (Vol. 2359, pp. 177-185). Berlin, Heidelberg: Springer. https://doi.org/10.1007/3-540-47917-1_18

Gonzalez-Rodriguez, J., Gil, J., Perez, R., & Franco-Pedroso, J. (2014). What are we missing with ivectors? A perceptual analysis of i-vector-based falsely accepted trials. *Proceedings of Odyssey 2014: The Speaker and Language Recognition Workshop. Joensuu, Finland,* (pp. 33-40). https://doi.org/10.21437/Odyssey.2014-6

Good, I. J. (1991). Weight of evidence and the Bayesian likelihood ratio. In C. G. Aitken, & D. A. Stoney, *The Use Of Statistics In Forensic Science* (pp. 85-106). Chichester: Ellis Horwood Ltd.

Gordon, M., & Ladefoged, P. (2001). Phonation types: a cross-linguistic overview. *Journal of Phonetics, 29*(4), 383-406. https://doi.org/10.1006/jpho.2001.0147

Gracco, C., Sasaki, C. T., McGowan, R., Tierney, E., & Gore, J. (1994). Magnetic resonance imaging (MRI) in vocal tract research: Clinical application. *The Journal of the Acoustical Society of America, 95,* 2821. https://doi.org/10.1121/1.409684

Greisbach, R. (1999). Estimation of speaker height from formant frequencies. *Forensic Linguistics, 6*(2), 265-277. https://doi.org/10.1558/sll.1999.6.2.265

Grieve, J., Speelman, D., & Geeraerts, D. (2013). A multivariate spatial analysis of vowel formants in American English. *Journal of Linguistic Geography, 1*(1), 31-51. https://doi.org/10.1017/jlg.2013.3

Grigoras, C. (2005). Digital audio recording analysis: The electric network frequency (ENF) criterion. *The International Journal of Speech, Language and the Law, 12*(2), 63-76. https://doi.org/10.1558/sll.2005.12.1.63

Grosjean, F. (1982). *Life with Two Languages: An Introduction to Bilingualism.* Cambridge, Mass: Harvard University Press.

Grozdić, Đ., Jovičić, S. T., & Rajković, M. (2011). Uticaj kvaliteta glasa na verbalnu ekspresiju emocija. *Proceedings of the 9th Telecommunications forum TELFOR 2011, Belgrade, Serbia,* (pp. 663-666).

Guillemin, B. J., & Watson, C. (2008). Impact of the GSM mobile phone network on the speech signal: some preliminary findings. *The International Journal of Speech, Language and the Law, 15*(2), 193-218. https://doi.org/10.1558/ijsll.v15i2.193

Gwet, K. L. (2019). irrCAC: Computing Chance-Corrected Agreement Coefficients (CAC). *R package version 1.0*. Retrieved May 2023, from https://CRAN.R-project.org/package=irrCAC

Gwet, K. L. (2021). *Handbook of Inter-Rater Reliability, 5th Edition. Volume 1: Analysis of Categorical Ratings.* Maryland, USA: AgreeStat Analytics.

Hammarberg, B. (2000). Voice research and clinical needs. *Folia Phoniatrica et Logopaedica, 52*(1-3), 93-102. https://doi.org/10.1159/000021517

Han, J.-Y., Hsiao, C.-J., Zheng, W.-Z., Weng, K.-C., Ho, G.-M., Chang, C.-Y., . . . Lai, Y.-H. (in press). Enhancing the Performance of Pathological Voice Quality Assessment System Through the Attention-Mechanism Based Neural Network. *Journal of Voice*. https://doi.org/10.1016/j.jvoice.2022.12.026

Hansen, J. H., & Clements, M. A. (1987). Evaluation of speech under stress and emotional conditions. *The Journal of the Acoustical Society of America, 82*(S1), S17-S18. https://doi.org/10.1121/1.2024686

Hanson, H. M. (1997). Glottal characteristics of female speakers: Acoustic correlates. *Journal of the Acoustical Society of America, 101*(1), 466-481. https://doi.org/10.1121/1.417991

Hardcastle, W. J., & Gibbon, F. (1997). Electropalatography and its Clinical Applications. In M. J. Ball, & C. Code (Eds.), *Instrumental Clinical Phonetics* (pp. 149-193). London: Whurr Publishers. https://doi.org/10.1002/9780470699119.ch6

Harmegnies, B., & Landercy, A. (1985). Language features in the long term average spectrum. *Revue de Phonétique Appliquée, 73-74-75*, pp. 69-80.

Harmegnies, B., & Landercy, A. (1988). Intra-speaker variability of the long term speech spectrum. *Speech Communication, 7*(1), 81-86. https://doi.org/10.1016/0167-6393(88)90023-4

Harmegnies, B., Bruyninckx, M., Llisteri, J., & Pock, D. (1989). Effects of language change on voice quality. An experimental contribution to the study of Catalan-Castilian case. *Proceedings of the First European Conference on Speech Communication and Technology EUROSPEECH 1989, Paris, France*, (pp. 2489-2492).

Havenhill, J. (2019). Articulatory Strategies for Back Vowel Fronting in American English. *Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS2019), Melbourne, Australia, 5-9 August 2019*, (pp. 1540-1544).

He, L., Zhang, Y., & Dellwo, V. (2019). Between-speaker variability and temporal organization of the first formant. *The Journal of the Acoustical Society of America, 145*, EL209. https://doi.org/10.1121/1.5093450

Heeren, W. (2020). The contribution of dynamic versus static formant information in conversational speech. *The 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. *Proceedings of the International Association for Forensic Phonetics and Acoustics conference, Zurich, Switzerland*, (pp. 39-40).

Heinrich, N., D'Alessandro, C., Doval, B., & Castellengo, M. (2004). On the use of the derivative of electroglottographic signals for characterization of non-pathological phonation. *The Journal of the Acoustical Society of America, 115*(3), 1321–1332. https://doi.org/10.1121/1.1646401

Henton, C., & Bladon, A. (1985). Breathiness in normal female speech: inefficiency versus desirability. *Language and Communication, 5*(3), 221-227. https://doi.org/10.1016/0271-5309(85)90012-6

Henton, C., & Bladon, A. (1988). Creak as a Sociophonetic Marker. In L. M. Hyman, & C. N. Li (Eds.), *Language, Speech, and Mind: Studies in Honour of Victoria A. Fromkin* (pp. 3-29). New York, London: Routledge.

Herbst, C. T. (2020). Electroglottography – An Update. *Journal of Voice, 34*(4), 503-526. https://doi.org/10.1016/j.jvoice.2018.12.014

Hewlett, N., & Beck, J. (2006). *An Introduction to the Science of Phonetics.* London and New York: Routledge.

Hillenbrand, J., & Houde, R. A. (1996). Acoustic Correlates of Breathy Vocal Quality: Dysphonic Voices and Continuous Speech. *Joumal of Speech and Hearing Research, 39*(6), 311-321. https://doi.org/10.1044/jshr.3902.311

Hillenbrand, J., Cleveland, R. A., & L, E. R. (1994). Acoustic Correlates of Breathy Vocal Quality. *Joumal of Speech and Hearing Research, 37*(4), 769-778. https://doi.org/10.1044/jshr.3704.769

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America, 97*(5), 3099–3111. https://doi.org/10.1121/1.411872

Hirano, M. (1981). *Clinical examination of voice.* Vienna and New York: Springer-Verlag.

Hirose, H. (1999). Investigating the Physiology of Laryngeal Structures. In W. J. Hardcastle, & J. Laver (Eds.), *The Handbook of Phonetic Sciences* (pp. 1-16). Blackwell Reference Online.: Blackwell Publishing.

Hofstra University. (2022). *MASTER OF ARTS IN LINGUISTICS*. Retrieved October 28, 2021, from Hofstra University: https://www.hofstra.edu/forensic-linguistics-master/

Hollien, H. (1983). Forensic Communication: An Emerging Specialty. *Criminal Defense,, 10*, 22-29.

Hollien, H. (1990). *The Acoustics of Crime: The New Science of Forensic Phonetics.* New York: Springer.

Hollien, H. (2002). *Forensic Voice Identification.* London and New York: Academic Press.

Hollien, H. (2012). About Forensic Phonetics. *Linguistica. 52,* pp. 27-53. Ljubljana: Tiskana. https://doi.org/10.4312/linguistica.52.1.27-53

Hollien, H. (2013). Barriers to Progress in Speaker Identification. *Linguistic Evidence in Security, Law and Intelligence, 1*(1), 1-23. 10.5195/lesli.2013.3

Hollien, H., & Michel, J. F. (1968). Vocal fry as a phonational register. *Journal of Speech and Hearing Research, 11*(3), 600–604. https://doi.org/10.1044/jshr.1103.600

Hollien, H., & Schwartz, R. (2000). Aural-perceptual speaker identification: problems with noncontemporary samples. *Forensic Linguistics, 7*(2), 199–211. https://doi.org/10.1558/sll.2000.7.2.199

Hollien, H., & Shipp, T. (1972). Speaking Fundamental Frequency and Chronologic Age in Males. *Journal of Speech and Hearing research, 15*(1), 155-159. https://doi.org/10.1044/jshr.1501.155

Hollien, H., de Jong, G., Martin, C. A., Schwartz, R., & Liljegren, K. (2001a). Effects of ethanol intoxication on speech suprasegmentals. *The Journal of the Acoustical Society of America, 110*(6), 3198–3206. https://doi.org/10.1121/1.1413751

Hollien, H., Harnsberger, J. D., Martin, C. A., Hill, R., & Alderman, G. (2009). Perceiving the Effects of Ethanol Intoxication on Voice. *Journal of Voice, 23*(5), 552-559. https://doi.org/10.1016/j.jvoice.2007.11.005

Hollien, H., Liljegren, K., Martin, C. A., & de Jong, G. (2001b). Production of intoxication states by actors--acoustic and temporal characteristics. *Journal of Forensic Sciences, 46*(1), 68–73.

Hollien, H., Martin, C. A., & de Jong, G. (1998). Production of intoxication states by actors: perception by lay listeners. *Journal of forensic sciences, 43*(6), 1153–1162.

Hollien, H., Moore, P., Wendahl, R. W., & Michel, J. F. (1966). On the nature of vocal fry. *Journal of Speech and Hearing Research, 9*(2), 245–247. https://doi.org/10.1044/jshr.0902.245

Holmes, E. (2023). Towards Phonetically-Informed Automatic Speaker Recognition (ASR). *Doctoral dissertation. University of York.*

Home Office. (2003). Advice on the use of voice identification parades. *UK Home Office Circular 057/2003*. Crime Reduction and Community Safety Group, Police Leadership and Powers Unit.

Honikman, B. (1964). Articulatory settings. In D. Abercrombie, D. B. Fry, P. MacCarthy, N. C. Scott, & J. L. Trim (Eds.), *In honour of Daniel Jones: Papers contributed on the occasion of his eightieth birthday, 12 September 1961* (pp. 73-84). London: Longman.

Horiguchi, S., & Bell-Berti, F. (1987). The Velotrace: a device for monitoring velar position. *The Cleft Palate Journal, 24*(2), 104-111.

Hoskin, J. A. (2022). Shifting the Burden: towards new tests for Language Analysis in the Asylum Procedure. *PhD thesis, University of York.*

Hoskin, J., Cambier-Langeveld, T., & Foulkes, P. (2020). Improving objectivity, balance and forensic fitness in LAAP : a response to Matras. *International Journal of Speech, Language and the Law, 26*(2), 257-277. https://doi.org/10.1558/ijsll.39208

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., & Foulkes, P. (2014). Variability in analyst decisions during the computation of numerical likelihood ratios. *The International Journal of Speech, Language and the Law, 21*(2), 279-315. https://doi.org/10.1558/ijsll.v21i2.279

Hughes, V., Foulkes, P., Harrison, P., Wormald, J., Xu, C., van der Vloed, D., & Kelly, F. (2022a). Person-specific automatic speaker recognition: understanding the behaviour of individuals for applications of ASR. *A poster presented at the 30thInternational Association for Forensic Phonetics and Acoustics Conference, Prague, Czechia. 10-13 July 2022.*

Hughes, V., Foulkes, P., Harrison, P., Wormald, J., Xu, C., van der Vloed, D., & Kelly, F. (2022b). Person-specific automatic speaker recognition: understanding the behaviour of individuals for applications of ASR. *UK Speech Conference, University of Edinburgh, 5-6 September 2022.*

Hughes, V., Harrison, P. T., Foulkes, P., French, J. P., & Gully, A. J. (2020). Effects of formant settings and channel mismatch on semi-automatic systems in forensic voice comparison. *Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS)*, (pp. 3080-3084).

Hughes, V., Harrison, P., Foulkes, P., French, P., Kavanagh, C., & San Segundo, E. (2017). The complementarity of automatic, semi-automatic, and phonetic measures of vocal tract output in forensic voice comparison. *A paper presented at the 26th Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA), Split, Croatia.*

Hughes, V., Harrison, P., Foulkes, P., French, P., Kavanagh, C., & San Segundo, E. (2017b). Mapping across feature spaces in forensic voice comparison: The contribution of auditory-based voice quality to (semi-)automatic system testing. *Proceedings of Interspeech 2017*, (pp. 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 semiautomatic forensic voice compariosn system. *Proceedings of Interspeech*, (pp. 227-223). Hyderabad.

Hughes, V., Wood, S., & Foulkes, P. (2016). Strength of forensic voice comparison evidence from the acoustics of filled pauses. *The International Journal of Speech, Language and the Law, 23*(1), 99-132. https://doi.org/10.1558/ijsll.v23i1.29874

Hyman, L. M., & Plank, F. (Eds.). (2018). *Phonological typology. Phonology and Phonetics 23.* Berlin & Boston: De Gruyter Mouton.

IAFPA. (2020). *Code of Practice.* Retrieved February 2020, from IAFPA: http://www.iafpa.net/about/code-of-practice/

IAFPA. (2022). *Resolutions*. Retrieved September 2022, from International Association of Forensic Phonetics and Acoustics: https://www.iafpa.net/about/resolutions/

IELTS. (2023a). *IELTS and the CEFR (Common European Framework of Reference).* Retrieved June 2023, from IELTS.org: https://www.ielts.org/-/media/pdfs/ielts-and-the-cefr.ashx

IELTS. (2023b). *IELTS Speaking Key Assessment Criteria.* Retrieved June 2023, from IELTS.org: https://www.ielts.org/-/media/pdfs/ielts-speaking-key-assessment-criteria.ashx

IELTS. (2023c). *IELTS Speaking Band Descriptors.* Retrieved June 2023, from IELTS.org: https://www.ielts.org/-/media/pdfs/ielts-speaking-band-descriptors.ashx

In Park, N., Shim, K.-S., Lee, J., Kim, J.-H., Ho Lim, S., Byun, J., . . . Jeon, O.-Y. (2022). Advanced forensic procedure for the authentication of audio recordings generated by Voice Memos application of iOS14. *Journal of Forensic Sciences, 67*(4), 1534-1549. https://doi.org/10.1111/1556-4029.15016

Irfan, M., Ramdania, D., Hasni, N., Budiman, I., Maylawati, D., & Manaf, K. (2021). Similarity Level Analysis of the Voices of Twins Using the Analysis of Variance and Likelihood Ratio Methods. *Proceedings of the 7th International Conference on Wireless and Telematics (ICWT)*, (pp. 1-5). https://doi.org/10.1109/ICWT52862.2021.9678443

Iseli, M., Shue, Y.-L., & Alwan, A. (2007). Age, sex, and vowel dependencies of acoustic measures related to the voice source. *Journal of the Acoustical Society of America*, 2283-2295. https://doi.org/10.1121/1.2697522

Ishi, C. T., Ishiguro, H., & Hagita, N. (2010). Acoustic, Electroglottographic and Paralinguistic Analyses of "Rikimi" in Expressive Speech. *Proceedings of Speech Prosody 2010 Chicago, IL, USA.*

Ishihara, S. (2014). A likelihood ratio-based evaluation of strength of authorship attribution evidence in SMS messages using N-grams. *The International Journal of Speech, Language and the Law, 21*(1), 23-50. https://doi.org/10.1558/ijsll.v21i1.23

Ishihara, S. (2017). Strength of forensic text comparison evidence from stylometric features: a multivariate likelihood ratio-based analysis. *The International Journal of Speech, Language and the Law, 24*(1), 67-98. https://doi.org/10.1558/ijsll.30305

Ishihara, S., & Kinoshita, Y. (2008). How Many Do We Need? Exploration of the Population Size Effect on the Performance of Forensic Speaker Classification. *Proceedings of the Interspeech 2008, Brisbane Australia*, (pp. 1941-1944).

Ivanović, M., & Kašić, Z. (2011a). Intenzitetske varijacije u govornoj ekspresiji primarnih emocija. *Peti međunarodni naučni skup: Specijalna edukacija i rehabilitacija danas. Zbornik radova.* (pp. 306-314). Beograd: Univerzitet u Beogradu, Fakultet za specijalnu edukaciju i rehabilitaciju.

Ivanović, M., & Kašić, Z. (2011b). Variranje trajanja segmenata u govornoj ekspresiji emocija. *Specijalna edukacija i rehabilitacija, 10*(2), 341-353.

Ivić, P. (1956). *Dijalektologija srpskohrvatskog jezika. Uvod u štokavsko narečje.* Novi Sad: Matica Srpska.

Ivić, P., & Lehiste, I. (2002). *O srpskohrvatskim akcentima.* (D. Petrović, Ed.) Novi Sad: Izdavačka knjižarnica Zorana Stojanovića.

Jacewicz, E., Fox, R. A., O'Neill, C., & Salmons, J. (2009). Articulation rate across dialect, age, and gender. *Language Variation and Change, 21*(2), 233-256. https://doi.org/10.1017/S0954394509990093

Jerotijević Tišma, D. (2020a). Fonetski parametri izražavanja emocija na srpskom (J1) i engleskom (J2) jeziku. In M. Kovačević, & J. Petković (Ed.), *Ekspresivnost u srpskom jeziku, 1*, pp. 303-318.

Jerotijević Tišma, D. (2020b). Fonetsko-prozodijske karakteristike i komunikativna funkcija rečenice. In M. Kovačević, & J. Petković (Ed.), *Srpski jezik, književnost, umetnost.*

Zbornik radova sa XV međunarodnog naučnog skupa održanog na Filološko-umetničkom fakultetu u Kragujevcu, (pp. 141-157).

Jessen, M. (2007). Speaker Classification in forensic phonetics and acoustics. In C. Müller (Ed.), *Speaker Classification I: Fundamentals, Features, and Methods* (pp. 180-204). Berlin / Heidelberg / New York: Springer.

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. (2010). The forensic phonetician: Forensic speaker identification by experts. In M. Coulthard, & A. Johnson (Eds.), *The Routledge Handbook of Forensic Linguistics* (pp. 378-394). Abingdon and New York: Routledge.

Jessen, M. (2018). Forensic voice comparison. In J. Visconti (Ed.), *Handbook of Communication in the Legal Sphere* (pp. 219-255). Berlin, Boston: De Gruyter Mouton. https://doi.org/10.1515/9781614514664-012

Jessen, M. (2021). MAP adaptation characteristics in forensic long-term formant analysis. *Proceedings of Interspeech 2021*, (pp. 411-415). http://dx.doi.org/10.21437/Interspeech.2021-1697

Jessen, M., & Becker , T. (2010). Long-Term Formant Term Formant Distribution as a Forensic Phonetic Feature. *ASA - 2nd Pan-American / Iberian Meeting on Acoustics.* Cancún, Mexico, Nov 15-19. 10.1121/1.3508452

Jessen, M., Köster, O., & Gfroerer, S. (2005). Influence of vocal effort on average and variability of fundamental frequency. *The International Journal of Speech, Language and the Law, 12*(2), 174-233. https://doi.org/10.1558/sll.2005.12.2.174

Johnson, K. A., Babel, M., & Fuhrman, R. A. (2020). Bilingual acoustic voice variation is similarly structured across languages. *Proceedings of Interspeech 2020, Shanghai, China*, (pp. 2387-2391). https://doi.org/10.21437/Interspeech.2020-3095

Jovičić, S. T., Jovanović, N., Subotić, M., & Grozdić, Đ. (2015). Impact of mobile phone usage on speech spectral features: some preliminary findings. *The International Journal of Speech, Language and the Law, 22*(1), 111-126. https://doi.org/10.1558/ijsll.v22i1.17880

Jovičić, S., & Grozdić, Đ. (2014). Arguments for auditory-instrumetal approach in forensic speaker recognition. *Proceedings of The International Scientific Conference "Archibald Reiss Days, Belgrade, Serbia*, (pp. 355-64).

Jovičić, T. S. (2001). Forenzički aspekti prepoznavanja govornika. *Nauka, tehnika, bezbednost, 1*, 41-60.

Kahil, Y., Zergat, K. Y., & Amrouche, A. (2018). Forensic Voice Comparison with ALIZE/LIA_RAL toolkit. *Proceedings of the 4th IEEE International Conference on Signal, Image, Vision and their Applications, Guelma, Algeria.*

Kainada, E., & Lengeris, A. (2015). Native language influences on the production of second-language prosody. *Journal of the International Phonetic Association, 45*(3), 269-287. https://doi.org/10.1017/S0025100315000158

Kao, S.-M., & Wang, W.-C. (2014). Lexical and organizational features in novice and experienced ELF presentations. *Journal of English as a Lingua Franca, 3*(1), 49-79. https://doi.org/10.1515/jelf-2014-0003

Kašić, Z., & Đorđević, J. P. (2009). Zašto je lingvistika postala forenzička veština. U D. Radovanović (Ur.), *Istraživanja u specijalnoj pedagogiji* (str. 469-482). Beograd: Univerzitet u Beogradu - Fakultet za specijalnu edukaciju i rehabilitaciju.

Kašić, Z., & Đorđević, J. P. (2009a). Ostaci automatizma artikulacione baze kao forenzički markeri. *ETRAN 2009 Zbornik Radova*, (pp. AK4.2-1-4). Vrnjačka Banja.

Kašić, Z., & Ivanović, M. (2011). Govorni parametri i tuga. *Specijalna edukacija i rehabilitacija*, *10*(4), 745-763.

Kassambara, A. (2023). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.6.0. Retrieved December 2023, from https://CRAN.R-project.org/package=ggpubr

Keating, P., & Esposito, C. (2007). Linguistic Voice Quality. *UCLA Working papers in Phonetics, 105*, 85-91. Retrieved August 2022, from https://escholarship.org/uc/item/04r5q6qn

Keating, P., Esposito, C. M., Garellek, M., Khan, S., & Jianjing, K. (2010). Phonation Contrasts Across Languages. *UCLA Working Papers in Phonetics,, 108*, 188-202.

Keating, P., Garellek, M., & Kreiman, J. (2015). Acoustic properties of different kinds of creaky voice. *Acoustic properties of different kinds of creaky voice. Proceedings of the 18th International Conference of Phonetic Sciences (ICPhS) Glasgow, Scotland.*

Kelly, F., & Harte, N. (2015). Forensic comparison of ageing voices from automatic and auditory perspectives. *The International Journal of Speech, Language and the Law, 22*(2), 178-202. https://doi.org/10.1558/ijsll.v22i2.21760

Kempster, G. B., Gerratt, B. R., Verdolini Abbott, K., Barkmeier-Kraemer, J., & Hillman, R. E. (2009). Consensus Auditory-Perceptual Evaluation of Voice: Development of a Standardized Clinical Protocol. *American Journal of Speech-Language Pathology, 18*(2), 124-132. https://doi.org/10.1044/1058-0360(2008/08-0017)

Kersta, L. G. (1962). Voiceprint Identification. *The Journal of the Acoustical Society of America, 34*(5), 725-725. https://doi.org/10.1121/1.1937211

Khan, S. u. (2010). Breathy phonation in Gujarati: An acoustic and electroglottographic study. *The Journal of the Acoustical Society of America, 127*(3_supplement), 2021. https://doi.org/10.1121/1.3385281

Kinoshita, Y. (2001). Testing realistic forensic speaker identification in Japanese: a likelihood ratio based approach using formants. *PhD Thesis*. the Australian National University.

Kinoshita, Y. (2014). Looking into the real world: Likelihood ratio variability under forensically realistic conditions. *Proceedings of the 15th Australasian International Conference on Speech Science and Technology*.

Kinoshita, Y., & Ishihara, S. (2014). Background population: How does it affect LR-based forensic voice comparison? *The International Journal of Speech, Language and the Law, 21*(2), 191-224. https://doi.org/10.1558/ijsll.v21i2.191

Kleber, F., Harrington, J., & Reubold, U. (2011). The relationship between the perception and production of coarticulation during a sound change in progress. *Language and Speech, 55*(3), 383-405. https://doi.org/10.1177/0023830911422194

Klug, K. (2023). Assessing a speaker's voice quality for forensic purposes: Using the example of creaky voice and breathy voice. *PhD thesis, University of York*.

Klug, K., & Niermann, M. (2024). Assessing the suitability of F0 estimators with respect to recording condition and voice quality. *Unpublished manuscript*.

Klug, K., Kirchhübel, C., Foulkes, P., & French, P. J. (2019). Analysing breathy voice in forensic speaker comparison: Using acoustics to confirm perception. *Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS) 2019, Melbourne, Australia*, (pp. 795-799).

Klug, K., Kirchhübel, C., Foulkes, P., Braun, A., & French, P. (in press). Assessing creaky voice quality for forensic purposes. *Proceedings of the Aarhus International Conference on Voice Studies. Sciendo.*

Kockmann, M., Farrell, K., Colibro, D., Vair, C., Alexander, A., & Kelly, F. (2021). Voice biometrics: perspective from the industry. In C. García-Mateo, & G. Chollet (Eds.), *Voice Biometrics: Technology, trust and security* (pp. 163-185). The Institution of Engineering and Technology. https://doi.org/10.1049/PBSE012E_ch8

Köster, O., & Köster, J.-P. (2004). The Auditory-Perceptual Evaluation of Voice Quality in Forensic Speaker Recognition. *Phonetician, 89*, 9-37.

Köster, O., & Schiller, N. O. (1997). Different influences of the native language of a listener on speaker recognition. *Forensic Linguistics, 4*(1), 18–28. https://doi.org/10.1558/ijsll.v4i1.18

Köster, O., Jessen, M., Khairi, F., & Eckert, H. (2007). Auditory-perceptual identification of voice quality by expert and non-expert listeners. *Proceedings of the 16th International Congress of Phonetic Sciences,* (pp. 1845–1848). Saarbrücken.

Kostić, Đ., Vladisavljević, S., & Popović, M. (1983). *Testovi za ispitivanje govora i jezika.* Beograd: Zavod za udžbenike i nastavna sredstva.

Krebs, P., & Braun, A. (2015). Long Term Formant measurements in bilingual speakers. *Proceedings of the IAFPA Annual Conference, 8 - 10 July 2015.* Leiden.

Kreidler, C. W. (1989, 2004). *The Pronunciation of English: A Course Book* (2nd ed.). Oxford: Blackwell Publishing.

Kreiman, J., Gerratt, B. R., Garellek, M., Samlan, R., & Zhang, Z. (2014). Toward a unified theory of voice production and perception. *Loquens, 1*(1), e009. http://dx.doi.org/10.3989/loquens.2014.009

Kreiman, J., Gerratt, B., & Antoñanzas-Barroso, N. (2007). Measures of the Glottal Source Spectrum. *Journal of Speech, Language, and Hearing Research, 50*(3), 595–610. https://doi.org/10.1044/1092-4388(2007/042)

Kreiman, J., Shue, Y.-L., Chen, G. C., Iseli, M., Gerratt, B. R., Neubauer, J., & Alwan, A. (2012). Variability in the relationships among voice quality, harmonic amplitudes, open quotient, and glottal area waveform shape in sustained phonation. *The Journal of the Acoustical Society of America, 132*(4), 2625–2632. https://doi.org/10.1121/1.4747007

Kuang, J. (2010). An acoustic and electroglottographic study of phonation contrast in Yi. *The Journal of the Acoustical Society of America, 127*(3_supplement), 2022. https://doi.org/10.1121/1.3385285

Kubic, T. A., & Ann, B. J. (1991). Quality assurance in the forensic laboratory. In C. G. Aitken, & D. A. Stoney (Eds.), *The Use of Statistics in Forensic Science* (pp. 207-234). Chichester: Ellis Howood Ltd.

Kumar, R., Ranjan, R., Singh, S. K., Kala, R., Shukla, A., & Tiwari, R. (2009). Multilingual speaker recognition using neural network. *Proceedings of the Frontiers of Res. on Speech and Music, FRSM 2009,* (pp. 1-8). Gwalior, India.

Kummer, A. W., Curtis, C., Wiggs, M., Lee, L., & Strife, J. L. (1992). Comparison of velopharyngeal gap size in patients with hypernasality, hypernasality and nasal emission, or nasal turbulence (rustle) as the primary speech characteristic. *The Cleft*

palate-craniofacial journal, 29(2), 152-156. https://doi.org/10.1597/1545-1569_1992_029_0152_covgsi_2.3.co_2

Künzel, H. (1994). On the Problem of Speaker Identification by Victims and Witnesses. *Forensic Linguistics, 1*, 45-58. https://doi.org/10.1558/ijsll.v1i1.45

Künzel, H. (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

Künzel, H. J. (1989). How Well Does Average Fundamental Frequency Correlate with Speaker Height and Weight? *Phonetica, 46*(1-3), 117-125. https://doi.org/10.1159/000261832

Künzel, H. J. (2000). Effects of voice disguise on speaking fundamental frequency. *Forensic Linguistics, 7*(2), 150-179. https://doi.org/10.1558/sll.2000.7.2.149

Künzel, H. J. (2001). Beware of the 'telephone effect': the influence of telephone transmission on the measurement of formant frequencies. *Forensic Linguistics, 8*(1), 80-99. https://doi.org/10.1558/ijsll.v8i1.80

Labov, W., Ash, S., & Boberg, C. (2005). *The Atlas of North American English: Phonetics, Phonology and Sound Change.* Berlin, New York: De Gruyter Mouton. https://doi.org/10.1515/9783110167467

Ladefoged, P. (1971). *Preliminaries to Linguistic Phonetics.* Chicago, London: The University of Chicago Press.

Ladefoged, P. (2001). *Vowels and Consonants: An Introduction to the Sounds of Language.* Massachusetts, Oxford : Blackwell Publishers.

Ladefoged, P., & Johnson, K. (2011). *A Course in Phonetics* (6th ed.). Wadsworth: Cengage Learning.

Ladefoged, P., & Maddieson, I. (1996). *The Sounds of the World's Languages.* Oxford: Blackwell.

Laerd Statistics. (2023). *Fleiss' kappa using SPSS Statistics*. Retrieved June 2023, from Laerd Statistics: https://statistics.laerd.com/spss-tutorials/fleiss-kappa-in-spss-statistics.php

Lavan, N., Burston, L. F., Ladwa, P., Merriman, S. E., Knight, S., & McGettigan, C. (2019). Breaking voice identity perception: Expressive voices are more confusable for listeners. *Quarterly Journal of Experimental Psychology, 72*(9), 2240–8. https://doi.org/10.1177/1747021819836890

Laver, J. (1968). Voice Quality and Indexical Information. *British Journal of Disorders of Communication, 3*(1), 43-54. https://doi.org/10.3109/13682826809011440

Laver, J. (1980). *The Phonetic Description of Voice Quality.* Cambridge: Cambridge University Press.

Laver, J. (1991). *The gift of speech : papers in the analysis of speech and voice.* Edinburgh: Edinburgh University Press.

Laver, J. (1991a). The Semiotic Nature of Phonetic Data. In J. Laver, *The Gift of Speech: Papers in the Analysis of Speech and Voice* (pp. 162-170). Edinburgh: Edinburgh University Press.

Laver, J. (1991b). The description of voice quality in general phonetic theory. In J. Laver, *The Gift of Speech: Papers in the Analysis of Speech and Voice* (pp. 184-208). Edinburgh: Edinburgh University Press.

Laver, J. (1991c). The Concept of Articulatory Settings: An Historical Overwiev. In J. Laver, *The Gift of Speech: Papers in the Analysis of Speech and Voice* (pp. 372-384). Edinburgh: Edinburgh University Press.

Laver, J. (1994). *Principles of Phonetics.* Cambridge: Cambridge University Press.

Laver, J. (2000). Phonetic evaluation of voice quality. In R. D. Kent, & M. J. Ball (Eds.), *Voice Quality Measurement* (pp. 37-48). San Diego: Singular Publishing Group.

Laver, J., & Trudgill, P. (1979). Phonetic and linguistic markers in speech. In K. R. Scherer, & H. Giles (Eds.), *Social Markers in Speech* (pp. 1-32). Cambridge: Cambridge University Press.

Laver, J., Wirz, S., Mackenzie, J., & Hillier, S. M. (1981). A perceptual protocol for the analysis of vocal profiles. *Edinburgh University Department of Linguistics Work in Progress, 14*, 139-155.

Lawrence, S., Nolan, F., & McDougall, K. (2008). Acoustic and perceptual effects of telephone transmission on vowel quality. *The International Journal of Speech, Language and the Law, 15*(2), 161-192. https://doi.org/10.1558/ijsll.v15i2.161

Lee, Y., & Kreiman, J. (2019). Within- and between-speaker acoustic variability: Spontaneous versus read speech. *A poster presented at the 178th Meeting of the Acoustical Society of America, San Diego, USA, Journal of the Acoustic Society of America, 146*(4_supplement), 3011. https://doi.org/10.1121/1.5137431

Lee, Y., Keating, P., & Kreiman, J. (2019). Acoustic voice variation within and between speakers. *Journal of the Acoustic Society of America, 146*(3), 1568–1579. https://www.doi.org/10.1121/1.5125134

Leemann, A., & Kolly, M.-J. (2015). Speaker-invariant suprasegmental temporal features in normal and disguised speech. *Speech Communication, 75*, 97-122. https://doi.org/10.1016/j.specom.2015.10.002

Lennes, M. (2022). SpeCT - Speech Corpus Toolkit for Praat: align_transcribed_utterance_intervals_with_sound.praat (Praat script). *Available via the GitHub repository at:*. Retrieved July 2023, from https://github.com/lennes/spect

Leykum, H. (2021). Voice Quality in Verbal Irony: Electroglottographic Analyses of Ironic Utterances in Standard Austrian German. *Proceedings of Interspeech 2021, Brno, Czechia*, (pp. 991-995). https://doi.org/10.21437/Interspeech.2021-452

Li, J., & Rose, P. (2012). Likelihood Ratio-based Forensic Voice Comparison with F-pattern and Tonal F0 from the Cantonese /Oy/ Diphthong. *n Proceedings of the 14th Australasian International Conference on Speech Science and Technology (SST 2012)*, (pp. 201-204). Sydney, Australia.

Li, P., Li, G., Han, J., Zhi, T., & Wang, D. (2020). Channel Mismatch Speaker Verification Based on Deep Learning and PLDA. *Journal of Physics: Conference Series, 2020 International Conference on Machine Learning and Computer Application, Shangri-La, China, 1682*, 012056. https://doi.org/10.1088/1742-6596/1682/1/012056

Lin, S. (2022). Observing and Measuring Speech Articulation. In R.-A. Knight, & J. Setter (Eds.), *The Cambridge Handbook of Phonetics* (pp. 362-386). Cambridge: Cambridge University Press.

Lindley, D. V. (1991). Probability. In C. G. Aitken, & D. A. Stoney, *The Use of Statistics in Forensic Science* (pp. 27-50). Chichester: Ellis Horwood Ltd.

Llamas, C., Harrison, P., Donnelly, D., & Watt, D. (2009). Effects of different types of face coverings on speech acoustics and intelligibility. *York Papers in Linguistic Series 2*(9), pp. 80-104.

Lo, J. (2022). fvclrr: Likelihood Ratio Calculation and Testing in Forensic Voice Comparison. R package version 1.1.4. Retrieved March 2024, from https://github.com/justinjhlo/fvclrr

Lo, J. H. (2021). Issues of Bilingualism in Likelihood Ratio-based Forensic Voice Comparison. *Doctoral dissertaton, University of York*.

Lo, J. J. (2021b). Cross-linguistic Speaker Individuality of Long-term Formant Distributions: Phonetic and Forensic Perspectives. *Proceedings of INTERSPEECH 2021*, (pp. 416-420). http://dx.doi.org/10.21437/Interspeech.2021-1699

Loakes, D., & Gregory, A. (2022). Voice Quality in Australian English. *JASA Express Letters*, 085201. https://doi.org/10.1121/10.0012994

Lombard, É. (1911). Le signe de l'élévation de la voix. *Annales des Maladies de l'Oreille, du Larynx, du Nez et du Pharynx, 37*, 101-119.

Luengo, I., Navas, E., Sainz, I., Saratxaga, I., Sanchez, J., Odriozola, I., & Hernaez, I. (2008). Text Independent Speaker Identification in Multilingual Environments. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC '08).* Marrakech, Marocco: European Language Association (ERLA).

Mackenzie Beck, J. (1988). Organic Variation and Voice Quality. *PhD Thesis, University of Edinburgh*.

Mackenzie Beck, J. (2005). Perceptual Analysis of Voice Quality The Place of Vocal Profile Analysis. In W. J. Hardcastle, & J. Mackenzie Beck (Eds.), *A Figure of Speech: A Festschrift for John Laver* (pp. 285-322). Mahwah, New Jersey & London: Lawrence Erlbaum Associates.

Mackenzie Beck, J. (2010). Organic Variation of the Vocal Apparatus. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), *The Handbook of Phonetic Sciences* (Second ed., pp. 153-201). Chichester: Blackwell Publishing Ltd. https://doi.org/10.1002/9781444317251.ch5

Mackenzie Beck, J., & Schaeffler, F. (2015). Voice quality variation in Scottish adolescents: gender versus geography. *Proceedings of the 18th International Conference of Phonetic Sciences (ICPhS), Glasgow, Scotland.*

Mackenzie, J., Laver, J., & M, H. S. (1983). Structural pathologies of the vocal folds and phonation. *Edinburgh University Department of Linguistics Work in Progress, 16*, pp. 80-116.

Maddieson, I. (1984). *Patterns of Sound.* Cambridge: Cambridge University Press.

Marković, M. (2009a). Different Strategies in Acquiring L2 Vowels: The Production of High English Vowels /iː, ɪ, uː, ʊ/ by Native Speakers of Serbian. In B. Čubrović, & T. Paunović (Eds.), *Ta(l)king English Phonetics Across Frontiers,* (pp. 3-18). Cambridge: Cambridge Scholars Publishing.

Marković, M. (2009b). Perception and production of English vowels /e/ and /æ/ by native speakers of Serbian. In A. Tsangalidis (Ed.), *Selected Papers from the 18th International Symposium of Theoretical and Applied Linguistics,* (pp. 253-262). Thessaloniki: Aristotle University of Thessaloniki. Retrieved September 1st, 2018, from http://ejournals.lib.auth.gr/thal/article/view/5444

Marković, M. (2011). Acquiring Second Language Prosody - Fundamental Frequency. *Proceedings of the 1st International Conference on English Studies: English Language and Anglophone Literatures Today (ELALT), Novi Sad, Serbia*, (pp. 238-249).

Marković, M. (2012). *Uporedna proučavanja vokala engleskog i srpskog jezika: između univerzalnog i specifičnog.* Novi Sad: Filoofski fakultet u Novom Sadu.

Marković, M., & Jakovljević, B. (2016). Phonetic cue interpretation in the acquisition of a non-native vocalic contrast. *Годишњак Филозофског факултета у Новом Саду, 41*(1), pp. 215-227.

Marquez, E. (2018). Changes In Fundamental Frequency Of Voice In A Group Of Transwomen Following Voice Modification Therapy. *[Master's thesis, University of Texas at El Paso] Open Access Theses & Dissertations. 1477*.

Marwick, H., Mackenzie, J., Laver, J., & Trevarthen, C. (1984). Voice quality as an expressive system in mother-to-infant communication: A case study. *Work in Progress, University of Edinburgh, Department of Linguistics, 17*, pp. 85-97.

McDougall, K. (2013a). Assessing perceived voice similarity using Multidimensional Scaling for the construction of voice parades. *International Journal of Speech, Language and the Law, 20*(2), 163-172. https://doi.org/10.1558/ijsll.v20i2.163

McDougall, K. (2013b). Earwitness evidence and the question of voice similarity. *British Academy Review, 21,* 18-21. Retrieved May 2024, from https://www.thebritishacademy.ac.uk/documents/805/BAR21-06-McDougall.pdf

McDougall, K. (2021). Ear-catching versus eye-catching? Some developments and current challenges in earwitness identification evidence. *Associazione Italiana Scienze della Voce, Proceedings of XVII AISV Conference, Zürich, Switzerland*, (pp. 33-56). https://doi.org/10.17469/O2108AISV000002

McDougall, K., Pautz, N., Goodwin, P., Nolan, F., MüllerJohnson, K., Paver, A., & Smith, H. M. (2023). An investigation of the effect of warning strength on voice parade performance. *31st International Association for Forensic Phoneticians and Acoustics Annual Conference, Zürich, Switzerland, 9-13 July 2023.*

McDougall, K., Paver, A., & Nolan, F. (2022). The impact of duration of speech sample on listeners' judgements of voice similarity. *Poster presented at the British Association of Academic Phoneticians Colloquium, York, 4-8 April 2022.*

McLaren, M., Castan, D., & Ferrer, L. (2016). Analyzing the Effect of Channel Mismatch on the SRI Language Recognition. *Proceedings of Oddyssey 2016, Bilbao, Spain*, (pp. 188-195).

McMenamin, G. R. (2002). *Forensic Linguistics: Advances in Forensic Stylistics.* Boca Raton, London, New York, Washington, D.C.: CRC Press.

McMicken, B., Salles, F., Shelley, V. B., Vento-Wilson, M., Rogers, K., Asterios, T., & Narayanan, S. S. (2017). Bilabial substitution patterns during consonant production in a case of congenital aglossia. *Journal of Communication Disorders, Deaf Studies and Hearing Aids, 5*(2), e1000175. https://doi.org/10.4172/2375-4427.1000175

Meuwly, D., Heeren, W., & Bolck, A. (2015). Exploring the strength of evidence of long-term formants in bilingual speakers. *Proceedings of Annual Conference of the International Association for Forensic Phonetics and Acoustics*, (pp. 75-76).

Mewly, D., Ramos, D., & Haraksim, R. (2017). A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation. *Forensic Science International, 276*, 142-153. https://doi.org/10.1016/j.forsciint.2016.03.048

Milne, P., Cavanagh, C., van der Vloed, D., & Dellwo, V. (2019). A survey of voice-related cases in three forensic speech laboratories. *A paper presented at the 28th Annual Conference of the Internatonal Assocaton for Forensc Phonetcs and Acoustcs.* Istanbul, 14th -17th July 2019.

Moisik, S. R. (2013). The Epilarynx in Speech. *Doctoral dissertation, Department of Linguistics, University of Victoria*.

Moisik, S. R., & Esling, J. H. (2014). Modeling the biomechanical influence of epilaryngeal stricture on the vocal folds: a low-dimensional model of vocal-ventricular fold coupling. *Journal of speech, language, and hearing research, 57*(2), S687–S704. https://doi.org/10.1044/2014_JSLHR-S-12-0279

Moisik, S. R., Esling, J. H., Crevier-Buchman, L., & Amelot, A. (2015). Multimodal imaging of glottal stop and creaky voice: Evaluating the role of epilaryngeal constriction. *Proceedings of the 18th International Congress on Phonetic Sciences (ICPhS), August 2015, Glasgow, United Kingdom*, (p. paper 247).

Moisik, S. R., Lin, H., & Esling, J. H. (2014). A study of laryngeal gestures in Mandarin citation tones using simultaneous laryngoscopy and laryngeal ultrasound (SLLUS). *Journal of the International Phonetic Association, 44*(1), 21-58. https://doi.org/10.1017/S0025100313000327

Mok, P. P., Bo Xu, R., & 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. (2010). Long-term formant distribution as a measure of speaker characteristics in read and spontaneous speech. *The Phonetician, 101*, 7-24. Retrieved January 2019, from http://www.isphs.org/Phonetician/Phonetician_101.pdf

Morrison, G. S. (2009). Forensic voice comparison and the paradigm shift. *Science and Justice, 49*, 298–308. https://doi.org/10.1016/j.scijus.2009.09.002

Morrison, G. S. (2009b). Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs. *he Journal of the Acoustical Society of America, 125*, 2387-2397. https://doi.org/10.1121/1.3081384

Morrison, G. S. (2011). Measuring the validity and reliability of forensic likelihood-ratio systems. *Science and Justice, 51*(3), 91-98. https://doi.org/10.1016/j.scijus.2011.03.002

Morrison, G. S. (2013). Tutorial on logistic regression calibration and fusion: converting a score to a likelihood ratio. *Australian Journal of Forensic Science, 25*(2), 173-197. https://doi.org/10.1080/00450618.2012.733025

Morrison, G. S. (2016). Special issue on measuring and reporting the precision of forensic likelihood ratios: Introduction to the debate. *Science and Justice, 56*(5), 371-373. https://doi.org/10.1016/j.scijus.2016.05.002

Morrison, G. S., & Enzinger, E. (2018). Score based procedures for the calculation of forensic likelihood ratios – Scores should take account of both similarity and typicality. *Science and Justice, 58*(1), 47-58. https://doi.org/10.1016/j.scijus.2017.06.005

Morrison, G. S., & Enzinger, E. (2019). Introduction to Forensic Voice Comparison. In W. F. Katz, & P. F. Assmann (Eds.), *The Routledge Handbook of Phonetics* (pp. 599–634). Abingdon, UK: Taylor & Francis. https://doi.org/10.4324/9780429056253-22

Morrison, G. S., Enzinger, E., Hughes, V., Jessen, M., Meuwly, D., Neumann, C., . . . 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-Diez, 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. 1-50). Boca Raton, FL: CRC.

Morrison, G. S., Lindh, J., & Curran, J. M. (2014). Likelihood ratio calculation for a disputed-utterance analysis with limited available data. *Speech Communication, 58*, 81-90. https://doi.org/10.1016/j.specom.2013.11.004

Morrison, G. S., Ochoa, F., & Thiruvaran, T. (2012). Database selection for forensic voice comparison. *Proceedings of the Odyssey 2012, Singapore*, (pp. 62-77).

Morrison, G. S., Weber, P., Enzinger, E., Labrador, B., Lozano-Díez, A., Ramos, D., & González-Rodríguez, J. (2022). Forensic voice comparison – Human-supervised-automatic approach. In M. Houck, L. Wilson, S. Lewis, H. Eldridge, K. Lothridge, & P. Reedy (Eds.), *Encyclopedia of Forensic Sciences* (Third ed., Vol. 2, pp. 720-736). Elsevier. https://doi.org/10.1016/B978-0-12-823677-2.00182-3

Muralikrishna, H., & Dinesh, D. A. (2022). Spoken language identification in unseen channel conditions using modified within-sample similarity loss. *Pattern Recognition Letters, 158*, pp. 16-23. https://doi.org/10.1016/j.patrec.2022.04.018

Murphy, C. H., & Doyle, P. C. (1987). The effects of cigarette smoking on voice-fundamental frequency. *Otolaryngology--head and neck surgery : official journal of American Academy of Otolaryngology-Head and Neck Surgery, 97*(4), 376–380. https://doi.org/10.1177/019459988709700406

Nagaraja, B. G., & Jayanna, H. S. (2013). Kannada Language Parameters for Speaker Identification with The Constraint of Limited Data. *The International Journal of Image, Graphics and Signal Processing, 9*, 14-20. https://doi.org/10.5815/ijigsp.2013.09.03

Nagaraja, B. G., & Jayanna, H. S. (2016). Feature extraction and modelling techniques for multilingual speaker recognition: a review. *International Journal of Signal and Imaging Systems Engineering, 9*(2), 67-78. https://doi.org/10.1504/IJSISE.2016.075000

Nair, B. B., Alzqhoul, E. A., & Guillemin, B. J. (2016). Impact of the GSM and CDMA Mobile Phone Networks on the Strength of Speech Evidence in Forensic Voice Comparison. *Journal of Forensic Research: Open Access, 7*(2), 324. http://dx.doi.org/10.4172/2157-7145.1000324

Nair, B., Alzqhoul, E., & Guillemin, B. J. (2014). Determination of Likelihood Ratios for Forensic Voice Comparison Using Principal Component Analysis. *The International Journal of Speech, Language and the Law, 21*(1), 83-112. https://www.doi.org/10.1558/ijsll.v21i1.83

Neuhauser, S. (2008). Voice Disguise Using a Foreign Accent: Phonetic and Linguistic Variation. *The International Journal of Speech Language and the Law, 15*(2), 131-159. https://doi.org/10.1558/ijsll.v15i2.131

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, 24*(6), e171–e176. https://doi.org/10.1016/j.jvoice.2011.07.013

Ni Chasiade, A., & Gobl, C. (2005). On the Relation Between Phonatory Quality and Affect. In W. J. Hardcastle, & J. Mackenzie Beck (Eds.), *A figure of speech: A Festschrift for John Laver* (pp. 323-346). Mahwah, New Jersey: Lawrence Erlbaum Associates Publishers.

Nikolić, D. (2016). Acoustic Analysis of English Vowels Produced by American Speakers and Highly Competent Serbian L2 Speakers. *Facta Universitatis. Series: Linguistics and Literature, 14*(1), 85-101.

Nolan, F. (1983). *The Phonetic Basis of Speaker Recognition.* Cambridge: Cambridge University Press.

Nolan, F. (1990). The Limitations of Auditory-Phonetic Speaker Identification. In H. Kniffka (Ed.), *Texte zu Theorie und Praxis forensischer Linguistik* (pp. 457-479). Tübingen: De Gruyter. 10.1515/9783111356464.457

Nolan, F. (1999). Speaker Recognition and Forensic Phonetics. In W. J. Hardcastle, & J. Laver (Eds.), *The Handbook of Phonetic Sciences* (pp. 1-15). Blackwell Reference Online: Blackwell Publishing.

Nolan, F. (2001). Speaker identification evidence: its forms, limitations, and roles. *Proceedings of the conference "Law and Language: Prospect and Retrospect"*. Levi, Finland.

Nolan, F. (2002). The 'telephone effect' on formants: a response. *Forensic Linguistics, 9*(1), 74-82. https://doi.org/10.1558/ijsll.v9i1.74

Nolan, F. (2005). Forensic speaker identification and the phonetic description of voice quality. In W. J. Hardcastle, & J. M. Back (Eds.), *A Figure of Speech: a Festschrift for John Laver* (pp. 385-411). Mahwah, New Jersey: Erlbaum.

Nolan, F. (2007). Voice Quality and Forensic Speaker Identification. *Govor, 24*(2), 111-128. Retrieved January 2019, from https://hrcak.srce.hr/173611

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. 10.1558/sll.2005.12.2.143

Nolan, F., McDougall, K., & Hudson, T. (2008). Voice Similarity and the Effect of the Telephone: A Study of the Implications for Earwitness Evidence (VoiceSim). *Final report RES-000-22-2582m*. Swindon: ESRC.

Nolan, F., McDougall, K., & Hudson, T. (2011). Some acoustic correlates of perceived (dis)similarity between same-accent voices. *Proceedings of the 17th International Congress of Phonetic Sciences, Hong Kong*, (pp. 1506-1509).

https://www.internationalphoneticassociation.org/icphsproceedings/ICPhS2011/OnlineProceedings/RegularSession/Nolan/Nolan.pdf.

Nolan, F., McDougall, K., & Hudson, T. (2013). Effects of the telephone on perceived voice similarity: implications for voice line-ups. *The International Journal of Speech Language and the Law, 20*(2), 229-246. https://www.doi.org/10.1558/ijsll.v20i2.229

Nygren, U., Nordenskjöld , A., Arver, S., & Södersten, M. (2016). Effects on Voice Fundamental Frequency and Satisfaction with Voice in Trans Men during Testosterone Treatment-A Longitudinal Study. *Journal of Voice, 30*(6), 766.e23–766.e34. https://doi.org/10.1016/j.jvoice.2015.10.016

O'Brien, B., Meunier, C., Ghio, A., Fredouille, C., & Bonastre, J.-F. G. (2021). Discriminating speakers using perceptual clustering interface. *Speaker Individuality in Phonetics and Speech Sciences: Speech Technology and Forensic Applications* (pp. 97-111). Zurich: Officinaventuno. https://doi.org/10.17469/O2108AISV000005

Olsson, J. (2008). *Forensic Linguistics* (2nd ed.). London and New York: Continuum International Publishing Group.

Orena, A. J., Polka, L., & Theodore, R. M. (2019). Identifying bilingual talkers after a language switch: Language experience matters. *Journal of the Acoustical Society of America, 145*(4), EL303–EL309. https://doi.org/10.1121/1.5097735

Orena, A. J., Theodore, R. M., & Polka, L. (2015). Language exposure facilitates talker learning prior to language comprehension, even in adults. *Cognition, 143*, 36-40. 10.1016/j.cognition.2015.06.002

Pabst, F., & Sundberg, J. (1993). Tracking multi-channel electroglottograph measurement of larynx height in singers. *Scandinavian Journal of Logopedics and Phoniatrics, 18*(4), 143-152. https://doi.org/10.3109/14015439309101360

Papcun, G., Kreiman, J., & Davis, A. (1989). Long-term memory for unfamiliar. *Journal of the Acoustical Society of America, 85*(2), 913-925. https://doi.org/10.1121/1.397564

Passetti, R. R., & Constantini, A. C. (2019). The Effect of Telephone Transmission on Voice Quality Perception. *Journal of Voice, 33*(5), 649-658. https://doi.org/10.1016/j.jvoice.2018.04.018

Paunović, T. (2009). Sociolingvistički pogled u susedovo dvorište: stavovi prema jezičkim varijetetima. *Radovi Filozofskog fakulteta, 11*(1), 77-99.

Paunović, T. (2011). Sounds Serbian? Acoustic properties of Serbian EFL students' speech. In E. Kitis, N. Lavidas, N. Topintzi, & T. Tsangalidis (Ed.), *Selected Papers from the 19th International Symposium on Theoretical and Applied Linguistics (ISTAL19)* (pp. 357-

369). Thesaloniki: Aristotle University of Thesaloniky, School of English, Department of Theoretical & Applied Linguistics.

Paunović, T. (2013). Beginnings, endings and in-betweens: Prosodic signals of discourse topic in English and Serbian. In B. Čubrović, & T. Paunović (Eds.), *Focus on English Phonetics* (pp. 191-213). Newcastle: Cambridge Scholars.

Paunović, T. (2015). Pitch height and pitch range in Serbian EFL students' reading and speaking tasks. *Nasleđe, 32*, 73-95.

Paunović, T. (2019). Focus on focus: Prosodic signals of utterance-level information structure in L1 Serbian, L1 English, and Serbian L2 English. *Zbornik Matice srpske za filologiju i lngvistiku, LXII*(2), 213-238.

Paver, A., Braber, N., & Wright, D. (2022). Listener judgements for social traits and criminal behaviours as a function of speaker pitch and articulation rate. *The 29th Conference of the International Assocication of Forensic Phonetics and Acoustics, Marburg, 10th - 13th July, 2022.*

Perrachione, T. K. (2019). Recognizing speakers across languages. In S. Frühholz, & P. Belin (Eds.), *The Oxford Handbook of Voice Perception* (pp. 1-17). Oxford: Oxford University Press. 10.1093/oxfordhb/9780198743187.013.23

Perrot, P., Aversano, G., & Chollet, G. (2007). Voice disguise and automatic detection: review and perspectives. In Y. Stylianou, M. Faundez-Zanuy, & A. Esposito (Eds.), *Nonlinear Speech Processing. Lecture Notes in Computer Science* (Vol. 4391, pp. 101-117). https://doi.org/10.1007/978-3-540-71505-4_7

Pessoa, A. N., Novaes, B. C., Madureira, S., & Camargo, Z. A. (2012). Perceptual and acoustic correlates of speech in a bilateral cochlear implant user. *Proceedings of the Speech Prosody 2012, May 22-25, Shanghai, China*, (pp. 51-54).

Pessoa-Almeida, A. N., Meireles, A., Madureira, S., & Camargo, Z. (2014). Prosodic analysis of the speech of a child with cochlear implant. In N. Campbell, D. Gibbon, & D. Hirst (Ed.), *Speech Prosody, 7*, pp. 1115-1118.

Philbrick, F. A. (1949). *Language and the Law: The Semantics of Forensic English.* New York: Macmillan.

Pillot-Loiseau, C., Horgues, C., Scheuer, S., & Kamiyama, T. (2019). The evolution of creaky voice use in read speech by native-French and native-English speakers in tandem: a pilot study. *Anglophonia [Online]*, 27. https://doi.org/10.4000/anglophonia.2005

Pinar, D., Cincik, H., Erkul, E., & Gungor, A. (2016). Investigating the Effects of Smoking on Young Adult Male Voice by Using Multidimensional Methods. *Journal of Voice, 30*(6), 721-725. https://doi.org/10.1016/j.jvoice.2015.07.007

Pisoni, D. B., & Martin, C. S. (1989). Effects of Alcohol on the Acoustic-Phonetic Properties of Speech: Perceptual and Acoustic Analyses. *Alcoholism: Clinical and Experimental Research, 13*(4), 577–587. https://doi.org/10.1111/j.1530-0277.1989.tb00381.x

Podesva, R. J. (2013). Gender and the social meaning of non-modal phonation types. In C. Cathcart, I.-H. Chen, G. Finley, S. Kang, C. S. Sandy, & E. Stickles (Ed.), *Proceedings of the 37th Annual Meeting of the Berkeley Linguistics Society*, (pp. 427-448).

Pommée, T., & Morsomme, D. (in press). Voice Quality in Telephone Interviews: A preliminary Acoustic Investigation. *Journal of Voice*. https://doi.org/10.1016/j.jvoice.2022.08.027

Protopapas, A., & Lieberman, P. (1997). Fundamental frequency of phonation and perceived emotional stress. *The Journal of the Acoustical Society of America, 101*(4), 2267–2277. https://doi.org/10.1121/1.418247

Prozansky, S., & Mathews, M. V. (1964). Talker-recognition procedure based on analysis of variance. *The Journal of the Acoustical Society of America, 36*(11), 2041-2047. https://doi.org/10.1121/1.1919320

Rajković, M., Jovičić, S., Đorđević, M., & Kašić, Z. (2005). Prozodijske karakteristike emotivnog govora: analiza dinamičkog ponašanja grupe akustičkih obeležja. *Zbornik radova 49. Konferencije za ETRAN, Budva, 5-10. juna 2005, 2*, pp. 373-376.

Ramanarayanana, V., Goldstein, L., Byrd, D., & Narayanan, S. S. (2013). An investigation of articulatory setting using real-time magnetic resonance imaging. *Journal of the Acoustical Society of America, 134*(1), 510-519. https://doi.org/10.1121/1.4807639

Raming, L. A. (1983). Effects of physiological aging on speaking and reading rates. *Journal of Communication Disorders, 16*(3), 217-226. https://doi.org/10.1016/0021-9924(83)90035-7

Rappaport, D. L. (2000). Establishing a Standard for Digital Audio Authenticity: A Critical Analysis of Tools, Methodologies and Challenges. *MA thesis. University of Colorado*.

Redford, M., & Baese-Berk, M. (2023). Acoustic Theories of Speech Perception. *Oxford Research Encyclopedia of Linguistics*. https://doi.org/10.1093/acrefore/9780199384655.013.742

Reich, A. R., & Duke, J. E. (1979). Effects of selected vocal disguises upon speaker identification by listening. *The Journal of the Acoustical Society of America, 66*(4), 1023-8. https://doi.org/10.1121/1.383321

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. (2017). Aging effects on voice features used in forensic speaker comparison. *The International Journal of Speech, Language and the Law, 24*(2), 177-199. https://doi.org/10.1558/ijsll.34096

Roach, P. (1991). *English Phonetics and Phonology: A Practical Course* (2nd ed.). Cambridge: Cambridge University Press.

Robertson, B., & Vignaux, G. (1995). *Interpreting Evidence: Evaluating Forensic Science in the Courtroom.* Chichester: John Wiley and Sons.

Rose, P. (2002). *Forensic Speaker Identification.* London and New York: Taylor & Francis.

Rose, P. (2003). The Technical Comparison of Forensic Voice Samples. *Expert Evidence 99*. (H. Selby, & I. Freckelton, Eds.) Sydney: Thompson Lawbook Co.

Rose, P. (2004). Technical Forensic Speaker Identification from a Bayesian Linguist's Perspective. *Proceeding of Odyssey 2004, Toledo, Spain.*

Rose, P. (2006). Accounting for Correlation in Linguistic-Acoustic Likelihood Ratio-based Forensic Speaker Discrimination. *Speaker and Language Recognition Workshop, 2006. IEEE Odyssey, 28th - 30th June.*, (pp. 1-8). https://doi.org/10.1109/ODYSSEY.2006.248095

Rose, P. (2013). More is better: likelihood ratio-based forensic voice comparison with vocalic segmental cepstra frontends. *The International Journal of Speech Language and the Law, 30*(1), 77-116. https://doi.org/10.1558/ijsll.v20i1.77

Rose, P. (2013b). Where the science ends and the law begins: Likelihood ratio-based forensic voice comparison in a $150 million telephone fraud. *The International Journal of Speech, Language and the Law, 20*(2), 277-324. https://doi.org/10.1558/ijsll.v20i2.277

Rose, P. (2015). Forensic voice comparison with monophthongal formant trajectories - a likelihood ratio-based discrimination of "schwa" vowel acoustics in a close social group of young Australian females. *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* https://doi.org/10.1109/ICASSP.2015.7178886

Rose, P. (2017). Likelihood ratio-based forensic voice comparison with higher level features: research and reality. *Computer Speech and Language, 45*, 475-502. https://doi.org/10.1016/j.csl.2017.03.003

Rose, P. (2022). Likelihood ratio-based forensic semi-automatic speaker identification with alveolar fricative spectra in a real-world case. *Proceedings of the Eighteenth Australasian International Conference on Speech Science and Technology*, (pp. 6-10). Retrieved January 2023, from https://sst2022.files.wordpress.com/2022/12/rose-2022-likelihood-ratio-based-forensic-semi-automatic-speaker-identification-with-alveolar-fricative-spectra-in-real-world-case.pdf

Rose, P., & Morrison, G. S. (2009). A response to the UK Position Statement on forensic speaker comparison. *The International Journal of Speech, Language and the Law, 16*(1), 139-163. https://doi.org/10.1558/ijsll.v16i1.139

Rose, P., & Wang, X. (2016). Cantonese forensic voice comparison with higher-level features: likelihood ratio-based validation using F-pattern and tonal F0 trajectories over a disyllabic hexaphone. *Proceedings of Odyssey 2016*, (pp. 326-333). https://doi.org/10.21437/Odyssey.2016-47

Rose, Y., McAllister, T., & Inkelas, S. (2022). Developmental Phonetics of Speech Production. In R.-A. Knight, & J. Setter (Eds.), *The Cambridge Handbook of Phonetics* (pp. 578-602). Cambridge: Cambridge University Press.

Rothenberg, M. (1992). A Multichannel Electroglottograph. *Journal of Voice, 6*(1), 36-43. https://doi.org/10.1016/S0892-1997(05)80007-4

Rubin, D. B., & Schenker, N. (1986). Multiple Imputation for Interval Estimation From Simple Random Samples With Ignorable Nonresponse. *Journal of the American Statistical Association, 81*(394), 366-374. https://doi.org/10.2307/2289225

Russell, A., Penny, L., & Pemberton, C. (1995). Speaking Fundamental Frequency Changes Over Time in Women: A Longitudinal Study. *Journal of Speech and Hearing Research, 38*(1), 101-109. https://doi.org/10.1044/jshr.3801.101

Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: People with extraordinary face recognition ability. *Psychonomic Bulletin & Review, 16*, 252–257. https://doi.org/10.3758/PBR.16.2.252

Růžičková, A., & Skarnitzl, R. (2017). Voice disguise strategies in Czech male speakers. *AUC Philologica, 2017*(3), 19-34. https://doi.org/10.14712/24646830.2017.30

Saleem, S., Subhan, F., Naseer, N., Bais, A., & Imtiaz, A. (2020). Forensic speaker recognition: A new method based on extracting accent and language information from short

utterances. *Forensic Science International: Digital Investigation, 34*, 300982. https://doi.org/10.1016/j.fsidi.2020.300982

San Segundo, E. (2021). International survey on voice quality: Forensic practitioners versus voice therapists. *Estudios de Fonética Experimental, 29*, 8-34.

San Segundo, E., & Delgado Hernández, J. (2021). A preliminary approach to the acoustic-perceptual characterization of dysarthria. *Proceedings of the 3rd International Symposium on Applied Phonetics (ISAPh 2021), 6-8 September 2021, Tarragona, Spain*, (pp. 63-66). https://doi.org/10.21437/ISAPh.2021-11

San Segundo, E., & Mompeán, J. A. (2017). A Simplified Vocal Profile Analysis Protocol for the Assessment of Voice Quality and Speaker Similarity. *Journal of Voice, 31*(5), 644.e11–644.e27. 10.1016/j.jvoice.2017.01.005

San Segundo, E., Braun, A., Hughes, V., & Foulkes, P. (2017). Speaker-similarity perception of Spanish twins and non-twins by native speakers of Spanish, German and English. *A paper presented at the International Association of Forensic Phonetics and Acoustics (IAFPA) conference, Split, Croatia.*

San Segundo, E., Foulkes, P., & Hughes, V. (2016). Holistic perception of voice quality matters more than L1 when judging speaker similarity in short stimuli. *Proceedings of the 16th Australasian Conference on Speech Science and Technology (ASSTA). University of Western Sydney, Australia.*

San Segundo, E., Foulkes, P., French, P., Harrison, P., Hughes, V., & Kavanagh, C. (2018). Cluster analysis of voice quality ratings: Identifying groups of perceptually similar speakers. *Proceedings of the Conference on Phonetics & Phonology in German-speaking Countries (P&P 13)*, (pp. 173-176).

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 proposal. *Journal of the International Phonetic Association, 49*(3), 353-380. 10.1017/S0025100318000130

Sanz, C., Pallavicini, C., Carrillo, F., Zamberlan, F., Sigman, M., Mota, N., . . . Tagliazucchi, E. (2021). The entropic tongue: Disorganization of natural language under LSD. *Consciousness and Cognition, 87*, 103070. https://doi.org/10.1016/j.concog.2020.103070

Schiel, F., & Heinrich, C. (2015). Disfluencies in the speech of intoxicated speakers. *The International Journal of Speech, Language and the Law, 22*(1), 19-34. https://doi.org/10.1558/ijsll.v22i1.24767

Schiller, N. O., & Köster, O. (1998). The ability of expert witnesses to identify voices: a comparison between trained and untrained listerners. *Forensic Linguistics, 5*(1), 1-9. https://doi.org/10.1558/ijsll.v5i1.1

Schlamann, M., Lehnerdt, G., Maderwald, S., & Ladd, S. (2009). Dynamic MRI of the vocal cords using phased-array coils: A feasibility study. *The Indian Journal of Radiology and Imaging, 19*(2), 127-131. https://doi.org/10.4103/0971-3026.50830

Schwab, S., & Goldman, J.-P. (2016). Do speakers show different F0 when they speak in different languages? The case of English, French and German? *Speech Prosody, Boston, 31 May 2016 - 3 June 2016, ISCA.*

Schwartz, G. (2019). Voice quality and L2 proficiency in the English tense-lax contrast. *Anglophonia [Online], 27*.

Scrucca, L., Fraley, C., Murphy, B. T., & Raftery, A. E. (2023). *Model-Based Clustering, Classification, and Density Estimation Using mclust in R.* New York: Chapman and Hall/CRC. https://doi.org/10.1201/9781003277965

Selamtzis, A. (2018). Analyses of voice and glottographic signals in singing and speech. *Doctoral thesis in speech and music communication, Stockholm, Sweden*.

Shewell, C. (1998). The effect of perceptual training on ability to use the Vocal Profile Analysis Scheme. *International journal of language and communication disorders, 33*(Supplement), 322-326. https://doi.org/10.3109/13682829809179444

Shue, Y.-L., Keating, P., Vicenik, C., & Yu, K. (2011). VoiceSauce: A program for voice analysis. *Proceedings of the ICPhS XVII, Hong Kong*, (pp. 1846-1849).

Simić, R., & Ostojić, B. (1996). *Osnovi fonologije srpskog književnog jezika.* Beograd: Univerzitet u Beogrdu.

Simpson, A. P., & Ericsdotter, C. (2007). Sex-specific differences in f0 and vowel space. *Proceedings of the XVIth International Congress of Phonetic Sciences*, (pp. 933–936). Saarbrücken.

Sjölander, K. (2004). The snack sound toolkit [computer program].

Sjöström, M., Eriksson, E. J., Zetterholm, E., & Sullivan, K. P. (2006). A Switch of Dialect as Disguise. *Lund University, Centre for Languages & Literature, Dept. of Lmguistics & Phonetics. Working Papers, 52*, pp. 113-116.

Skvortsov, A. (2021). Call Recorder skvalex v. 3.4.9. *Software*. Retrieved September 2021, from https://callrecorder.skvalex.com/get

Smith, A. (2010). Development of Neural Control of Orofacial Movements for Speech. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), *The Handbook of Phonetic Sciences* (Second ed., pp. 251-296). Chichester: Wiley-Blackwell.

Smith, H. M., Baguley, T. S., Robson, J., Dunn, A. K., & Stacey, P. C. (2018). Forensic voice discrimination by lay listeners: The effect of speech type and background noise on performance. *Applied Cognitive Psychology, 33*(2), 272-287. https://doi.org/10.1002/acp.3478

Smrkovski, L. L. (1975). Collaborative Study of Speaker Identification by the Voiceprint Method. *Journal of Association of Official Analytical Chemists, 58*(3), 453-456. https://doi.org/10.1093/jaoac/58.3.453

Soares, C., & Grosjean, F. (1984). Bilinguals in a monolingual and a bilingual speech mode: The effect on lexical access. *Memory and Cognition*, 380-386. https://doi.org/10.3758/BF03198298

Sobell, L. C., & Sobell, M. B. (1972). Effects of Alcohol on the Speech of Alcoholics. *Journal of Speech and Hearing Research, 15*(4), 861-868. https://doi.org/10.1044/jshr.1504.861

Sóskuthy, M., & Stewart-Smith, J. (2020). Voice quality and coda /r/ in Glasgow English in the early 20th century. *Language Variation and Change, 32*(2), 133-157. https://doi.org/10.1017/S0954394520000071

Soskuthy, M., Foulkes, P., Haddican, W., Hay, J., & Hughes, V. (2015). Word-level distributions and structural factors code-termine GOOSE fronting. *Proceedings of the 18th International Congress of Phonetic Sciences, Glasgow, UK*. Retrieved May 2024, from https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS1001.pdf

Speech Technology Center. (n.d.). SIS II forensic sound software;. Retrieved from http://www.speechpro.com/

Sredojević, D. (2017). *Fonetsko-fonološki opis akcenata u standardnom srpskom jeziku: od specifičnog ka opštem.* Novi Sad: Sajnos.

Sretenović, M. (2015). Zašto se rugamo govoru „južne pruge". *Politika*. Retrieved November 2022, from https://www.politika.rs/sr/clanak/344086/Kultura/Zasto-se-rugamo-govoru-juzne-pruge

Stanojčić, Ž., & Popović, L. (1986). *Gramatika srpskoga jezika za gimnazije i srednje škole.* Beograd: Zavod za udžbenike.

Steeneken, H. J., & Hansen, J. H. (1999). Speech under stress conditions: overview of the effect on speech production and on system performance. *1999 IEEE International Conference*

on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258). *4*, pp. 2079-2082. IEEE. https://www.doi.org/10.1109/ICASSP.1999.758342

Stevenage, S. V., Clarke, G., & McNeill, A. (2012). The "other-accent" effect in voice recognition. *Journal of Cognitive Psychology, 24*(6), 647-653. 10.1080/20445911.2012.675321

Stevens, K. N. (1971). Sources of inter-and intra-speaker variability in the acoustic properties of speech sounds. *Proceedings of the seventh International Congress of Phonetic Sciences/Actes du Septième Congrès international des sciences phonétiques*, (pp. 206-232). Montreal.

Stone, M. (2010). Laboratory Techniques for Investigating Speech Articulation. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), *The Handbook of Phonetic Sciences* (Second ed., pp. 9-38). Chichester: Blackwell Publishing Ltd. https://doi.org/10.1002/9781444317251.ch1

Stoney, D. A. (1991). Transfer evidence. In C. G. Aitken, & D. A. Stoney (Eds.), *The Use of Statistics in Forensic Science* (pp. 107-138). Chichester: Ellis Horwood Ltd.

Stuart-Smith, J. (1999). Glasgow: Accent and voice quality. In G. J. Docherty, & P. Foulkes (Eds.), *Urban voices: Accent study in the British Isles* (pp. 203-222). London: Edward Arnold.

Suárez, J. A. (1983). *The Mesoamerican Indian Languages.* Cambridge: Cambridge University Press.

Subotić, L., Sredojević, D., & Bjelaković, I. (2012). *Fonetika i fonologija: ortoepska i ortografska norma standardnog srpskog jezika.* Novi Sad: Filozofski fakultet Novi Sad.

Sullivan, K. P., & Kügler, F. (2001). Was the knowledge of the second language or the age difference the determining factor? *Forensic Linguistics, 8*(2), 1-8. https://doi.org/10.1558/sll.2001.8.2.1

Sullivan, K. P., & Schlichting, F. (2000). Speaker discrimination in a foreign language: first language environment,second language learners. *Forensic Linguistics, 7*(1), 95-112. https://doi.org/10.1558/sll.2000.7.1.95

Sundström, E., & Oren, L. (2019). Sound production mechanisms of audible nasal emission during the sibilant /s/. *Journal of the Acoustical Society of America, 146*(6), 4199–4210. https://doi.org/10.1121/1.5135566

Svartvik, J. (1968). *Evans Statements: A Case for Forensic Linguistics.* Gothenburg: University of Gothenburg Press.

Sztahó, D., & Szaszák, G. (2021). Deep Learning Methods in Speaker Recognition: A Review. *Periodica Polytechnica Electrical Engineering and Computer Science, 65*(4), 310-328. https://doi.org/10.3311/PPee.17024

Tan, T. (2010). The effect of voice disguise on Automatic Speaker Recognition. *Proceedings of the 3rd International Congress on Image and Signal Processing (CISP). IEEE*, (pp. 3538–3541). Yantai, China. http://doi.org/10.1109/CISP.2010.5647131

Taroni, F., Aitken, C., Garbolino, P., & Biedermann, A. (2006). *Bayesian Networks and Probabilistic Inference in Forensic Science.* Hoboken, New Jersey: John Wiley & Sons Ltd.

Tisljár-Szabó, E., Rossu, R., Varga, V., & Pléh, C. (2014). The Effect of Alcohol on Speech Production. *Journal of Psycholinguistic Research, 43*(6), 737–748. https://doi.org/10.1007/s10936-013-9278-y

Titze, I. R. (1990). Interpretation of the electroglottographic signal. *Journal of Voice, 4*(1), 1-9. https://doi.org/10.1016/S0892-1997(05)80076-1

Tomić, K. (2017). Temporal Parameters of Spontaneous Speech in Forensic Speaker Identification in Case of Language Mismatch: Serbian as L1 and English as L2. *32*, 117-143. https://doi.org/10.14746/cl.2017.32.5

Tomić, K. (2020). Acoustic Analysis of Pitch Accent as a Regional Forensic Marker in Serbian. *Facta Universitatis Series: Linguistics and Literature, 18*(2), 235-256. https://doi.org/10.22190/FULL2002235T

Tomić, K. (2020). Međujezična forenzička komparacija glasova - dugoročne frekvencije formanata. *Primenjena lingvistika, 21*, 7-24. https://doi.org/10.18485/primling.2020.21.1

Tomić, K. (2020). The effect of linguistic context on speaker recognition by earwitnesses in voice line-ups. In V. Lopičić, & B. Mišić Ilić (Eds.), *Jezik, književnost, kontekst. Tematski zbornik radova.* (pp. 135-148). Niš: Izdavački centar Filozofski fakultet Niš.

Tomić, K., & French, P. (2019). Long-term Formant Frequencies in Cross-language Forensic Voice Comparison under Likelihood Ratio Framework. *A paper Presented at The 28th Annual Conference of the International Association for Forensic Phonetics and Acoustics in Istanbul, July 13th to 17th.*

Tomić, K., & French, P. (2023). Vocal Profile Analysis as a Tool in Cross-Language Forensic Speaker Comparison. *A paper presented at the 31st Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA), Zurich, July 9–12, 2023.*

Tomić, K., & Milenković, K. (2019). Forenzičko profilisanje govornika iz uzorka na engleskom kao stranom jeziku - analiza kvaliteta vokala. *Zbornik Matice srpske za filologiju i lingvistiku, 62*(1), 151-170.

Tosi, O. J. (1979). *Voice Identification: Theory and Legal Applications.* Baltimore: University Park Press.

Trudgill, P. (2000). *Sociolinguistics: An Introduction to Language and Society.* London: Penguin Books Limited.

Tuhanioğlu, B., Erkan, S. O., Özdaş, T., Derici, Ç., Tüzün, K., & Şenkal, Ö. A. (2019). The Effect of Electronic Cigarettes on Voice Quality. *Journal of Voice, 33*(5), 811.e13-811.e17. https://doi.org/10.1016/j.jvoice.2018.03.015

University of York. (2022). *MA Forensic Phonetics*. Retrieved October 28, 2021, from University of York: https://www.york.ac.uk/study/postgraduate-taught/courses/ma-forensic-phonetics/

Valenzuela, M. G., & French, P. (2023). Production of English Vowel Contrasts in Spanish L1 Learners: A Longitudinal Study. *Loquens, 10*(1-2), e102. https://doi.org/10.3989/loquens.2023.e102

van As-Brooks, C., Hilgers, F. J., Koopmans-van Beinum, F. J., & Pols, L. C. (2005). Anatomical and Functional Correlates of Voice Quality in Tracheoesophageal Speech. *Journal of Voice, 19*(3), 360-372. https://doi.org/10.1016/j.jvoice.2004.07.011

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software, 45*(3), 1-67. https://doi.org/10.18637/jss.v045.i03

van den Berg, J. (1958). Myoelastic-aerodynamic theory of voice production. *Journal of Speech and Hearing Research, 1*, 227-244. https://doi.org/10.1044/jshr.0103.227

van der Vloed, D., Jessen, M., & Gfroerer, S. (2017). Experiments with two forensic automatic speaker comparison systems using reference populations that (mis)match the test language. *A paper presented at the Conference on Audio Forensics 2017, June 15–17 Arlington, VA, USA.*

van Leeuwen, D. A., & Brümmer, N. (2007). An Introduction to Application-Independent Evaluation of Speaker Recognition Systems. In C. Muller (Ed.), *Speaker Classifications I* (pp. 330-353). Heidelberg: Springer-Verlag. https://doi.org/10.1007/978-3-540-74200-5_19

Vaňková, J., & Skarnitzl, R. (2014). Within- and Between-Speaker Variability of Parameters Expressing Short-Term Voice Quality. *Proceedings of the 7th international conference on Speech Prosody*, (pp. 1081-1085). https://doi.org/10.13140/2.1.2377.1520

Vladisavljević, S. (1981). *Poremećaji izgovora.* Beograd: Privredni pregled.

Vogel, A. P., Pearson-Dennett, V., Magee, M., Wilcox, R. A., Esterman, A., Thewlis, D., . . . Todd, G. (2021). Adults with a history of recreational cannabis use have altered speech production. *Drug and Alcohol Dependence, 227*, 108963. https://doi.org/10.1016/j.drugalcdep.2021.108963

Vuković, B., Ćalasan, S., & Vegar, A. (2022). Influence of smoking on voice quality. *Biomedicinska istraživanja, 13*(1), 20-26. https://www.doi.org/10.5937/BII2201020V

Wagner, I. (2019). Examples of Casework in Forensic Speaker Comparison. *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019*, (pp. 721-725).

Wang, H., & Zhang, C. (2015). Forensic Automatic Speaker Recognition Based on Likelihood Ratio Using Acoustic-phonetic Features Measured Automatically. *Journal of Forensic Science and Medicine, 1*(2), 119-123. https://doi.org/10.4103/2349-5014.169617

Watt, D., Harrison, P., Hughes, V., French, P., Llamas, C., Braun, A., & Robertson, D. (2020). Assessing the effects of accent-mismatched reference population databases on the performance of an automatic speaker recognition system. *The International Journal of Speech, Language and the Law, 27*(1), 1-34. https://doi.org/10.1558/ijsll.41466

Webb, A. L., Carding, P. N., Deary, I. J., MacKenzie, K., N, S., & Wilson, J. A. (2004). The reliability of three perceptual evaluation scales for dysphonia. *European Archives of Oto-Rhino-Laryngology and Head & Neck, 261*, 429–434. https://doi.org/10.1007/s00405-003-0707-7

Wells, J. C. (1982). *Accents of English I: An Introduction.* Cambridge: Cambridge University Press.

Wells, J. C. (1982). Accents of English: Volume 1. *Cambridge University Press.*

Wenndta, S. J. (2016). Human recognition of familiar voices. *The Journal of the Acoustical Society of America, 140*(2), 1172-1183. https://doi.org/10.1121/1.4958682

Westbury, J., Milenkovic, P., Weismer, G., & Kent, R. (1990). X-ray microbeam speech production database. *The Journal of the Acoustical Society of America, 88*(S1), S56. https://doi.org/10.1121/1.2029064

Wester, M. (2012). Talker discrimination across languages. *Speech Communication, 54*(6), 781-790. https://doi.org/10.1016/j.specom.2012.01.006

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., . . . Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software, 4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). dplyr: A Grammar of Data Manipulation. R package version 1.1.4. https://CRAN.R-project.org/package=dplyr

Wilhelm, S. (2019). Voice Quality in British English. Its Nature, Functions and Applications. *Anglophonia [Online], 27*.

Williams, C. E. (1972). Emotions and Speech: Some Acoustical Correlates. *The Journal of the Acoustical Society of America, 52*(4), 1238-1250. https://doi.org/10.1121/1.1913238

Wilson, K. D. (1987). *Voice problems of children.* Baltimore, MA: Williams & Wilkins.

Winters, S. J., Levi, S. V., & Pisoni, D. B. (2008). Identification and discrimination of bilingual talkers across languages. *Journal of the Acoustical Society of America, 123*(6), 4524–4538. 10.1121/1.2913046

Wirz, S. (1991). The voice of the deaf. In M. Fawcus (Ed.), *Voice Disorders and their Management* (Second ed., pp. 283-303). New York: Springer.

Wolf, J. J. (1972). Eficient acoustic parameters for speaker recognition. *The Journal of the Acoustical Society of America*, 2044-2055.

Wrench, A., & Beck, J. (2022). Physiological Foundations. In T. C. Phonetics, R.-A. Knight, & J. Setter (Eds.). Cambridge: Cambridge University Press.

Wright, R., Mansfiel, C., & Panfili, L. (2019). Voice quality types and uses in North American English. *Anglophonia [Online], 27*. https://doi.org/10.4000/anglophonia.1952

Xiao Wang, B., & Hughes, V. (2022). Reducing uncertainty at the score-to-LR stage in likelihood ratio-based forensic voice comparison using automatic speaker recognition systems. *Proceedings of Interspeech 2022*, (pp. 5243-5247). https://doi.org/10.21437/Interspeech.2022-518

Xiao Wang, B., Hughes, V., & Foulkes, P. (2019). The effect of speaker sampling in likelihood ratio based forensic voice comparison. *The International Journal of Speech, Language and the Law, 26*(1), 97-120. https://doi.org/10.1558/ijsll.38046

Xiao Wang, B., Hughes, V., & Foulkes, P. (in press). Effect of Score Sampling on System Stability in Likelihood Ratio based Forensic Voice Comparison. *Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS)*, (pp. 3065-3069). Retrieved January 2023, from https://eprints.whiterose.ac.uk/142645/

Xie, S., Yan, N., Yu, P., Ng, M. L., Wang, L., & Ji, Z. (2016). Deep Neural Networks for Voice Quality Assessment based on the GRBAS Scale. *Proceedings of Interspeech 2016, San Francisco, USA*, (pp. 2656-2660). https://doi.org/10.21437/Interspeech.2016-986

Xu, Y., Tu, M., & Zhang, L. (2022). The Application of Authentication of Audio Recordings Technology in the Public Security System. *2022 International Conference on Computation, Big-Data and Engineering (ICCBE)*, (pp. 290-295). https://www.doi.org/10.1109/ICCBE56101.2022.9888179.

Yanushevskaya, I., Gobl, C., & Ní Chasaide, A. (2013). Voice quality in affect cueing: does loudness matter? *Frontiers in Psychology, 4*, Article 335. https://doi.org/10.3389/fpsyg.2013.00335

Yanushevskaya, I., Gobl, C., & Ní Chasaide, A. (2018). Cross-language differences in how voice quality and f contours map to affect. *The Journal of the Acoustical Society of America, 144*(5), 2730–2750. https://doi.org/10.1121/1.5066448

Yarmey, D. A. (1995). Earwitness speaker identification. *Psychology, Public Policy, and Law, 1*(4), 792-816. http://dx.doi.org/10.1037/1076-8971.1.4.792

Yarmey, D. A., & Matthys , E. D. (1992). Voice identification of an abductor. *Applied Cognitive Psychology, 6*(5), 367-377. https://doi.org/10.1002/acp.2350060502

Yarmey, D. A., Yarmey, L. A., Yarmey, M., & Parliament, L. (2001). Commonsense Beliefs and the Identification of Familiar Voices. *Applied Cognitive Psychology, 15*(3), 283 - 299. 10.1002/acp.702

Yu, M. (2019). Re-evaluating the Other Accent Effect in Talker Recognition. *MA thesis, University of Toronto*.

Yuasa, I. P. (2010). Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile American women? *American Speech, 85*(3), 315-337. https://doi.org/10.1215/00031283-2010-018

Yuasa, I. P. (2010). Creaky Voice: A New Feminine Voice Quality for Young Urban-Oriented Upwardly Mobile American Women? *American Speech, 85*(3), 315–337. https://doi.org/10.1215/00031283-2010-018

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

Zarić, M. (2004). Tajna Matejevačkog slučaja. *Novosti*. Retrieved February 2022, from https://www.novosti.rs/dodatni_sadrzaj/clanci.119.html:276554-Tajna-Matejeva269kog-slu269aja

Zec, D., & Zsiga, E. (2009). Interaction of Tone and Stress in Standard Serbian: The Cornell Meeting 2008. In W. Browne, A. Cooper, A. Fisher, E. Kesici, N. Predolac, & D. Zec (Eds.). Ann Arbor: Michigan Slavic Publications.

Zeidenberg, P., Clark, W. C., Jaffe, J., Anderson, S. W., Chin, S., & Malitz, S. (1973). Effect of Oral Administration of Δ9 Tetrahydrocannabinol on Memory, Speech, and Perception of Thermal Stimulation: Results With Four Normal Human Volunteer Subjects. Preliminary Report. *Comprehensive Psychiatry, 14*(6), 549-556. https://doi.org/10.1016/0010-440X(73)90040-0

Zeljkovic, I., Haffner, P., Amento, B., & Wilpon, J. (2008). GMM/SVM N-best speaker identification under mismatch channel conditions. *2008 IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, USA*, (pp. 4129-4132). https://doi.org/10.1109/ICASSP.2008.4518563

Zhong, C. (2019). Cross-Linguistic Forensic Voice Comparison Based on Vowel Formants. *Journal of Foreign Languages, 42*(1), 61-71.

Zuo, D., & Mok, P. P. (2015). Formant dynamics of bilingual identical twins. *Journal of Phonetics, 52*, 1-12. https://doi.org/10.1016/j.wocn.2015.03.003

Appendices

Appendix 1 – Interview discussion topics

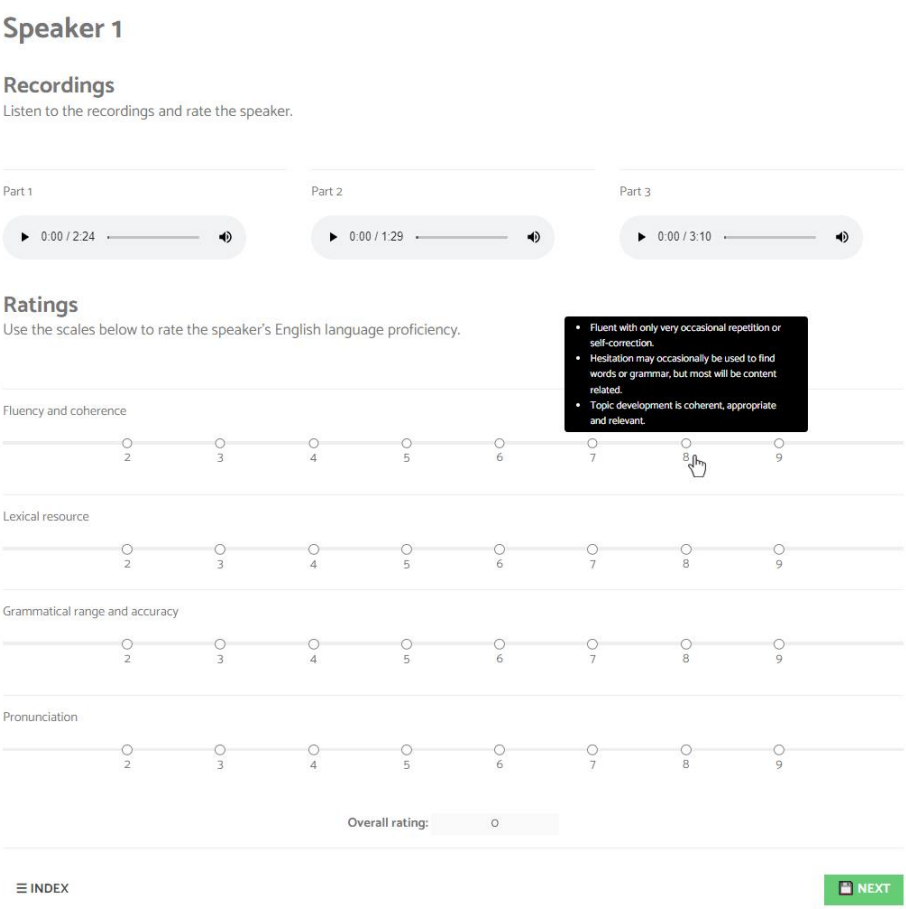
Primary questions

Serian	**Part 1**	Koliko dobro poznaješ svoje komšije? Koliko često ih viđaš? Zašto? Kakve probleme ljudi ponekad imaju sa komšijama? Šta misliš, kako komšije mogu da pomgonu jedni drugima? ... Da li imaš dovoljno slobodnog vremena? Da li provodiš svoje slobodno vreme u kući ili napolju? Koje aktivnosti voliš kada si napolju? Da li bi volela da isprobaš neku novu aktivnost u budućnosti, koju?
	Part 2	Opiši neki elektronski uređaj koji je po tebi veoma koristan? Koji je to uređaj? Koliko često ga koristiš i kako? Zašto misliš da je koristan?
	Part 3	Da li misliš da tehnologija koristi čovečanstvu i kako? Da li upotreba tehnologije može da ima negativne posledice na ljudsko društvo? Koje? Šta misliš, koliko će tehnologija da utiče na naš život u budućnosti? Da li tehnologija može da koristi u obrazovanju? Kako?
English	**Part 1**	What kind of place do you live in: a house or an apartment? Do you think it's better to live in a house or in an apartment? Why? Describe your neighbourhood. Do you like it and why? ... Do you like animals? Do you have any animals in your home as a pet? Did you have a pet when you were a child? Would you like to have a pet in the future?
	Part 2	Descibe a city you have visited which has impressed you. Where is it? When did you visit it? Why did you go there? Why did the city impress you?
	Part 3	What are the advantages and disadvantages of living in a big city? What are some of the major problems that big cities are facing nowadays? What are the possible solutions to these problems? How has citylife changed in the past 20 years?

Additional questions

Serbian	English
Speaking part 1	
Da li voliš da putuješ? Zašto? Kako obično putuješ? Gde si putovala u poslednje vreme? Kakva mesta voliš da posetiš? Koje države bi volela da posetiš i zašto?	Do you enjoy giving and receiving gifts? Why or why not? When do people usually give gifts? What gifts do people usually give? Have you ever given someone a gift that you made by yourself? Is it easy for you to choose a gift for someone?
Možeš li mi reći nešto o mestu gde živiš? Koje su prednosti i mane života u tom mestu? Da li voliš tu da živiš? Zašto? Čime se ljudi u tom mestu uglavnom bave? Po čemu bi rekla da je to mesto poznato?	How often do you watch television? Why? Which television channel do you usually watch? Why? Do you think most television programmes are good? Why?
Speaking part 2	
Opiši proslavu/žurku na kojoj si bila i koju ćeš zauvek pamtiti. Čija proslava je to bila? Šta se slavilo? Šta su ljudi radili na proslavi? Zašto ćeš zauvek pamtiit ovu proslavu?	Describe a teacher from your past that left an impact on you. What were this teacher's special qualities? Why do you remember this teacher?
Speaking part 3	
Zašto ljudi organizuju porodična slavlja u tvojoj zemlji? Da li odobravaš to što ljudi potroše puno novca na zabave i proslave? Zašto? Kakve nacionalne proslave postoje u tvojoj zemlji? Da li se slažete sa argumentom da bi novac koji se potroši na organizaciju ovih proslava (kao npr Nova godina) trebalo da se da u dobrotvorne svrhe? Zašto?	What kind of person makes a good teacher? Why do people choose to become teachers? Do you think education will change in the future? How? How does technology affect education?

Appendix 2 – English proficiency scoring experiment



Appendix 3 – IELTS speaking band descriptors

Band score	Fluency and Coherence
9	Fluent with only very occasional repetition or self-correction. Any hesitation that occurs is used only to prepare the content of the next utterance and not to find words or grammar. Speech is situationally appropriate and cohesive features are fully acceptable. Topic development is fully coherent and appropriately extended.
8	Fluent with only very occasional repetition or self-correction. Hesitation may occasionally be used to find words or grammar, but most will be content related. Topic development is coherent, appropriate and relevant.
7	Able to keep going and readily produce long turns without noticeable effort. Some hesitation, repetition and/or selfcorrection may occur, often mid-sentence and indicate problems with accessing appropriate language. However, these will not affect coherence. Flexible use of spoken discourse markers, connectives and cohesive features.
6	Able to keep going and demonstrates a willingness to produce long turns. Coherence may be lost at times as a result of hesitation, repetition and/or self-correction. Uses a range of spoken discourse markers, connectives and cohesive features though not always appropriately.
5	Usually able to keep going, but relies on repetition and self-correction to do so and/or on slow speech. Hesitations are often associated with mid-sentence searches for fairly basic lexis and grammar. Overuse of certain discourse markers, connectives and other cohesive features. More complex speech usually causes disfluency but simpler language may be produced fluently.
4	Unable to keep going without noticeable pauses. Speech may be slow with frequent repetition. Often self-corrects. Can link simple sentences but often with repetitious use of connectives. Some breakdowns in coherence.
3	Frequent, sometimes long, pauses occur while candidate searches for words. Limited ability to link simple sentences and go beyond simple responses to questions. Frequently unable to convey basic message.
2	Lengthy pauses before nearly every word. Isolated words may be recognisable but speech is of virtually no communicative significance.
Band score	**Lexical Resource**
9	Total flexibility and precise use in all contexts. Sustained use of accurate and idiomatic language.
8	Wide resource, readily and flexibly used to discuss all topics and convey precise meaning. Skilful use of less common and idiomatic items despite occasional inaccuracies in word choice and collocation. Effective use of paraphrase as required.
7	Resource flexibly used to discuss a variety of topics. Some ability to use less common and idiomatic items and an awareness of style and collocation is evident though inappropriacies occur. Effective use of paraphrase as required.
6	Resource sufficient to discuss topics at length. Vocabulary use may be inappropriate but meaning is clear. Generally able to paraphrase successfully.
5	Resource sufficient to discuss familiar and unfamiliar topics but there is limited flexibility. Attempts paraphrase but not always with success.

4	Resource sufficient for familiar topics but only basic meaning can be conveyed on unfamiliar topics. Frequent inappropriacies and errors in word choice. Rarely attempts paraphrase.
3	Resource limited to simple vocabulary used primarily to convey personal information. Vocabulary inadequate for unfamiliar topics.
2	Very limited resource. Utterances consist of isolated words or memorised utterances. Little communication possible without the support of mime or gesture.
Band score	**Grammatical Range and Accuracy**
9	Structures are precise and accurate at all times, apart from 'mistakes' characteristic of native speaker speech.
8	Wide range of structures, flexibly used. The majority of sentences are error free. Occasional inappropriacies and non-systematic errors occur. A few basic errors may persist.
7	A range of structures flexibly used. Error-free sentences are frequent. Both simple and complex sentences are used effectively despite some errors. A few basic errors persist.
6	Produces a mix of short and complex sentence forms and a variety of structures with limited flexibility. Though errors frequently occur in complex structures, these rarely impede communication.
5	Basic sentence forms are fairly well controlled for accuracy. Complex structures are attempted but these are limited in range, nearly always contain errors and may lead to the need for reformulation.
4	Can produce basic sentence forms and some short utterances are error-free. Subordinate clauses are rare and, overall, turns are short, structures are repetitive and errors are frequent.
3	Basic sentence forms are attempted but grammatical errors are numerous except in apparently memorised utterances.
2	No evidence of basic sentence forms.
Band score	**Pronunciation**
9	Uses a full range of phonological features to convey precise and/or subtle meaning. Flexible use of features of connected speech is sustained throughout. Can be effortlessly understood throughout. Accent has no effect on intelligibility.
8	Uses a wide range of phonological features to convey precise and/or subtle meaning. Can sustain appropriate rhythm. Flexible use of stress and intonation across long utterances, despite occasional lapses. Can be easily understood throughout. Accent has minimal effect on intelligibility.
7	Displays all the positive features of band 6, and some, but not all, of the positive features of band 8.
6	Uses a range of phonological features, but control is variable. Chunking is generally appropriate, but rhythm may be affected by a lack of stress-timing and/or a rapid speech rate. Some effective use of intonation and stress, but this is not sustained. Individual words or phonemes may be mispronounced but this causes only occasional lack of clarity. Can generally be understood throughout without much effort.
5	Displays all the positive features of band 4, and some, but not all, of the positive features of band 6.
4	Uses some acceptable phonological features, but the range is limited. Produces some acceptable chunking, but there are frequent lapses in overall rhythm.

	Attempts to use intonation and stress, but control is limited. Individual words or phonemes are frequently mispronounced, causing lack of clarity. Understanding requires some effort and there may be patches of speech that cannot be understood.
3	Displays some features of band 2, and some, but not all, of the positive features of band 4.
2	Uses few acceptable phonological features (possibly because sample is insufficient). Overall problems with delivery impair attempts at connected speech. Individual words and phonemes are mainly mispronounced and little meaning is conveyed. Often unintelligible.

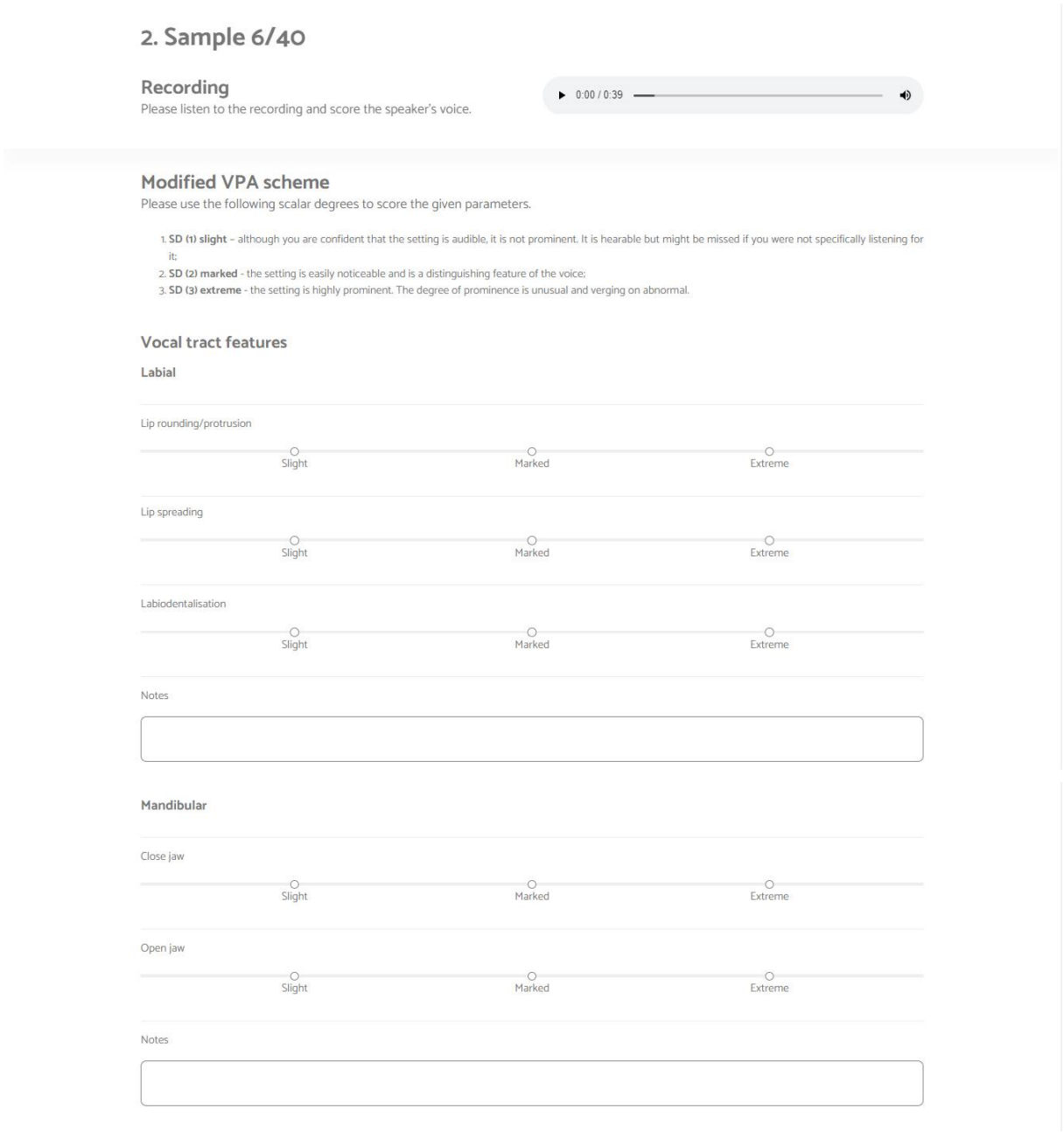
Note: The table is reproduced from IELTS Speaking Band Descriptors document at https://www.ielts.org/-/media/pdfs/ielts-speaking-band-descriptors.ashx (IELTS, 2023c)

Appendix 4 – Truncated Vocal Profile Analysis protocol

Vocal tract features		***Slight***	***Marked***	***Extreme***	***Notes***
Labial	Lip rounding/ protrusion				
	Lip spreading				
	Labiodentalisation				
Mandibular	Close jaw				
	Open jaw				
Tongue tip/blade	Advanced tongue tip/blade				
	Retracted tongue tip/blade				
	Retroflexion				
Lingual body	Raised tongue body				
	Lowered tongue body				
	Fronted tongue body				
	Backed tongue body				
	Extensive lingual range				
	Minimised lingual range				
Pharyngeal	Pharyngeal constriction				
Velopharyngeal	Nasal				
	Denasal				
Larynx height	Raised larynx				
	Lowered larynx / pharyngeal expansion				
Phonation features		***Absent***	***Present***		***Notes***
Falsetto					
Creak					
Whisper					
		Slight	***Marked***	***Extreme***	***Notes***
Creaky					
Whispery					
Breathy					
Harsh					
Tremor					
General notes					

Note: The protocol is a modified, truncated version of the original Vocal Profile Analysis Scheme by Laver et al. (1981).

Appendix 5 – VPA scoring experiment



2. Sample 6/40

Recording

Please listen to the recording and score the speaker's voice.

0:00 / 0:39

Modified VPA scheme

Please use the following scalar degrees to score the given parameters.

1. **SD (1) slight** – although you are confident that the setting is audible, it is not prominent. It is hearable but might be missed if you were not specifically listening for it;
2. **SD (2) marked** - the setting is easily noticeable and is a distinguishing feature of the voice;
3. **SD (3) extreme** - the setting is highly prominent. The degree of prominence is unusual and verging on abnormal.

Vocal tract features

Labial

Lip rounding/protrusion

Slight Marked Extreme

Lip spreading

Slight Marked Extreme

Labiodentalisation

Slight Marked Extreme

Notes

Mandibular

Close jaw

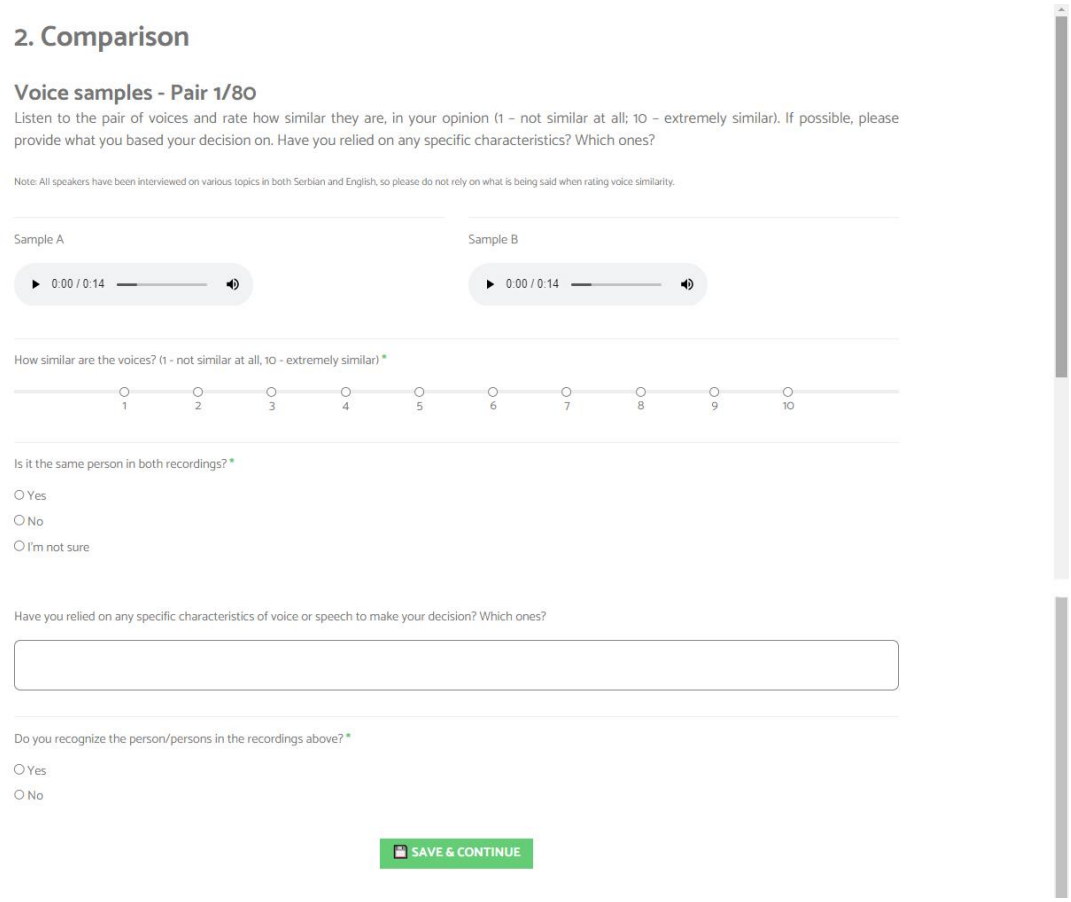
Slight Marked Extreme

Open jaw

Slight Marked Extreme

Notes

Appendix 6 – Voice similarity: naïve listeners



2. Comparison

Voice samples - Pair 1/80

Listen to the pair of voices and rate how similar they are, in your opinion (1 – not similar at all; 10 – extremely similar). If possible, please provide what you based your decision on. Have you relied on any specific characteristics? Which ones?

Note: All speakers have been interviewed on various topics in both Serbian and English, so please do not rely on what is being said when rating voice similarity.

Sample A

0:00 / 0:14

Sample B

0:00 / 0:14

How similar are the voices? (1 - not similar at all, 10 - extremely similar) *

1 2 3 4 5 6 7 8 9 10

Is it the same person in both recordings? *

- Yes
- No
- I'm not sure

Have you relied on any specific characteristics of voice or speech to make your decision? Which ones?

Do you recognize the person/persons in the recordings above? *

- Yes
- No

SAVE & CONTINUE

Appendix 7 – Paired t-test comparisons of Serbian and English formant values

Speaker	f1_t	f1_pvalue	f1_d
S1	-3.9021	0.0001	-0.0522
S10	-3.3960	0.0007	-0.0454
S11	12.3360	0.0000	0.1649
S12	9.4687	0.0000	0.1271
S13	25.1111	0.0000	0.3382
S14	-11.3864	0.0000	-0.1522
S15	5.0722	0.0000	0.0679
S16	-3.5461	0.0004	-0.0474
S17	12.5208	0.0000	0.1673
S19	27.4455	0.0000	0.3670
S20	3.2521	0.0011	0.0435
S21	-15.0436	0.0000	-0.2017
S23	14.3951	0.0000	0.1925
S24	3.1318	0.0017	0.0419
S25	-1.8993	0.0575	-0.0256
S26	-0.5891	0.5558	-0.0079
S27	-6.6732	0.0000	-0.0919
S28	-20.4301	0.0000	-0.2731
S29	12.7219	0.0000	0.1704
S30	6.1516	0.0000	0.0823
S32	-5.9806	0.0000	-0.0799
S33	9.2601	0.0000	0.1238
S35	-11.2464	0.0000	-0.1517
S36	-19.0331	0.0000	-0.2561
S37	24.4112	0.0000	0.3351
S38	9.4880	0.0000	0.1272
S39	26.5935	0.0000	0.3782
S40	0.9866	0.3238	0.0132
S41	-5.4480	0.0000	-0.0728
S42	-14.0240	0.0000	-0.1874
S43	-1.4924	0.1356	-0.0199
S44	4.3197	0.0000	0.0578
S45	-5.0487	0.0000	-0.0676
S46	5.2919	0.0000	0.0711
S47	14.4956	0.0000	0.1944
S48	-6.0225	0.0000	-0.0806
S49	6.8730	0.0000	0.0924
S51	8.8190	0.0000	0.1266
S52	-12.3739	0.0000	-0.1654
S53	-0.0369	0.9706	-0.0005
S54	-2.7710	0.0056	-0.0371
S55	-11.5698	0.0000	-0.1547
S56	-9.3312	0.0000	-0.1270
S57	-5.9935	0.0000	-0.0801
S58	-1.5861	0.1127	-0.0212
S59	-2.2698	0.0232	-0.0303
S6	3.5079	0.0005	0.0470
S60	-1.0117	0.3117	-0.0135
S7	15.1546	0.0000	0.2036
S8	-6.6362	0.0000	-0.0901

Speaker	f2_t	f2_pvalue	f2_d
S1	25.2311	0.0000	0.3382
S10	31.1743	0.0000	0.4221
S11	6.0298	0.0000	0.0812
S12	10.9724	0.0000	0.1468
S13	19.2711	0.0000	0.2576
S14	24.2209	0.0000	0.3269
S15	22.5419	0.0000	0.3018
S16	23.9392	0.0000	0.3199
S17	11.5985	0.0000	0.1553
S19	8.0168	0.0000	0.1073
S20	11.0294	0.0000	0.1480
S21	30.3613	0.0000	0.4132
S23	34.2240	0.0000	0.4574
S24	19.1986	0.0000	0.2566
S25	32.0801	0.0000	0.4319
S26	34.1136	0.0000	0.4566
S27	17.0020	0.0000	0.2272
S28	20.0733	0.0000	0.2726
S29	25.2680	0.0000	0.3392
S30	20.2111	0.0000	0.2709
S32	38.4022	0.0000	0.5133
S33	23.0800	0.0000	0.3087
S35	12.8047	0.0000	0.1718
S36	37.6747	0.0000	0.5175
S37	28.5411	0.0000	0.3814
S38	16.7657	0.0000	0.2246
S39	18.7699	0.0000	0.2509
S40	17.2214	0.0000	0.2301
S41	22.6907	0.0000	0.3042
S42	35.2510	0.0000	0.4717
S43	30.2489	0.0000	0.4042
S44	5.9566	0.0000	0.0797
S45	21.4465	0.0000	0.2868
S46	8.6452	0.0000	0.1157
S47	18.8333	0.0000	0.2519
S48	16.8632	0.0000	0.2254
S49	23.7961	0.0000	0.3180
S51	24.3745	0.0000	0.3266
S52	25.2984	0.0000	0.3390
S53	27.8181	0.0000	0.3719
S54	30.9673	0.0000	0.4142
S55	30.1180	0.0000	0.4027
S56	7.0413	0.0000	0.0942
S57	20.7274	0.0000	0.2773
S58	18.0775	0.0000	0.2416
S59	25.6418	0.0000	0.3428
S6	38.4760	0.0000	0.5228
S60	33.5198	0.0000	0.4480
S7	12.5602	0.0000	0.1684
S8	15.4720	0.0000	0.2069

Speaker	f3_t	f3_pvalue	f3_d
S1	-11.6482	0.0000	-0.1676
S10	3.8743	0.0001	0.0518
S11	-6.6520	0.0000	-0.0989
S12	-11.7613	0.0000	-0.1584
S13	2.5396	0.0111	0.0341
S14	5.4685	0.0000	0.0751
S15	2.5022	0.0123	0.0336
S16	-0.4087	0.6828	-0.0055
S17	-16.6947	0.0000	-0.2254
S19	2.6511	0.0080	0.0373
S20	4.4127	0.0000	0.0593
S21	17.8223	0.0000	0.2473
S23	-4.0768	0.0000	-0.0568
S24	-14.0795	0.0000	-0.1933
S25	11.5911	0.0000	0.1549
S26	-38.2101	0.0000	-0.5321
S27	15.2879	0.0000	0.2148
S28	15.6567	0.0000	0.2142
S29	-5.5187	0.0000	-0.0763
S30	33.2046	0.0000	0.4545
S32	4.3088	0.0000	0.0584
S33	18.6862	0.0000	0.2549
S35	20.3847	0.0000	0.2749
S36	38.1377	0.0000	0.5102
S37	27.6742	0.0000	0.3699
S38	8.3755	0.0000	0.1125
S39	-9.9278	0.0000	-0.1428
S40	-15.7380	0.0000	-0.2174
S41	19.8862	0.0000	0.2757
S42	27.1870	0.0000	0.3713
S43	31.6700	0.0000	0.4243
S44	11.1451	0.0000	0.1491
S45	3.9610	0.0001	0.0582
S46	-34.4564	0.0000	-0.4811
S47	-6.3634	0.0000	-0.0873
S48	-2.1188	0.0341	-0.0289
S49	-3.4118	0.0006	-0.0460
S51	-3.0220	0.0025	-0.0406
S52	-3.4858	0.0005	-0.0473
S53	1.7280	0.0840	0.0233
S54	-5.6040	0.0000	-0.0784
S55	5.0257	0.0000	0.0681
S56	9.6034	0.0000	0.1332
S57	-8.6410	0.0000	-0.1169
S58	-13.9991	0.0000	-0.1971
S59	29.7433	0.0000	0.4005
S6	10.8289	0.0000	0.1561
S60	9.5089	0.0000	0.1294
S7	-40.7000	0.0000	-0.5721
S8	-1.2608	0.2074	-0.0169

Speaker	Cov_t	Cov_pvalu	Cov_d
S1	3.7788	0.0002	0.2398
S10	1.5481	0.1219	0.0951
S11	4.0198	0.0001	0.2549
S12	1.9822	0.0477	0.1193
S13	1.1559	0.2480	0.0702
S14	-0.2056	0.8372	-0.0124
S15	-0.5823	0.5605	-0.0349
S16	0.8494	0.3959	0.0510
S17	2.8259	0.0048	0.1793
S19	-1.6581	0.0976	-0.1004
S20	-1.3912	0.1644	-0.0837
S21	-1.8303	0.0675	-0.1099
S23	2.8399	0.0046	0.1880
S24	-2.6918	0.0072	-0.1611
S25	1.3887	0.1652	0.0862
S26	1.7827	0.0749	0.1068
S27	2.6956	0.0071	0.1648
S28	-2.7468	0.0061	-0.1679
S29	1.5263	0.1272	0.0913
S30	-2.1769	0.0297	-0.1352
S32	0.1262	0.8996	0.0076
S33	1.3767	0.1689	0.0887
S35	-1.4330	0.1521	-0.0857
S36	3.5008	0.0005	0.2116
S37	0.1694	0.8656	0.0103
S38	-2.3748	0.0177	-0.1427
S39	2.5304	0.0115	0.1575
S40	0.5864	0.5577	0.0355
S41	-2.4216	0.0156	-0.1464
S42	-0.2093	0.8342	-0.0131
S43	-2.2089	0.0274	-0.1322
S44	-2.3921	0.0169	-0.1464
S45	1.2223	0.2219	0.0740
S46	-2.7680	0.0057	-0.1656
S47	-4.5395	0.0000	-0.2754
S48	-1.8025	0.0718	-0.1132
S49	-0.1114	0.9113	-0.0067
S51	-0.2080	0.8353	-0.0124
S52	0.4875	0.6260	0.0298
S53	0.3506	0.7259	0.0216
S54	-1.0029	0.3161	-0.0600
S55	2.1145	0.0347	0.1376
S56	-0.0895	0.9287	-0.0056
S57	1.8142	0.0700	0.1162
S58	1.0072	0.3140	0.0607
S59	-0.3376	0.7357	-0.0210
S6	0.3060	0.7597	0.0184
S60	2.5077	0.0123	0.1533
S7	-3.0846	0.0021	-0.1854
S8	-5.3481	0.0000	-0.3319

Appendix 8 – Bootstrapped t-test with 100 replications of 200 random measurements

Speaker	F1_MeanT	F1_SDofT	F2_MeanT	F2_SDofT	F3_MeanT	F3_SDofT	Cov_MeanT	Cov_SDofT
S1	-0.3334	1.0050	2.4505	1.0387	-1.1572	0.9655	1.5635	0.8323
S10	-0.2880	0.9099	2.9460	1.0557	0.4323	1.0459	0.5982	0.9546
S11	1.2704	1.0072	0.6913	0.9747	-0.4467	0.9209	1.9171	0.8289
S12	0.8125	0.9353	0.8316	1.1030	-1.0798	1.0180	0.7791	0.9632
S13	2.3299	1.0220	1.6537	1.0486	0.0663	1.0590	0.4675	0.9372
S14	-1.1588	0.9538	2.4380	0.9071	0.4263	1.0329	0.1059	0.9460
S15	0.5828	0.9911	2.2237	1.0238	0.1929	0.9807	-0.2680	1.0520
S16	-0.2935	1.0133	2.4415	1.0226	0.0152	0.9937	0.1262	0.9057
S17	1.1374	0.9694	1.0446	1.1624	-1.5147	1.0507	1.1551	0.9959
S19	2.6466	0.9801	0.9370	1.0933	0.3067	1.0095	-0.7392	0.9757
S20	0.3840	1.0043	1.0674	1.1293	0.2210	1.1422	-0.5364	0.9333
S21	-1.3398	1.0789	2.8139	1.0031	1.6830	0.9591	-0.8178	0.9265
S23	1.3849	1.0026	3.2051	1.0055	-0.2857	1.1019	1.2688	0.7410
S24	0.4150	1.0519	1.9821	1.0218	-1.2609	0.9320	-1.1517	0.9819
S25	-0.0919	0.9126	3.0160	1.0237	1.3352	1.0166	0.6077	0.9591
S26	-0.0302	0.9167	3.1012	1.0652	-3.4685	0.8964	0.7720	0.9741
S27	-0.5900	1.0495	1.5324	1.0595	1.5143	0.8774	1.1793	0.9100
S28	-1.9372	1.0796	1.8942	1.0219	1.4423	1.0683	-1.1572	0.8871
S29	1.2438	0.9544	2.3891	1.1398	-0.5192	1.0354	0.5992	0.9537
S30	0.7074	1.0011	2.0124	0.9585	3.0459	1.0537	-0.7683	0.8431
S32	-0.6590	1.0531	3.5457	0.9151	0.4101	1.0473	0.0412	0.9384
S33	0.9973	0.9746	2.3576	1.0556	1.8288	0.9545	0.4810	0.8601
S35	-1.0738	0.9966	1.0439	0.9975	1.9871	1.0366	-0.5986	0.9045
S36	-1.7766	0.9294	3.5094	1.1536	3.5228	0.8988	1.4279	0.9274
S37	2.4482	0.8533	2.7884	1.0187	2.6477	1.0180	0.0290	0.9078
S38	0.9007	0.9882	1.6274	1.1087	0.8713	1.0299	-1.0743	0.8889
S39	2.7652	0.9607	1.8514	0.9326	-0.8424	1.1912	1.0616	1.0092
S40	0.1250	1.1236	1.6935	1.0593	-1.3742	1.0567	0.2837	0.8980
S41	-0.4226	1.0178	2.1429	0.9901	1.8025	0.8114	-1.0665	0.8886
S42	-1.3625	1.0775	3.3977	1.0722	2.5837	1.1008	-0.0901	0.9478
S43	-0.2027	1.0806	2.7702	1.0979	2.7831	0.8727	-1.1178	0.8760
S44	0.5024	1.0313	0.6092	0.9848	1.0620	0.9844	-1.0128	0.8836
S45	-0.4942	1.0021	2.0859	0.9590	0.3856	0.9332	0.4348	0.9221
S46	0.4161	1.0246	0.8595	1.0355	-3.3052	1.0214	-1.1029	0.9997
S47	1.1588	1.0584	1.7758	1.0213	-0.4907	0.8881	-2.0227	0.8800
S48	-0.7365	1.1313	1.5605	1.1049	-0.2172	1.1599	-0.8108	0.9299
S49	0.5003	1.0074	2.0764	1.0011	-0.3776	0.9544	-0.0602	0.9150
S51	0.9308	1.0011	2.4158	1.1037	-0.3736	0.9959	-0.1528	0.8874
S52	-1.1698	1.0180	2.3549	0.9889	-0.3330	0.9805	0.1218	0.8427
S53	0.0765	1.0784	2.7484	0.9997	0.1518	1.0230	0.0726	0.9531
S54	-0.4228	0.9740	2.8754	1.2037	-0.4815	0.9949	-0.4653	0.9478
S55	-1.1063	0.9502	2.8030	1.0944	0.4974	1.0695	0.7535	0.9034
S56	-0.7199	1.0594	0.6142	1.0396	0.8199	1.1552	-0.1058	1.0306
S57	-0.6798	1.0841	1.9497	1.0021	-0.8716	0.9615	0.7940	0.8496
S58	-0.1662	1.0575	1.6114	0.8724	-1.2434	0.8352	0.5219	0.9236
S59	-0.2738	0.9341	2.5373	0.9412	2.7639	1.1711	-0.0079	0.9119
S6	0.3492	0.9933	3.6225	0.9820	1.0299	1.1387	0.0474	0.9354
S60	-0.0944	1.1024	3.3122	1.0754	0.8588	1.0413	1.0453	0.8000
S7	1.4537	1.0071	1.2751	1.0021	-3.8084	0.9001	-1.3978	0.8351
S8	-0.7437	0.8930	1.3516	1.0868	-0.1204	0.9898	-2.3269	0.9144
Average	0.1474	1.0060	2.1168	1.0352	0.2623	1.0075	-0.0119	0.9163

Appendix 9 – Two-factor ANOVA of formant values

LTF1

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Language	1	1.416e+06	1415770	65.84	4.89e-16	***
Speaker	98	1.225e+09	12500173	581.35	< 2e-16	***
Residuals	1119900	2.408e+10	21502			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

LTF2

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Language	1	6.285e+09	6.285e+09	24642.5	<2e-16	***
Speaker	98	7.058e+09	7.202e+07	282.4	<2e-16	***
Residuals	1119900	2.856e+11	2.550e+05			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

LTF3

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Language	1	3.120e+07	31196894	335.5	<2e-16	***
Speaker	98	2.081e+10	212380503	2284.1	<2e-16	***
Residuals	1119900	1.041e+11	92983			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Covariance

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Language	1	2.606e+08	2.606e+08	0.108	0.743	
Speaker	98	2.869e+12	2.927e+10	12.092	<2e-16	***
Residuals	55900	1.353e+14	2.421e+09			

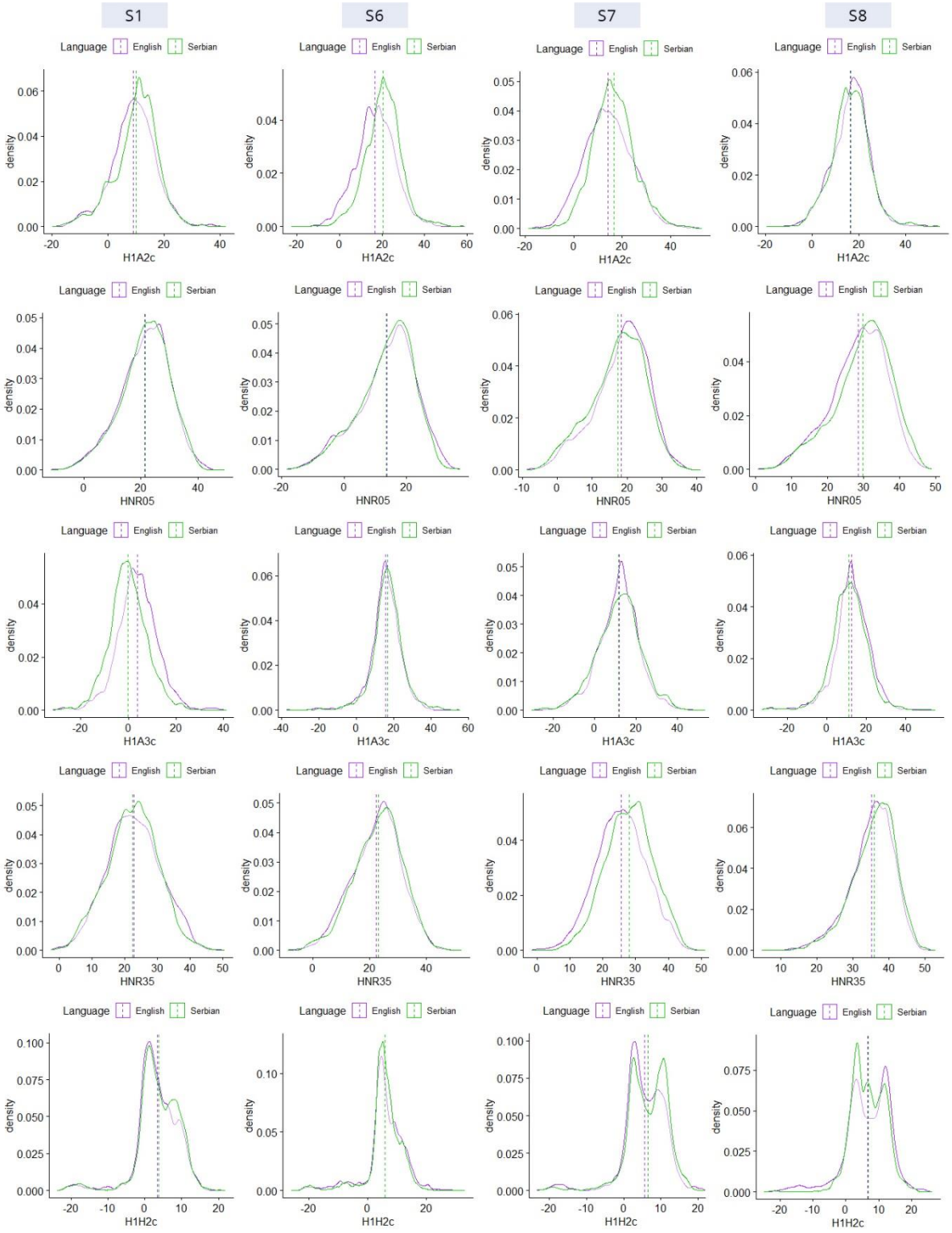
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Frontness**

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Language	1	4.703e+07	47032358	160.5	<2e-16	***
Speaker	98	6.224e+09	63505611	216.7	<2e-16	***
Residuals	1119900	3.282e+11	293086			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Appendix 10 – Density distribution of phonatory parameters per speaker



Appendix 11 - Two-factor ANOVA of phonatory parameters

H1*-H2*

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Language	1	2032	2032	52.58	4.13e-13	***
Speaker	98	12924266	131880	3413.50	< 2e-16	***
Residuals	3999900	154535725	39			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

H2*-H4*

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Language	1	140846	140846	2591	<2e-16	***
Speaker	98	9706879	99050	1822	<2e-16	***
Residuals	3999900	217446324	54			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

H1*-A1*

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Language	1	710000	710000	13845	<2e-16	***
Speaker	98	19799511	202036	3940	<2e-16	***
Residuals	3999900	205119262	51			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

H1*-A2*

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Language	1	3805813	3805813	54787	<2e-16	***
Speaker	98	55791016	569296	8195	<2e-16	***
Residuals	3999900	277854388	69			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

H1*-A3*

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Language	1	113770	113770	1192	<2e-16	***
Speaker	98	87958830	897539	9401	<2e-16	***
Residuals	3999900	381885119	95			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

H4*-2K*

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Language	1	438899	438899	5273	<2e-16	***
Speaker	98	13917405	142014	1706	<2e-16	***
Residuals	3999900	332924972	83			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

HNR05

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Language	1	2154574	2154574	29730	<2e-16	***
Speaker	98	60055954	612816	8456	<2e-16	***
Residuals	3999900	289876664	72			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

HNR15

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Language	1	2278373	2278373	35909	<2e-16	***
Speaker	98	43778281	446717	7041	<2e-16	***
Residuals	3999900	253786130	63			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

HNR25

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Language	1	266308	266308	4320	<2e-16	***
Speaker	98	50940724	519803	8432	<2e-16	***
Residuals	3999900	246583333	62			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

HNR35

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Language	1	24478	24478	466.4	<2e-16	***
Speaker	98	57381039	585521	11157.2	<2e-16	***
Residuals	3999900	209910722	52			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

CPP

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Language	1	63429	63429	9908	<2e-16	***
Speaker	98	5640189	57553	8990	<2e-16	***
Residuals	3999900	25605779	6			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Appendix 12 – Likelihood Ratio results (EER and C_{llr})

GMM-UBM likelihood ratio results

Parameter	Sr_eer	Sr_cllr	En_eer	En_cllr	CL_eer	CL_cllr	CL_eer_sr	CL_cllr_sr	CL_eer_er	CL_cllr_en
LTF1	0.125	0.51	0.1875	0.67	0.3125	1.09	0.31	1.08	0.3167	1.1
LTF2	0.1875	0.57	0.246	0.81	0.375	0.88	0.377	0.89	0.3229	0.89
LTF3	0.253	0.71	0.2396	0.7	0.3146	0.84	0.3729	0.85	0.25	0.83
Covariance	0.25	0.69	0.3125	0.81	0.3125	0.8	0.2646	0.81	0.31225	0.8
Frontness	0.1875	0.54	0.2417	0.71	0.3125	0.86	0.375	0.87	0.3125	0.87
Frontness*	0.1875	0.54	0.2375	0.69	0.433	0.97	0.375	0.96	0.375	0.99
H1*-H2	0.1875	0.62	0.1958	0.73	0.2	0.78	0.1938	0.78	0.2604	0.81
H2*-H4	0.1875	0.62	0.25	0.64	0.2542	0.88	0.3229	0.88	0.25	0.87
H1*-A1*	0.2333	0.71	0.252	0.8	0.2354	1.03	0.2438	1.04	0.1917	1.01
H1*-A2*	0.1875	0.64	0.2458	0.73	0.3125	0.91	0.3146	0.9	0.3146	0.92
H1*-A3*	0.2625	0.68	0.25	0.72	0.3229	0.81	0.3229	0.81	0.325	0.81
H4*2K*	0.25	0.7	0.25	0.7	0.31	0.86	0.3625	0.87	0.375	0.87
CPP	0.1875	0.71	0.2375	0.69	0.2979	0.84	0.2646	0.84	0.31	0.85
HNR_{05}	0.3	0.75	0.3125	0.76	0.3125	0.78	0.3125	0.85	0.2542	0.75
HNR_{15}	0.25	0.79	0.2979	0.72	0.2583	0.77	0.3062	0.81	0.2625	0.78
HNR_{25}	0.25	0.84	0.26	0.75	0.2958	0.75	0.3	0.79	0.3	0.74
HNR_{35}	0.25	0.78	0.25	0.72	0.2583	0.74	0.3	0.76	0.26	0.74
Formants	0.065	0.31	0.125	0.47	0.25	0.97	0.25	0.95	0.25	1.01
Formants + Cov	0.0625	0.26	0.075	0.35	0.25	0.76	0.25	0.75	0.25	0.82
Phonation - all	0.11	0.31	0.0771	0.38	0.125	0.64	0.1896	0.59	0.1312	0.91
Cov, H1*-H2*, HNR_{35}	0.125	0.51	0.1875	0.78	0.1312	0.5	0.1896	0.55	0.1875	0.54
Cov, H1*-A3*, HNR_{35}	0.1333	0.47	0.1833	0.56	0.1396	0.52	0.179	0.54	0.1708	0.53
f3, H1*-A3*, HNR_{35}	0.1771	0.49	0.1271	0.42	0.19375	0.58	0.25	0.59	0.1875	0.59
Sr_eer - questioned, known and background sample in Serbian										
En_eer - questioned, known and background sample in English										
CL_eer - questioned sample in English, known sample in Serbian, background in both languages										
CL_eer_sr - questioned sample in English, known sample in Serbian, background in Serbian										
CL_eer_en - questioned sample in English, known sample in Serbian, background in English										

MVKD likelihood ratio results for tokens averaged over 1 second of speech

Parameter	Sr_eer	Sr_cllr	En_eer	En_cllr	CL_eer	CL_cllr	CL_cllr_calib	CL_eer_sr	CL_cllr_sr	CL_cllr_sr_calib	CL_eer_er	CL_cllr_en	CL_cllr_en_calib
LTF1	0.304286	0.890042	0.319592	1.078011	0.36	1.35449	0.885039133	0.31102	1.246229	0.780146058	0.299388	1.2220139	0.781330584
LTF2	0.334898	0.800343	0.380204	0.927203	0.5	2.907815	0.983637968	0.379796	3.049604	0.866854123	0.399592	2.735799	0.881472632
LTF3	0.202041	0.741207	0.25449	0.754256	0.276122	2.294456	0.797791554	0.220204	2.646902	0.6604176	0.220816	1.5899217	0.660489633
Covariance	0.319592	0.996102	0.329592	0.900685	0.400612	0.990265	0.914608584	0.384082	1.144946	0.881966622	0.395714	1.0825027	0.883297565
Frontness	0.339388	0.842002	0.399796	0.983072	0.502041	2.655475	0.990106087	0.38	2.47876	0.877522058	0.401633	2.3891481	0.894802891
Frontness*	0.33551	0.839996	0.400408	0.989071	0.399796	1.039415	0.945964456	0.42	1.08253	0.885764612	0.396122	2.5374362	0.881926343
H1*-H2	0.259796	0.690818	0.261837	0.822026	0.260204	1.067609	0.79997346	0.240204	1.045975	0.686601725	0.244286	0.9451328	0.693141651
H2*-H4	0.294082	0.900873	0.299796	0.742759	0.382857	1.224797	0.886337844	0.337959	1.302658	0.817599604	0.339592	1.3316579	0.8208186
H1*-A1*	0.219184	0.614472	0.199592	0.623403	0.339796	1.988019	0.870909008	0.28	1.612531	0.736119971	0.279796	1.6228507	0.739261764
H1*-A2*	0.180408	0.509906	0.22	0.640102	0.359796	3.335539	0.861174359	0.256327	3.372963	0.738920141	0.259592	2.7747363	0.737979822
H1*-A3*	0.195102	0.520796	0.219592	0.747464	0.280408	1.369701	0.725292334	0.224694	1.316354	0.609563462	0.224898	1.1687664	0.611475104
H4*2K*	0.340612	0.81541	0.321633	0.900871	0.320408	1.371357	0.936104882	0.337959	1.218649	0.830809393	0.341633	1.101082	0.826756936
CPP	0.180204	0.60612	0.200204	0.60042	0.257143	1.353851	0.668500646	0.200408	1.468824	0.585516984	0.200816	1.5420459	0.582047787
HNR_{05}	0.199388	0.678735	0.236327	0.796165	0.279388	1.399536	0.71911399	0.265306	1.861645	0.652027678	0.27551	1.9058027	0.651040476
HNR_{15}	0.239592	0.856845	0.238367	0.728187	0.336327	2.090738	0.81640522	0.262449	2.407922	0.717336692	0.271837	2.3713809	0.717742786
HNR_{25}	0.22102	1.002605	0.218367	0.693355	0.3	1.731858	0.778694992	0.220204	1.659682	0.651102351	0.22	1.5824432	0.651909204
HNR_{35}	0.220204	0.94015	0.18449	0.642644	0.279592	1.619448	0.715421838	0.20449	1.38745	0.578718498	0.20449	0.9381803	0.586875796
Formants	0.120204	0.384732	0.18	0.532465	0.260204	0.716975	0.716975174	0.124082	0.456905	0.456905164	0.125306	0.4691088	0.469108841
Formants + Cov	0.119796	0.367258	0.155306	0.484309	0.201429	0.684359	0.684359437	0.120204	0.416786	0.4167861	0.13	0.4288984	0.42889843
Phonation - all	0.016327	0.049283	0.022653	0.093016	0.082041	0.29822	0.298220046	0.038776	0.142519	0.142518539	0.03898	0.147338	0.147337992
Cov, H1*-H2*, HNR_{35}	0.139388	0.431885	0.14	0.388728	0.199592	0.552232	0.552231827	0.117755	0.364118	0.364117673	0.119388	0.3736272	0.373627175
Cov, H1*-A3*, HNR_{35}	0.1	0.306598	0.103469	0.326406	0.169796	0.50055	0.500549894	0.103061	0.327788	0.327787887	0.103265	0.3333211	0.333321056
f3, H1*-A3*, HNR_{35}	0.04102	0.155213	0.057347	0.203068	0.135918	0.435003	0.435003394	0.060204	0.243009	0.243009411	0.06102	0.2430094	0.244595941

Sr_eer - questioned, known and background sample in Serbian

En_eer - questioned, known and background sample in English

CL_eer - questioned sample in English, known sample in Serbian, background in both languages

CL_eer_sr - questioned sample in English, known sample in Serbian, background in Serbian

CL_eer_en - questioned sample in English, known sample in Serbian, background in English

CL_cllr_en_calib - questioned sample in English, known sample in Serbian, background in English - calibrated scores

MVKD likelihood ratio results for tokens averaged over 2 seconds of speech

Parameter	Sr_eer	Sr_cllr	En_eer	En_cllr	CL_eer	CL_cllr	CL_cllr_calib	CL_eer_sr	CL_cllr_sr	CL_cllr_sr_calib	CL_eer_er	CL_cllr_en	CL_cllr_en_calib
LTF1	0.301837	0.847693	0.319796	1.047647	0.360204	1.265587	0.884559825	0.312041	1.144192	0.781854742	0.300408	1.1697827	0.7824634
LTF2	0.323673	0.800334	0.38	0.904667	0.499592	2.583953	0.983516708	0.380408	2.649628	0.866318193	0.398367	2.4345703	0.88177901
LTF3	0.201837	0.66511	0.244898	0.707945	0.276327	1.987353	0.797076723	0.22	2.226092	0.660986129	0.22102	1.406423	0.661304625
Covariance	0.323265	0.974747	0.319388	0.8987	0.401633	0.970876	0.913545774	0.394898	1.110126	0.883253579	0.397959	1.0553366	0.884413338
Frontness	0.339796	0.840717	0.400816	0.9572	0.502653	2.378671	0.990020192	0.38	2.179677	0.877562088	0.403673	2.1487973	0.895100123
Frontness*	0.339796	0.838955	0.399796	0.962869	0.400204	1.010639	0.945479823	0.420204	1.037215	0.887765443	0.420816	1.0496227	0.8912619
H1*-H2	0.259796	0.674398	0.26	0.787296	0.260612	0.994739	0.79792152	0.241224	0.970879	0.687912607	0.247143	0.8869189	0.694845575
H2*-H4	0.284286	0.867889	0.300612	0.737397	0.383061	1.161759	0.885711119	0.340612	1.463344	0.817781143	0.339796	1.2905787	0.821586752
H1*-A1*	0.219796	0.60926	0.201837	0.618958	0.32551	1.800785	0.870257432	0.280408	1.438022	0.736609671	0.280408	1.5051093	0.739929647
H1*-A2*	0.179796	0.505588	0.209796	0.610731	0.360204	2.851954	0.860885317	0.256327	2.948564	0.739395981	0.260408	2.3231053	0.738970521
H1*-A3*	0.180816	0.512352	0.219388	0.704284	0.28	1.193839	0.724045195	0.235102	1.132733	0.611379862	0.235306	1.0454216	0.613573257
H4*2K*	0.339796	0.79347	0.323061	0.883864	0.321633	1.294634	0.935566615	0.328776	1.134708	0.831920475	0.343469	1.0634943	0.827510756
CPP	0.180612	0.562742	0.19898	0.568065	0.258571	1.148803	0.666965231	0.200816	1.221612	0.58699537	0.200816	1.3070334	0.582926053
HNR_{05}	0.199592	0.616806	0.237347	0.72015	0.279388	1.167122	0.717111972	0.275306	1.486762	0.653610611	0.275918	1.5931171	0.652355156
HNR_{15}	0.237551	0.771851	0.240408	0.681411	0.336735	1.775948	0.815773221	0.263061	2.002937	0.718463625	0.262653	2.0305407	0.718909958
HNR_{25}	0.221837	0.866917	0.218571	0.66043	0.3	1.495557	0.778305435	0.219592	1.380282	0.653157942	0.220816	1.4146317	0.653455819
HNR_{35}	0.220204	0.816079	0.18449	0.610424	0.28	1.388914	0.71524939	0.20449	1.144722	0.581786275	0.215306	1.1584514	0.581862886
Formants	0.120204	0.383067	0.18	0.53086	0.260204	0.71581	0.715810311	0.125102	0.457431	0.457430541	0.134694	0.4704559	0.470455887
Formants + Cov	0.120204	0.365995	0.147347	0.484358	0.200204	0.682654	0.682654094	0.12	0.417374	0.417373887	0.12102	0.4305805	0.430580501
Phonation - all	0.016735	0.049522	0.022245	0.093121	0.080612	0.296069	0.296068558	0.038776	0.143958	0.143958294	0.039184	0.1486507	0.148650693
Cov, H1*-H2*, HNR_{35}	0.12449	0.430452	0.139796	0.388355	0.199388	0.550515	0.550515037	0.118367	0.367535	0.36753525	0.117959	0.3726663	0.372666337
Cov, H1*-A3*, HNR_{35}	0.099796	0.307513	0.104082	0.326189	0.179592	0.498928	0.498927749	0.114898	0.330589	0.330588816	0.102653	0.3327746	0.332774575
f3, H1*-A3*, HNR_{35}	0.04102	0.154673	0.057551	0.201528	0.135714	0.434073	0.434072609	0.06102	0.244442	0.244441689	0.061429	0.2441073	0.244107314

Sr_eer - questioned, known and background sample in Serbian

En_eer - questioned, known and background sample in English

CL_eer - questioned sample in English, known sample in Serbian, background in both languages

CL_eer_sr - questioned sample in English, known sample in Serbian, background in Serbian

CL_eer_en - questioned sample in English, known sample in Serbian, background in English

CL_cllr_en_calib - questioned sample in English, known sample in Serbian, background in English - calibrated scores

About the Author

Kristina Tomić has obtained her Bachelor's and Master's degrees in English language and literature at the Faculty of Philosophy, University of Niš, with an average grade of 9.31 and 9.86, respectively. She developed an interest in forensic phonetics during her studies, and she entered the vast world of science in 2014 by defending her Master's thesis titled "Temporal parameters of spontaneous speech in cross-language forensic speaker comparison" under the supervision of Prof Dr Tatjana Paunović. She was a trustee of the Studentship of the Republic of Serbia as well as the competitive Studentship of the Foundation for the Development of Young Scientists and Artists. Since 2010, she has been entitled to membership in the association for people with high intelligence quotient - MENSA.

From 2014 to 2021, Kristina worked as a distance-based teacher of English for "Tutor ABC", a company registered in Taipei, Taiwan, later known as "iTutorGroup", based in Hong Kong. During this period, Kristina has taught over 8,600 classes, and more than 8,200 students have enjoyed lectures in her digital classroom. In 2021, she began working as an independent writer and language consultant while also engaging in international casework as a forensic linguist. Since 2023, Kristina has established continuous collaboration with JP French International, a forensic audio and speech laboratory which emerged as an amalgamation of the casework division of the Centre for Forensic Phonetics and Acoustics of the University of Zurich and JP French Associates, a company from York, Great Britain, with a thirty-five-year tradition in speech and audio forensics. Since 2024, in addition to participating in forensic casework, she has also assumed the role of a general manager, overseeing a team of six people and the flow of dozens of forensic cases at any given time.

She has authored numerous scientific papers and regularly participates in international scientific conferences. In addition, Kristina is a member of professional associations such as the Applied Linguistics Association of Serbia (under Association Internationale de Linguistique Appliquée) and the International Association for Forensic Phonetics and Acoustics.

Fun facts: She likes to play computer games and has been an inspiration for the creation of a fictional character, linguist Ketrin Kovačević, in the novel "Death on the Emerald Coast" by Zlata Ljubenović Tomić.

Biografija autora

Kristina Tomić je završila Osnovne i Master akademske studije engleskog jezika i književnosti na Filozofskom fakultetu Univerziteta u Nišu sa prosekom 9,31 i 9,86. Interesovanje za forenzičkom fonetikom razvila je još tokom studija a u svet nauke otisnula se 2014. godine odbranivši master tezu sa naslovom „Temporalni parametri spontanog govora u međujezičnoj forenzičkoj komparaciji govornika" pod mentorstvom prof dr Tatjane Paunović. Bila je nosilac Stipendije Republike Srbije kao i prestižne Stipendije Fondacije za razvoj naučnog i umetničkog podmlatka. Od 2010. godine stiče pravo na članstvo u udruženju za osobe sa visokom inteligencijom – MENSA.

Od 2014. do 2021. godine radila je kao nastavnik engleskog jezika na daljinu za kompaniju „Tjutor Ej Bi Si" (Tutor ABC) registrovanu u Tajpeju u Tajvanu, a kasnije poznatiju kao „Aj Tjutor Grup" (iTutorGroup) sa sedištem u Hong Kongu. Tokom ovog perioda, Kristina je održala preko 8600 časova a kroz njenu digitalnu učionicu prošlo je preko 8200 polaznika. Od 2021. godine počinje samostalno da se bavi pisanjem i konsultantskim uslugama u oblasti jezika kao i da sarađuje na međunarodnim slučajevima kao veštak. Kristina od 2023. godine postaje stalni saradnik Laboratorije za forenziku zvuka i govora „Džej Pi Frenč Internešnal" (JP French International) koja je nastala udruživanjem Centra za forenzičku fonetiku i akustiku Univerziteta u Cirihu i „Džej Pi Frenč Asosiets" (JP French Associates), kompanije iz Jorka u Velikoj Britaniji sa tridesetpetogodišnjom tradicijom u oblasti veštačenja zvuka i govora. Od 2024. godine pored forenzičkog rada, preuzima i ulogu generalnog menadžera te je tako danas odgovorna za tim od šest ljudi i tok na desetine forenzičkih slučajeva u svakom trenutku.

Autor je brojnih naučnih radova, redovan učesnik međunarodnih naučnih konferencija i član profesionalnih udruženja kao što su Društvo za primenjenu lingvistiku Srbije (pod Međunarodnom asocijacijom za primenjenu lingvistiku) i Međunarodna asocijacija za forenzičku fonetiku i akustiku.

Zanimljivosti: voli da igra kompjuterske igre i poslužila je kao inspiracija za lik lingviste Ketrin Kovačević u romanu „Smrt na Smaragdnoj obali" Zlate Ljubenović Tomić.

ИЗЈАВА О АУТОРСТВУ

Изјављујем да је докторска дисертација, под насловом

КВАЛИТЕТ ГЛАСА У МЕЂУЈЕЗИЧНОЈ ФОРЕНЗИЧКОЈ КОМПАРАЦИЈИ ГОВОРНИКА

која је одбрањена на Филозофском факултету Универзитета у Нишу:

- резултат сопственог истраживачког рада;
- да ову дисертацију, ни у целини, нити у деловима, нисам пријављивао/ла на другим факултетима, нити универзитетима;
- да нисам повредио/ла ауторска права, нити злоупотребио/ла интелектуалну својину других лица.

Дозвољавам да се објаве моји лични подаци, који су у вези са ауторством и добијањем академског звања доктора наука, као што су име и презиме, година и место рођења и датум одбране рада, и то у каталогу Библиотеке, Дигиталном репозиторијуму Универзитета у Нишу, као и у публикацијама Универзитета у Нишу.

У Нишу, 2024. године

Потпис аутора дисертације:

Кристина Томић

(Кристина Д. Томић)

ИЗЈАВА О ИСТОВЕТНОСТИ ЕЛЕКТРОНСКОГ И ШТАМПАНОГ ОБЛИКА ДОКТОРСКЕ ДИСЕРТАЦИЈЕ

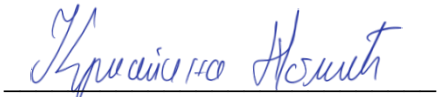
Наслов дисертације:

КВАЛИТЕТ ГЛАСА У МЕЂУЈЕЗИЧНОЈ ФОРЕНЗИЧКОЈ КОМПАРАЦИЈИ ГОВОРНИКА

Изјављујем да је електронски облик моје докторске дисертације, коју сам предао/ла за уношење у **Дигитални репозиторијум Универзитета у Нишу,** истоветан штампаном облику.

У Нишу, 2024. године

Потпис аутора дисертације:



(Кристина Д. Томић)

ИЗЈАВА О КОРИШЋЕЊУ

Овлашћујем Универзитетску библиотеку „Никола Тесла” да у Дигитални репозиторијум Универзитета у Нишу унесе моју докторску дисертацију, под насловом:

КВАЛИТЕТ ГЛАСА У МЕЂУЈЕЗИЧНОЈ ФОРЕНЗИЧКОЈ КОМПАРАЦИЈИ ГОВОРНИКА

Дисертацију са свим прилозима предао/ла сам у електронском облику, погодном за трајно архивирање.

Моју докторску дисертацију, унету у Дигитални репозиторијум Универзитета у Нишу, могу користити сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Creative Commons), за коју сам се одлучио/ла.

1. Ауторство **(CC BY)**
2. Ауторство – некомерцијално **(CC BY-NC)**
3. **Ауторство – некомерцијално – без прераде (CC BY-NC-ND)**
4. Ауторство – некомерцијално – делити под истим условима **(CC BY-NC-SA)**
5. Ауторство – без прераде **(CC BY-ND)**
6. Ауторство – делити под истим условима **(CC BY-SA)**

У Нишу, 2024. године

Потпис аутора дисертације:

Кристина Томић

(Кристина Д. Томић)