The Individual Lay Listener as a Variable in Speaker Individualisation Tasks

Sascha Thomas Schäfer

Doctor of Philosophy

University of York

Department of Language and Linguistic Science

October 2023

Abstract

In some legal cases investigators rely on the testimony of witnesses who have heard rather than seen a crime in order to establish the identity of the perpetrator. These 'earwitnesses' may be invited to take part in a voice parade (VP). In contrast to their well-established visual counterpart, however, VPs are rare, being difficult to design, implement and interpret.

Most recent attempts to improve earwitness reliability have focused on the VP procedure itself, by finding optimal settings for variables that can be controlled by an investigator, such as the quality, duration, and presentation of the stimuli. While such approaches help establish confidence in the procedure, they cannot establish confidence in the individual witness, whose general ability to correctly recognise voices may be questioned in court. The present study therefore takes a different approach by characterising the individual witness as a variable in the identification procedure.

Two hundred and fifty phonetically untrained listeners took part in three psychometric voice processing tests of increasing difficulty: An unfamiliar voice discrimination test (N = 100), an unfamiliar voice recognition test in which participants were aware of the task (N = 100), as well as an unfamiliar voice recognition test in which participants were unaware of the task (N = 50).

The tests were capable of discriminating between a wide range of performances, including a total of four 'super recognisers', who markedly outperformed other participants. Results are discussed in relation to the relevant linguistic, psychological, and legal literature. Moreover, the feasibility of a screening test for earwitnesses is discussed that may provide a trier of fact with nuanced information about an individual witness's ability, thus helping assess the weight of VP evidence.

In Erinnerung an

Alfred Becker & Adolf Schäfer

Table of Contents

List of Tables

List of Figures

Acknowledgements

I am first and foremost grateful to my supervisor Paul Foulkes for his support and guidance over the past four years! Thank you for reminding me to be "clear, accurate and thorough" and for giving me advice when I needed it. Thank you also for the occasional last-minute proofreading![I]

I would also like to thank Cusanuswerk e.V. for their generous support of my doctoral research, as well as the German Academic Scholarship Foundation for awarding me their 'Exposé Scholarship'.

I can't express my gratitude to Jürgen Trouvain for giving me so many opportunities to pursue my interest in phonetics when I was an undergraduate student! Thank you for letting me work on your projects, for allowing me to make mistakes, and for opening countless doors for me over the years. I would never have made it this far without your help!

Many thanks to Ghada Khattab for giving me the chance to gain teaching experience at Newcastle University alongside my PhD. I've probably learned more about phonetics by teaching it than ever before. Many thanks also to Dominic Watt for allowing me to assist with the Forensic Linguistics module at the University of York.

I am grateful to Eleanor Chodroff. You have been a fantastic TAP member! I would also like to thank Shayne Sloggett for his valuable feedback on the experimental design.

I am immensely grateful to my family for supporting me over these past years. I am especially grateful to my parents, Andrea and Uwe, my grandmothers, Luise and Roswitha, my brother Philipp, my aunt Sandra, and my cousin Ida, for their encouragement and for believing in me.

Finally, a special thanks to my friends for keeping my spirits up throughout this journey! I could not have made it through this project without all of your support! I won't even try to list names because I can't risk forgetting anyone. You know who you are, and I am immensely grateful to all of you!

Thank you, Nicky,
हमने इस छोटे से जीवन से एक पूरा जीवन चुरा लिया।

[I] Sorry for making you read so many footnotes…

Declaration

I, Sascha Thomas Schäfer, declare that this thesis is a presentation of original work and that I am the sole author. This work has not previously been presented for a degree or other qualification at this University or elsewhere. All sources are acknowledged as references.

Some of the work presented in Chapters 3 and 6 has already been published in the following journal article:

Schäfer, S. & Foulkes, P. (2023). Towards a screening test for earwitnesses: Investigating the individual voice recognition skills of lay listeners. *International Journal of Speech, Language and the Law* 30(2). 234 - 267. *https://doi.org/10.1558/ijsll.25638*

1. Introduction

The present thesis addresses the reliability of 'earwitnesses', a term referring to witnesses who have heard rather than seen the perpetrator of a crime. This applies to crimes in which there is no face-to-face contact between perpetrator and witness, e.g. crimes committed over the telephone. It also applies to crimes in which the perpetrator's face is concealed. In a wider context, earwitnesses are a particular kind of phonetically untrained listeners, commonly referred to as 'lay' or 'naïve' listeners in the literature. Only the former term will be used in this thesis as the word 'naïve' may be misunderstood as an evaluation. The results of the present thesis thus have implications for lay voice processing beyond the specific field of forensic speech science.

1.1 Motivation and objectives

Earwitness testimony is powerful as it can be decisive in the verdict in legal cases (Robson, 2017). This is potentially problematic, as unreliable earwitness testimony has played a pivotal role in several wrongful convictions across various jurisdictions (Edmond et al., 2011; McGorrery & McMahon, 2017; Sherrin, 2015; Yarmey et al., 2001). The forensic phonetician Harry Hollien therefore concluded in 1996 that "a rather gloomy picture can be painted about the capability of lay individuals to carry out speaker identification of any type" (1996: 14).

Finding ways of improving the elicitation of earwitness testimony has therefore been a recurrent goal of phonetic, psychological, and legal research. This research has led to the development of "voice parades" (VPs), which test the witness's ability to pick the voice of a suspect from a line-up that also includes several 'foil recordings' (Nolan, 2003). VPs have been adopted by various legal systems, including the United Kingdom, the United States, Australia, Canada, and several European countries (McGorrery & McMahon, 2017; Sherrin, 2015; Smith et al., 2020). In contrast to their visual counterpart, however, VPs are not used frequently, being costly as well as difficult to design, implement and interpret. When 43 police forces in England and Wales were asked about their use of VPs in 2015, for example, less than 10% of them reported that they had used the procedure in the previous decade (Robson, 2017).

Considerable research has since been undertaken to improve and standardise VPs. Most recent efforts to do so have focused on "system variables", i.e., the variables that can be controlled by the investigator who sets up the parade (Wells, 1978). This entails, but is not limited to, the quality (McDougall, 2021), duration (Kerstholt et al., 2004; McDougall, 2021;

Pautz et al., 2023; Smith et al., 2020), number (Pautz et al., 2023), and presentation of the stimuli (Smith et al., 2020). While such stimulus-driven approaches help establish credibility in the VP procedure itself, they do not necessarily establish credibility in the individual witness, whose general voice recognition skills may be questioned in court. The present study therefore approaches the problem from a different angle by exploring between-listener differences. The underlying assumption is that listeners differ markedly in their ability to recognise and memorise unfamiliar voices and might therefore not be equally suited for a standardised VP. The individual listener's capabilities can be characterised as an "estimator variable", as they are beyond the control of the investigator (Wells, 1978).

There is a demand for information on witness-specific variables on the part of the legal system. In the UK-based case of *R –v– Flynn and St John* [2008], for instance, Lord Justice Gage characterised "the ability of the individual lay listener to identify voices in general" as one of five variables that have a bearing on the success of an earwitness identification (*R v Flynn and St John* [2008]: §16,3). He further acknowledged that "the ability of an individual to identify voices varies from person to person" (*R v Flynn and St John* [2008]: §16,3,iii). While the main function of a VP is the elicitation of evidence, information on the witness's general capabilities could help determine the weight of such evidence. The Crown Court Compendium's guidelines for dealing with earwitness testimony already account for listener capability to some extent in that a jury must be warned if the witness suffers from any form of "hearing disability or other impediment" (Judicial College, 2023). Such a warning is crucial as it assists the jury "in assessing the weight to be given to a piece of evidence that the jury may feel is a necessary element in finding proof beyond a reasonable doubt" (Sherrin, 2015: 9). It is, however, noteworthy that the guidelines do not consider between-listener differences beyond the presence or absence of a medical condition (Judicial College, 2023).

Examples from the linguistic and legal literature show that the existence of between-listener differences is widely accepted. According to Baldwin and French (1990: 65) it is "certainly true that human beings vary considerably in their auditory attention and memory". This view is seconded by Jeremy Robson (2017: 39), a barrister, who asserts that the ability to accurately identify a speaker "varies greatly from individual to individual". In contrast, experiments conducted on the reliability of earwitness identification have largely focused on the interpretation of performance averages across participants, in order to draw conclusions that are true for an average listener, and hence, for a majority of witnesses (assuming that abilities are normally distributed across the population).

Recently, psychological studies have provided new impulses for the investigation of between-listener differences. Several psychometric voice processing tests have shown that lay listeners differ substantially in their abilities to discriminate, recognise and memorise voices; most notably the *Glasgow Voice Memory Test* (GVMT; Aglieri et al., 2017), the *Bangor Voice Matching Test* (BVMT; Mühl et al., 2018), and the *Jena Voice Learning and Memory Test* (JVLMT; Humble et al., 2022). While these tests differ in various aspects, including the way in which they address memory and learning processes, their common goal is to place listeners on a performance spectrum, ranging from "super recognition", i.e. exceptional voice processing skills, to (developmental) phonagnosia (Roswandowitz et al., 2017; Van Lancker & Canter, 1982) at the lower end, which is the auditory equivalent of 'face blindness' (prosopagnosia). Applicability of these findings to earwitness testimony, however, cannot be taken for granted. A primary reason for this is that the stimuli used in these tests were not created from naturalistic voice samples, but from isolated vowels (GVMT), isolated syllables (BVMT), or strings of pseudo-words (JVLMT). Moreover, the difficulty of the BVMT's and JVLMT's test items was predominantly determined by means of item response theory (IRT) rather than by means of phonetic criteria, i.e., only those items were included in the final version of the test that had demonstrated a certain level of difficulty with listeners in preliminary versions of the test, which included more test items.

Yet, such findings provide an impetus for phonetically informed empirical research on individual differences in voice processing, with a higher ecological validity for earwitness testimony. The present work addresses this problem by presenting and discussing findings from three voice processing tests in which the individual listener was the independent variable. The general hypothesis is that lay listeners differ markedly in their voice processing capabilities and are thus not equally suited for a standardised VP procedure. Three different test designs were employed to simulate three different frame conditions:

1. A baseline voice discrimination test in which the role of memory was reduced to a minimum and participants were aware of the task (Chapter 6).
2. A baseline voice recognition test in which participants were aware of the task (Chapter 7).
3. A baseline voice recognition test in which participants were not aware of the task (Chapter 8).

The difficulty was assumed to increase between tests, so that average performances were hypothesised to decline between tests. What sets the tests apart from existing voice memory tests is the use of naturalistic stimuli, sourced from the DyViS database (Nolan et al., 2009), as well as a method for defining stimulus difficulty on a phonetic basis. While the primary goal is the characterisation of between-listener differences, it is a secondary goal to find explanatory variables for these differences, such as listener age, sex, or reaction time. A further relevant secondary goal is the connection between listener performance and confidence.

The thesis discusses practical implications of between-listener differences for the elicitation of earwitness testimony. In particular, the feasibility of a screening test for earwitnesses is discussed, which could complement the findings of a VP by providing a trier of fact with nuanced information on an individual witness's overall capabilities. The tests conducted in this study fall short of the requirements for such a screening test, as further variables would need to be considered to emulate the complexity of an authentic earwitness scenario. Nonetheless, the results shed light on the general importance of listener-dependent estimator variables in earwitness research. In addition to the implications of this work for forensic phonetic research, the results may be of interest to psychologists studying individual voice processing differences, as the stimuli used in this series of studies are significantly more complex from a phonetic point of view than the stimuli used in most psychological studies addressing this issue.

1.2 Earwitness testimony within forensic speech science

The problems addressed in this thesis are embedded in the wider context of forensic speech science (FSS), which is the application of phonetic, phonological, and general linguistic expertise to legal investigations. Inherent to all problems in FSS is the presence of at least one vocal trace produced by a perpetrator during a criminal event. This trace is commonly referred to as the "questioned sample" (QS; Rose, 2002: 24). Depending on the characteristics of the crime, common tasks performed by forensic speech scientists include:

- **Speaker profiling**: A task conducted in the absence of a suspect. The expert analyses the QS(s) to gather information about the perpetrator, e.g. regional or socioeconomic background. This entails comparisons with reference populations.
- **Forensic speaker comparison:** Recording(s) of a suspect exist, commonly referred to as "known sample(s)" (KS(s)). The expert compares QS(s) and KS(s) before the

background of a reference population to assess the likelihood that they were produced by the same speaker.

- **Administering of VPs: A** recording of a suspect (KS) exist, but no recording (QS) of the perpetrator. The only access to the QS is through an earwitness's memory. The expert facilitates a comparison of the witness's impression of the perpetrator's voice and the KS; usually by presenting the witness with a KS and samples of a reference population.
- **Content identification**: The expert assesses what was said by the perpetrator when the QS is recorded but unintelligible, e.g. for technical reasons, due to a particular accent or pathological speech.
- **Recording authentication**: The expert determines if the QS has been technically manipulated and/or that the QS originated from the criminal event.

A more detailed overview of these tasks is provided in e.g. Hollien (2002), Rose (2002), and Jessen (2012). What the first three entries on this list have in common is that they are focused on the identity of the perpetrator, i.e. the question of who produced the QS. Out of these, both forensic speech comparison and the administering of VPs benefit from the existence of specific reference material in the form of a KS. They can therefore both be characterised as comparison tasks. The role of the expert, however, differs drastically between these tasks. While the existence of a voice recording allows the expert to directly compare the QS and KS, the VP relies on a lay listener to make this comparison, and – on top of that – from memory. The role of the expert is rather that of a facilitator who creates a framework for the layperson's voice comparison, which uses optimal settings for the controllable variables (in line with the present state of research) and also provides safeguards against misidentifications.

1.3 Outline of the thesis

Chapters 2 to 4 of this thesis provide an overview and discussion of the relevant literature, building on insights from phonetics, psychology, and law. Each chapter approaches the topic of earwitness evidence from a different angle.

Chapter 2 investigates the nature of vocal traces, which are the prerequisite for any form of voice comparison. The chapter introduces the concept of speaker identity and addresses the overarching research question of how vocal trace is linked to the individual who produced it. In a first step, the speaker-specific information conveyed by the vocal trace is characterised.

There is a particular focus on the accessibility and saliency of this information to a lay listener. In a second step, the type of reasoning is described that allows for drawing conclusions about speaker identity from observations about the trace. A characterisation and evaluation of the information contained in the speech signal is a prerequisite for assessing the stimulus difficulty of voice perception studies.

Chapter 3 characterises the earwitness's role as a source of information in legal proceedings. This entails differences between earwitness testimony and expert witness testimony. Voice parades are discussed as a way of eliciting earwitness testimony, including their purpose, legal requirements, their current form, and their shortcomings. A detailed literature review is provided regarding the variables that affect earwitness testimony before the background of the already established distinction between system and estimator variables. Lastly, the feasibility of a complementary screening test for earwitnesses is discussed, which may help assess whether an individual witness is suitable for a standardised VP.

Chapter 4 concludes the theoretical observations by summarising what is known about individual voice processing capabilities, mainly from a psychological point of view. General theories are introduced about how voices are perceived, encoded, and stored. Most importantly, a continuum of proficiency is defined, ranging from super recognition to developmental phonagnosia. Three well-established psychological tests that aim to place individuals on this continuum are discussed with regard to their applicability for forensic applications.

Chapter 5 establishes a link between the theoretical observations made earlier and the empirical part of the thesis. It is explained why and how the conducted tests differ from existing empirical studies. The investigated variables are defined, and their choice justified. Moreover, general design principles are explained that apply to all three conducted experiments.

Chapters 6 to 8 report on the voice processing tests conducted. Each chapter addresses a different test. The chapters follow a similar structure. The research questions, hypotheses and methodology are outlined, followed by a presentation and discussion of the findings.

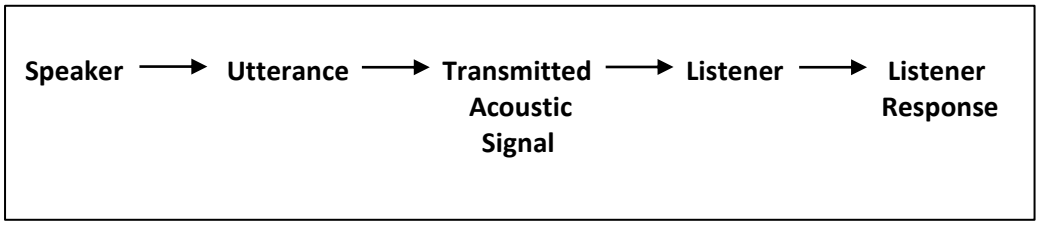
Chapter 9 draws an overall conclusion and provides an outlook for future research.

2. Speech and speaker identity

An investigation of lay listeners' ability to retrieve identity information from another individual's speech requires a general understanding of the information contained in the speech signal. This chapter breaks human speech down into observable constituents and discusses their 'discriminatory power' (Rose, 2002), i.e. the usefulness of a particular feature for the purpose of voice recognition. Rose uses the term "discriminatory power" to refer to an inherent property of an object of investigation, in this case vocal features. Note that this is different from the concept of "discriminating power" as it is commonly used in forensic sciences, which denotes the "the ability of a forensic technique to differentiate between individuals or items" (Robertson et al., 2016: 62).

The purpose of this chapter can be illustrated with the help of the classic speech chain model (Figure 1), which describes voice perception as "a succession of stages that transmit information from a speaker to a listener" (Kreiman, 1997: 86).

Figure 1: Stages of transmitting information from a speaker to a listener (Kreiman, 1997: 87)



The main research questions of this thesis concern the last two links of the chain by asking about the role of the listener in speech perception and the listener's ability to form a particular type of response, i.e. a judgement regarding the speaker's identity. Contrary to this, the present chapter focuses on the speech material based on which such a judgement is made by a listener. It is the overall aim of the chapter to specify the logical relationship between speech and the individual producing it. This entails articulatory characteristics of the speaker who produces the speech, the auditory and acoustic characteristics of the produced utterance, as well as possible acoustic changes to the signal during transmission, e.g. during a telephone call. Moreover, the role of the context in which speech was produced must be considered, as different contexts will have an impact on speaker behaviour and listener expectations. The central question is to what extent the transmitted utterance can be 'traced back' to its source, the speaker.

2.1 Conceptualising vocal identity

2.1.1 Uniqueness in a forensic context

One of the fundamental principles of modern forensic disciplines is Locard's exchange principle. It states that a perpetrator will either leave traces at the crime scene or carry traces from the crime scene (Robertson et al., 2016: 1). In the context of FSS, the traces of interest are vocal traces; particularly, speech that was produced by a perpetrator during a criminal event and that left a trace on a recording medium or in the memory of a witness. It is the goal of some tasks within FSS to establish the identity of the perpetrator based on the comparison of such traces with material from a suspect. This applies to forensic speech comparisons and also to VPs (as established in Section 1.2). The technically correct term for the act of inferring that a trace was produced by a particular source is *individualisation*, rather than the commonly used term *identification* (Kirk, 1963: 236).

For a legal practitioner, especially for a trier of fact, it is essential to have a notion of the robustness of individualisation evidence. Voice-based individualisation procedures have therefore often been compared to individualisation techniques that are based on other biometric features. In this connection, it is generally accepted by forensic experts that voice-based individualisation cannot establish source identity with the same certainty as e.g. fingerprints or DNA patterns (Foulkes & French, 2012: 558). The voice is therefore often characterised as an 'imperfect biometric' (Evans et al., 2014: 126). The precise reasons for the comparatively low robustness of individualisation by voice are less clearly communicated in the literature. Hollien (2002: 7) summarises the "basic problem" of speaker individualisation[1] as follows:

> "It is simply *not* known whether or not every one of the 5 - 6 billion people in the world produces utterances which are unique to them and different from those of all others. That is to say (technically), we really do not know if intraspeaker variability is always less (or smaller) than interspeaker variability and if this relationship is true for all situations and under all conditions. In other words, once the patterns are established for a given speaker, are they actually unique to them or will they vary around the resulting configuration in a manner that causes them to substantially overlap with those of other speakers?" (Hollien, 2002: 7)

[1] Hollien uses the term "speaker identification".

Following the logic of this quote, a speaker's voice exhibits observable "patterns" and the individual products of that source, which are referred to as "utterances", may share these patterns to a certain extent. Vocal traces at a crime scene would be such utterances. The "basic problem" of speaker individualisation is now characterised as the inability to demonstrate the uniqueness of these voice patterns. This presupposes that individualisation is a process based on deductive reasoning, i.e. a conclusion that is drawn from a major and a minor premise, which must both be true for the conclusion to be valid. Thus, a positive individualisation by voice would ideally follow this line of reasoning:

Major premise:	Voices have unique patterns.
Minor premise:	The vocal trace shares the unique vocal patterns of the suspect (to a sufficient degree).
Conclusion:	The suspect's voice produced the vocal trace.

Across disciplines, forensic scientists widely assume syllogisms of this kind to be at the core of the individualisation process (Meuwly, 2006: 207).

The point made by Hollien (2002) in the passage cited above is that this syllogism does not apply to individualisations by voice because there is doubt about the correctness of the major premise. However, a major premise of this sort cannot be established for any type of biometrical individualisation (Meuwly, 2006: 207f.). Individualisation is therefore never based on strict deductive reasoning. For instance, the major premise that the patterns of ridges on fingertips are unique has simply never been falsified. It is merely assumed plausible[2] based on an ever-growing number of observed fingerprints that were different from each other. It will be argued at a later point in this chapter that the assumption of vocal uniqueness does indeed reach a similarly high level of plausibility. In sum, uncertainty about whether voices are unique sources does not set vocal individualisation processes apart from any other biometry-based individualisation procedures. There is thus no 'perfect biometric'.

Maybe more importantly, uniqueness of the source is not a prerequisite for the actual type of reasoning used in individualisation. Up to this point, the term "unique" has been used in this chapter to denote the distinguishability of objects. A voice is unique in this sense if it can be distinguished from (all) other voices, while it is not unique if another voice is found which is so similar that both are indistinguishable. This relationship would be one of qualitative identity.

[2] Meuwly uses the term "verisimilitude", a term coined by Karl Popper (Meuwly, 2006: 208). For the purpose of the present argument, the term "plausibility" is considered sufficiently equivalent.

Qualitative identity: The relation between two objects that share properties. There is a continuum on which objects are "more or less qualitatively identical" depending on their similarity (Noonan & Curtis, 2022: n.p.). At the high end of this continuum is indistinguishability.

It is, of course, unequivocal that even two indistinguishable voices remain separate entities, belonging to different speakers. It can therefore be said that indistinguishably similar (qualitatively identical) voices are still unique in that they are not numerically identical.

Numerical identity: The "relation everything has to itself and to nothing else" (Noonan & Curtis, 2022: n.p.). In other words, numerical identity is the uniqueness of an object in the sense of it being a different entity form all other objects in the universe (Kirk, 1963: 236).

This distinction between numerical and qualitative identity is the basis of the *principle of individuality* in forensic sciences, according to which objects may be indistinguishable but no objects are numerically identical (Robertson et al., 2016: 2).

Crucially, this principle applies to all objects, i.e. to the source as well as to the trace. This means that a particular vocal trace is unique and numerically different from every other utterance, including utterances produced by the same speaker. While this does not rule out that traces produced by the same source are indistinguishable, real-world examples show that most of them are not. Rose, for instance, observes that the same speaker can never repeat "the same thing in exactly the same way" (Rose, 1996: 307, 2002: 10). Although it was shown in this section that such a statement cannot be proven, it demonstrates that the general distinguishability of traces is deemed highly plausible by forensic speech scientists.

The numerical individuality of the trace is particularly important as biometric individualisation is based on a comparison of traces. A fingerprint is, for instance, not compared to a finger (the source), but to another fingerprint (trace) of known origin. The same applies to individualisation by voice. Therefore, even if all voices are unique, it does not follow that an expert or lay listener will correctly establish that two unique utterances were produced by the same voice. The task of individualisation is thus to determine whether necessarily unidentical speech samples are different because they were produced by different sources, or despite the fact that they were produced by the same source. This is, again, not limited to vocal evidence as e.g., two fingerprints produced by the same finger are different objects and very likely to be

dissimilar to some degree (Robertson et al., 2016: 2). Even identification by fingerprint is therefore by no means infallible (Thompson et al., 2014).

In summary, the term "unique" is used in two conflicting ways in the literature; in the sense of distinguishability and in the sense of being a single numerical identity. Neither interpretation does, however, allow for the conclusion that the unanswered question of source uniqueness is a particular problem of individualisation by voice.

It will be demonstrated in the following section that the actual challenge with speaker individualisation is the variability of the source.

2.1.2 The relationship between trace and source

It was shown in the previous section that individualisation is not based on strict deductive reasoning. It will be argued here that individualisation is instead to a large extent the consequence of inductive reasoning as conclusions about the general, i.e. the perpetrator's voice, must be drawn from observations of the particular, i.e. the compared individual samples. While the conclusion of deductive reasoning is considered a fact – provided that the premises are true –, the conclusion of inductive reasoning is merely a probability.

In the quote cited earlier, Hollien (2002:7) metaphorically referred to the voice as a pattern or as having a pattern. The term "pattern" is widely used in connection with evidence, be it the pattern of lines that make up a fingerprint or a pattern of genes that make up a DNA profile. However, unlike a fingerprint or DNA, voices are not made up of unchanging biological properties (Foulkes & French, 2012: 558). Consequently, the hypothetical voice pattern must be fluctuating as it comprises the biological properties of the vocal tract as well as (more or less) systematic ways of using it when speaking. Variability therefore exists within the pattern of a speaker (within-speaker variability) and between the patterns of different speakers (between-speaker variability), causing potential overlaps (Hollien, 2002: 7).

It is difficult to establish a straightforward connection between this idea of source as a fluctuating pattern and the reflection of this pattern on a trace. Therefore, the present section is an attempt to demonstrate the logical relation between vocal traces and sources by means of an explanatory framework that is based on the individual observable utterance (e.g. a trace) as a minimal unit. All other concepts and objects that play a role in speaker individualisation will be described in relation to this unit.

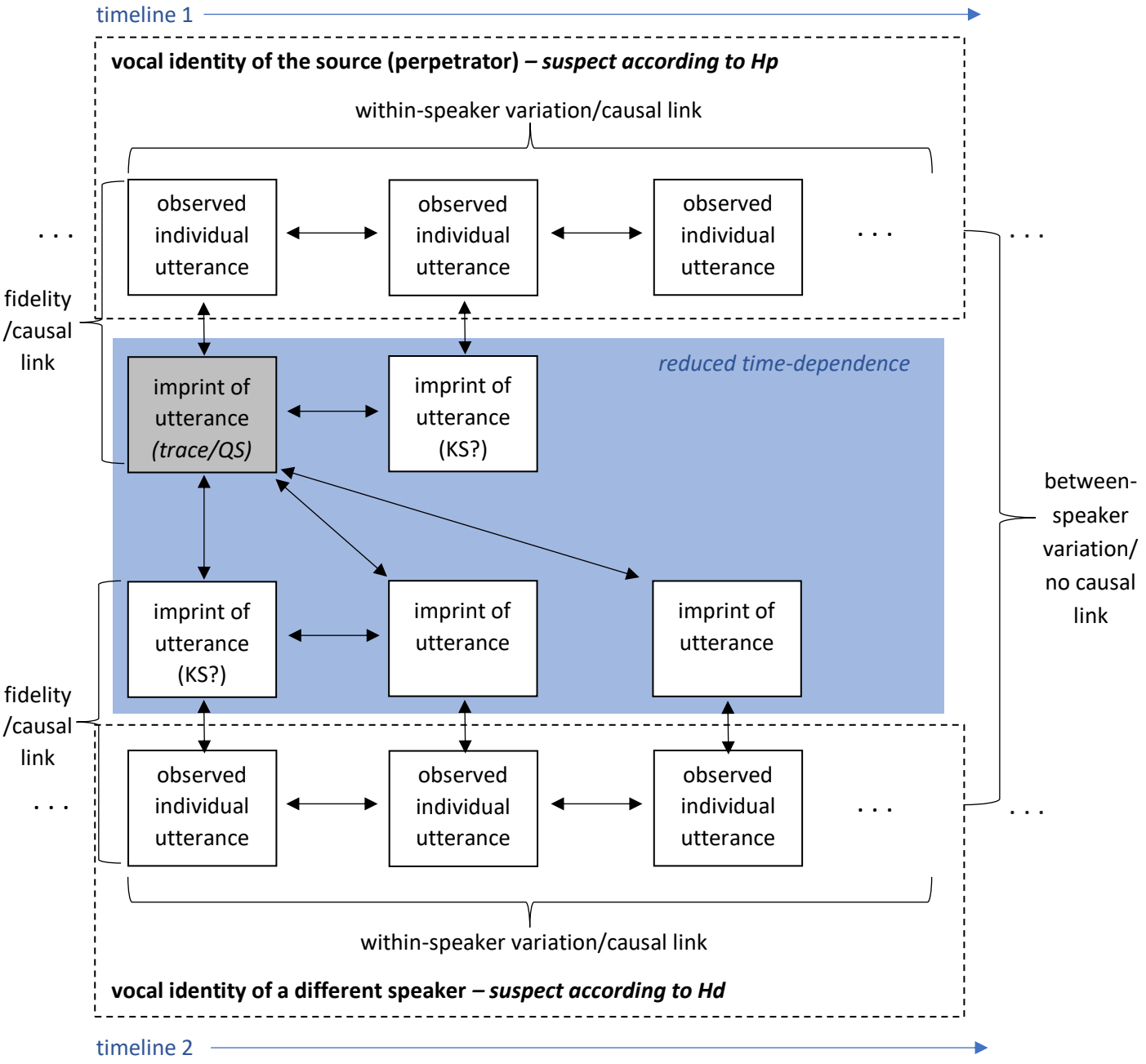
In such a framework, voice (the source), can be defined as the sum of all individual utterances produced by a speaker; including past, present and future utterances. This understanding of voice as an ever-growing sum of utterances will be referred to as "vocal identity". Such a summative vocal identity is, of course, not observable in its entirety by anyone. The problem of observability and the resulting lack of information is, however, the central point being made here. It may be questioned on philosophical grounds whether such a vocal identity is in fact a single numerical identity that changes over the course of time, or rather showing that individuals have numerically different voices at different times. However, it is generally accepted in forensic individualisations that numerical identities can change as long as they maintain organisational continuity. To provide a specific example, an old passport photo is accepted to be a representation of a person's face, although this face might have changed by the time that it is compared to the photograph (Meuwly, 2006: 207). If Rose's claim that speakers cannot repeat an utterance in an indistinguishable manner twice holds true, vocal identities must logically be qualitatively unique because their constituent utterances are unique as well.

Figure 2 illustrates the relations between a vocal trace (greyed out), its source and other objects. Each box represents a numerical identity. In the case of boxes with a solid outline this is an individual utterance, while boxes with a dashed outline represent a speaker's vocal identity. The boxes with a solid outline are of the same size in this general diagram for reasons of simplification. If applied to a specific case, the size could be manipulated to reflect the length or complexity of an utterance. The only indispensable size difference is that a speaker's overall vocal identity must be larger than the utterances of which it is constituted. Three dots inside the dashed boxes indicate that the depicted number of utterances in the diagram is arbitrary. While, in theory, a speaker's vocal identity comprises all utterances produced by the speaker, Rose (2002: 22) rightly points out that for the purpose of a voice comparison large parts of this total identity are not informative, e.g. what the perpetrator sounded like as a child or what the perpetrator will sound like in the future. Consequently, a temporal delimitation is required when comparing voices, based on the timeline of the investigation. Three dots on each side of the dashed boxes indicate that irrelevant parts of the vocal identity were excluded.

Black arrows represent a relationship of qualitative identity (similarity). In this general diagram no distinction is being made between different degrees of qualitative identity. If applied to a specific case, the degree could again be reflected by means of e.g. line width. Speakers produce qualitatively different utterances in different contexts. The differences across

utterances within a vocal identity are referred to as within-speaker variability and differences between vocal identities are referred to as between-speaker variability.

Figure 2: A summative framework of vocal identity



In a forensic context, vocal traces typically exist in the form of what is referred to as an "imprint" in the diagram. The term "imprint" was chosen as an analogy to physical pieces of evidence and the traces that they leave behind at a crime scene. For instance, a shoe worn by the perpetrator can make an imprint at the crime scene. While physical pieces of evidence may additionally be left behind at the crime scene themselves, an imprint is the only form in which speech evidence can persist. This imprint can be a recording, which may then serve as the basis for a forensic speech comparison, or an imprint in the memory of a witness. The imprint is a new numerical identity that is only qualitatively identical with the utterance itself as it is the result of inevitable alterations. For instance, the process of recording can subtract information due to technical limitations (Watkinson, 2002: 160) but at the same time add information in the form of sonic artifacts (Watkinson, 2002: 161). Similarly, memory is generally accepted to be a constructive process, whereby information is lost during memorisation, e.g. due to a lack of attention or due to cognitive limitations, but also added by inference from either the context or from prior experience (Conway & Howe, 2022: 3).

Imprints are not depicted as constituent parts of a speaker's overall identity in the diagram. The rationale for this decision is that differences between the original utterance and its imprint cannot be explained by within-speaker variation, nor are they attributable to between-speaker variation. Consequently, they introduce a third kind of variability, which can be characterised as a confounding variable as it impedes the task of individualisation. For instance, an expert conducting a forensic speech comparison will have difficulty performing the task when the compared recordings are mismatched in terms of recording quality (Alexander et al., 2005; Hughes et al., 2019; Nechanský et al., 2022). In this case, the expert has to abstract away from the differences that were induced by the recording process before they can assess whether the remaining differences are best explained by within- or between-speaker variation. To provide an extreme example, two recordings of the same original utterance can differ acoustically just by virtue of recording quality. This presupposes that imprints are not equally true to the original, which is why this type of variability is referred to as "fidelity" in the diagram.

With the help of the framework provided in Figure 2, an earwitness's situation can be characterised as follows: The witness was exposed to a QS at a crime scene. The QS persists as an imprint in the witness's memory that is a more or less accurate representation of the original trace. Consequently, the memorised QS is in a copy-original relationship with the utterance that it represents. An analogy could be drawn to the relationship between fingerprint and fingertip.

The memorised QS is also in a part-whole relationship with its source, the perpetrator's vocal identity, being one of many utterances produced by that speaker. An analogy could be drawn to a partial fingerprint that does not represent the entire source. At a later point, the witness is exposed to a KS, which is a recorded utterance of a suspect. The KS is also in a copy-original and a part-whole relationship with its own source. Both samples are compared outside of the contexts and timelines in which they were originally produced. It is not clear how representative the characteristics of the particular utterances (QS and KS) are of the entire vocal identity of the speaker(s). Note that additionally the imprint fidelity of the QS may have degraded in the meantime.

The witness's testimony is needed because no one can demonstrate continuity in time between both samples and a common source, which would be the only way of proving both the QS and the KS are part of the same vocal identity (Meuwly, 2006: 209). For this reason, two potential positions for the KS are present in the diagram, one linked to the same vocal identity as the QS and one linked to another speaker's vocal identity. The purpose of the witness testimony is to gather evidence for one of two hypotheses: First, the hypothesis that the suspect is numerically identical with the perpetrator, in which case both samples are part of the same vocal identity. Second, the hypothesis that the suspect is numerically different from the perpetrator, in which case both samples are part of different vocal identities. Since establishing that two recordings were produced by the same source is the incriminating finding in most cases, it is usually the prosecution's hypothesis (Hp) that this is the case, while the defence's hypothesis (Hd) is the assumption that the samples have different sources.

As shown in the diagram, the observable relation between QS and utterances by the same and different speakers is one of qualitative identity (similarity). That is, the QS can share characteristics, e.g. the pitch range or some accent features, with utterances from the same source and with utterances of another source. The only causal relations known is that the QS is an imprint of an utterance produced by the perpetrator and the KS an utterance produced by the suspect. A causal relationship between QS and KS is, however, not observable. This causal relationship is only inferred once either the Hp or the Hd has been accepted. If the Hp is true, all qualitative differences between QS and KS are explained by within-speaker variability (and imprint fidelity). If the Hd is true, all qualitative differences are explained by between-speaker variability (and imprint fidelity). While one of these explanations involves a causal continuity with the source, the other one does not. There is consequently no reflection of causality in observable qualitative identity (similarity) alone.

2.1.3 The role of time

The impact of time on speaker individualisation is threefold, as time is a defining characteristic of comparisons, of voice, as well as of forensic conditions.

First, time is the distinguishing criterion between numerical and qualitative identity (Meuwly, 2006: 207). No two objects can be numerically identical at the same time, while they can be qualitatively identical simultaneously (Meuwly, 2006: 207). Numerical identities can be compared to each other as they are necessarily different (although potentially indistinguishable). A single object on the other hand can only be observed, because the comparison of an object with itself would not be informative (Noonan & Curtis, 2022: n.p.). However, different parts of the same object can be compared to each other; moreover, a copy of a numerical identity can be compared to the original or to another copy. In both cases new numerical identities are created that are qualitatively identical to the original. If perpetrator and suspect are identical, the individualisation is the act of comparing the speech of the same individual to itself. The summative approach to vocal identity demonstrates the unavoidability of a part-whole relation and a copy-original relation between trace and source, which render this type of comparison possible in the first place.

Second, voice has a temporal dimension. The vocal identity changes over time as each new utterance adds further variability to it. Even if the perpetrator is identical with the suspect, the source will have altered in between the production of the QS and the production of the KS (Hollien, 2002: 31-36). This may apply to the biological, cultural and habitual characteristics of the produced utterances. Blue arrows in Figure 2 symbolise the temporal progression of each relevant vocal identity. This differs from predominantly biological biometrics like a fingerprint, which is relatively stable over the course of time.

Third, it is in the nature of forensic investigations that the comparison between a trace and a source will be diachronic, as there is an inevitable delay between the productions of the compared samples. Both the QS and the KS are preserved in time in the form of an imprint (a copy). They are consequently compared outside of the original timelines in which they were produced (cf blue arrows in Figure 2). The imprint condition facilitates a less time-dependent analysis (cf. blue rectangle in Figure 2). The analysis is, however, not completely time-independent, as imprints can deteriorate. This is especially true for memory imprints in witness cases. In an early study by Frances McGehee (1944) for instance, lay listeners were able to recognise a voice previously heard on the telephone with an accuracy of 85% after a two-day delay, whereas a two-week delay reduced accuracy to 48%. After that, performance appeared

to plateau at 47% and 45% after four and after eight weeks, respectively. The deterioration of memory imprints will be discussed in greater detail in Chapter 4.

2.1.4 Implications for earwitness accuracy

Following the above observations, several tentative hypotheses about the accuracy of earwitness testimony can be formed on logical grounds alone:

1. **Lay listeners may run a higher risk of producing false identifications when confronted with similarity, and conversely, may not notice source identity when confronted with dissimilarity**. This is because similarity between two compared utterances is not necessarily an indication of source. An overheard utterance is only a more or less representative snapshot of a highly variable identity. The result of an individualisation task is therefore a statement of opinion and not a certain fact (Evett, 1996: 120). A lay listener may not be aware of this.
2. **A related hypothesis is that lay listeners may have difficulty when drawing conclusions from their observations**. A term that might cause confusion in this relationship is "match" (Robertson et al., 2016: 63), which should be carefully defined given that similarities between a QS and KS can exist independent of a logical relationship between the two. Therefore, the reasons for counting similarities as a match should be carefully defined.
3. **Given the part-whole relationship between trace and source, it is predicted that familiar listeners will be better at an individualisation task than unfamiliar listeners because they will have been exposed to more the variability within the source.** Chapters 3 and 4 will focus on the role of the listener and corroborate this claim with empirical findings.
4. **An increasing time delay between the production of the QS and the KS is hypothesised to have an adverse effect on recognition accuracy because time affects both vocal identity of the speaker and the memory of the listener.** While the QS is fixed in time by means of imprint, the source keeps changing, potentially altering the observable similarities between QS and KS in cases where the suspect was the perpetrator. The possible deterioration of imprint fidelity may add to this effect.

5. **Individualisation is likely to be easier, the longer the compared utterances are.** The underlying assumption is that longer utterances, all else being equal, will expose more within-speaker variability. Consequently, longer utterances have the potential of being more representative of the speaker's overall identity. For instance, an individual sound may predominantly provide acoustic information (e.g. pitch), while the next more complex structure, the syllable, may already contain some language-specific information in terms of permissible syllable structures (phonotactics). More complex structures with a semantic meaning, such as words or sentences, may also reflect more regional or cultural information, in the form of accent and dialect features. Experts conducting a voice comparison based on recordings are aware of this and usually establish a minimum duration for a recording to be admissible for the task (Hollien, 2002: 40), e.g. 30 seconds of net speech in case of the German Federal Criminal Police (BKA) (Künzel, 1995). The multidimensional character of speech evidence will be discussed in detail in the following section.

2.2 Potential vocal identifiers

Vocal objects are multidimensional. Rose (2002: 14) observes that a "speaker's voice is potentially characterisable in an exceedingly large number of different dimensions." There is, consequently, a great number of logically independent values by means of which a given utterance can be described. A comparison of two utterances can be more or less complex depending on the number of observed dimensions. In terms of the previously discussed identity relations, the number of observed dimensions will delimit the possible degree of qualitative identity that can be established between the utterances.

The vocal dimensions thus function as *identifiers*. (Note that this term acknowledges that these dimensions carry information about speaker identity, not that their analysis will definitely facilitate individualisation of a voice). It is the aim of this section to describe and categorise possible identifiers and to characterise how they are subject to variability between and within speaker identities.

The following overview of identifiers employs an adaptation of Jessen's categorisation into *organic*, *idiolectal* and *habitual* features of voice (2012: 38 - 40). The label "idiolectal" was changed to "cultural" for the purpose of this discussion, for reasons explained in the respective section. Note that the discussion only aims at highlighting the most salient identifiers and is far from exhaustive. Each category of identifiers will be analysed in two steps: First, the information that is theoretically present in the speech signal is described, and it is explained how an expert listener can use this information when comparing voices. In a second step, it is discussed to what extent and in what form the same information is available to a lay listener, e.g. an earwitness.

2.2.1 Organic vocal identifiers

Organic features of speech are the result of the speaker's physical characteristics, i.e. the characteristics of the organs involved in speech production (Jessen, 2012: 38). The totality of these organs is referred to as the "vocal apparatus". It is largely identical with the respiratory apparatus because the organs involved in speech production are also involved in breathing. The exchange of gases could in fact be characterised as the main function of these organs as it is more vital to the organism (Kreiman & Sidtis, 2011: 27).

2.2.1.1 Spectral and resonant frequencies

The production of speech sounds can be subdivided into three stages: initiation, phonation, and articulation. Initiation is the first stage of sound production, during which airflow is generated. Airflow is a necessary requirement for the production of sounds, which are changes of air pressure over time (cf. Section 2.1). Most speech sounds in all languages coincide with exhalation because they are produced on an egressive airstream originating in the lungs (Roach, 2009: 24).

During phonation, the airflow is for the first time transformed into an acoustic phenomenon (Meuwly, 2001: 4). Phonation occurs inside the larynx, where the pulmonic airflow causes the vocal folds to vibrate in a quasi-periodic manner. In its narrowest sense, the term "voice" refers to this oscillation of the vocal folds; which is why its presence or absence creates so-called "voicing contrasts" in consonants. The frequency of the vocal fold vibration is referred to as fundamental frequency (f0) and can be acoustically measured in the unit Hertz (cycles per second, Hz). Human voices speaking at a normal level have a fundamental frequency spectrum ranging from 75 to 10,000 Hz (Nolan, 2005: 397). The average adult male voice has a mean f0 between 90 and 140 Hz, whereas the mean f0 of female speakers ranges between 180 and 300 Hz. Small children typically have a mean f0 between 300 and 600 Hz (Meuwly, 2001: 4). The sound produced by the vocal folds is complex in nature and also conveys multiples of f0, the so-called harmonics. All of these frequencies can be summarised as "spectral frequencies".

The supralaryngeal portion of the vocal apparatus, i.e. the part above the larynx, is referred to as "vocal tract" (Ladefoged & Johnson, 2014: 4). It consists of three chambers: the pharynx, the oral cavity and the nasal cavities. The airflow can be further modified while travelling though these chambers during articulation, the third stage of sound production. Note that some classifications count the oro-nasal process as a stage of speech production that is separate from articulation, e.g. Ladefoged & Johnson (2014). The interplay of phonation and articulation can be described with the help of the source/filter model, which assumes that the "speech wave is the response of the vocal tract filter systems to one or more sound sources" (Fant, 1960: 15). This means that the quality of the sound produced by the vocal tract (source) is further modified by the resonant properties of the vocal tract. The shape of the vocal tract produces a series of bandpass filters, which block or dampen some frequencies and pass or amplify others. These filters are known as formants (Fant, 1960). Formants are particularly impacted by the length and shape of the vocal tract (Jessen, 2012: 38). According to the model,

source and filter are in principle independent, which means that the formant frequencies and the spectral frequencies can vary independently of one another (Fant, 1960).

Due to the direct link between these frequency measures and the speaker's anatomy, both the spectral frequencies and resonant frequencies provide important information about the vocal tract that produced the speech. Unsurprisingly, experts performing professional voice comparisons routinely analyse these features when comparing voice recordings (Jessen, 2012: 97). This was confirmed by two independent studies which compared the methods used by different international groups of forensic phoneticians (Cambier-Langeveld, 2007; Gold & French, 2011). Both studies found f0 (mean, range, and standard deviation) to be the only acoustic-phonetic feature that was analysed by all participants. Formant analysis was the second most common type of acoustic analysis and was performed by 97% of Gold and French's participant group and 90% of Cambier-Langeveld's participants.

Frequency measurements are subject to substantial within-speaker variability. Formants, for instance, which are indicative of the vocal tract's resonating properties, "are rapidly modified during speech by moving the articulators (tongue, lips, soft palate, etc.)" (Fitch, 2000: 259). They consequently differ for the same individual when different speech sounds are produced, as well as across utterances. This is particularly true for vowels, whose quality changes depending on the active manipulation of tongue position, jaw opening, and lip rounding. Similarly, spectral frequencies are highly adaptable to circumstances. Bradshaw et al. (2022) observed that speakers expose more of their f0-range during the opening of a telephone conversation as compared to mid-conversation utterances. They interpreted this behaviour as the speakers' (potentially non-conscious) attempt to provide the listener with identity-specific information. Correspondingly, perception studies (Belin et al., 2017; McAleer et al., 2014) found that listeners were able to rapidly form personality trait impressions (such as trustworthiness) after brief exposure to unfamiliar voices. The stimuli were shorter than a second and consisted of recordings of the word "hello", which is a typical conversation opener. Both studies found the f0-contour of the stimuli to be a key factor in the formation of personality judgements.

Due to the high degree of within-speaker variability, speech scientists often take long-term measurements, such as the mean f0, or measurements derived from the mean, such as the range and standard deviation (SD). It is the aim of this measure to increase the discriminatory power of the parameter by obtaining a measured value that is more representative of the speaker's overall identity (in the sense of Figure 2) rather than short-term variability (Rose,

2002: 45). The utterances on which these measurements are based can still be selected so that a similar speaking style (e.g. modal voice) is compared across samples; thus avoiding categorical differences in f0 due to the use of phonatory settings which require more vocal effort. It has been hypothesised, for instance, that there is a higher uniformity of f0 values between speakers when shouting, making it harder to tell speakers apart (Blatchford & Foulkes, 2006; Rostolland, 1982).

To this same end, long-term measurements, so called "long-term formant distributions" (LTFD) can be taken for the resonant frequencies (Nolan & Grigoras, 2005). These measurements are usually taken from the totality of the available speech material and aim to capture a speaker's formant dispersion pattern. In this connection, LTFDs can provide a good estimation of the type and degree of within-speaker variation (Moos, 2010). Samples should, however, be matched for speaking style for LTFDs as well, as LTF measurements are generally higher for read than for spontaneous speech, the reasons for which are not well-understood (Jessen, 2012: 116).

2.2.1.2 Voice quality

Another organic identifier is "voice quality" (VQ), which refers to "a speaker's vocal configurations beyond those for individual sounds, pitch and loudness" (Szczepek Reed, 2011: 199) and is an umbrella term for various individual voice qualities. Individual voice qualities can be the result of both phonation and articulation, i.e. of both the source and the filter. Examples of the former type ('phonatory settings') are, e.g. creaky voice, breathy voice, and whisper, which are dependent on the way in which the vocal folds are used during phonation. The latter type ('supralaryngeal settings') comprises e.g. pharyngeal settings, labial settings, and (de-) nasalisation (Laver, 2009).

Earlier studies regarded VQ as a long-term paralinguistic element of speech. Abercrombie (1967: 91), for instance, defines VQ as "those characteristics which are present more or less all the time that a person is talking: it is a quasi-permanent quality". Consequently, this notion of VQ applies to a speaker's overall vocal identity. An example of a long-term paralinguistic voice quality would be the damage to the vocal folds as a consequence of heavy smoking, the so-called "smoker's voice" (Watt, 2010: 78) (Long-term smoking also results in a lower f0 (Damborenea Tajada et al., 1999; Sorensen & Horii, 1982)). In more recent works, VQ is identified as a prosodic phenomenon with particular linguistic functions (e.g. Laver, 1994, 2009). For instance, concrete connections were drawn between individual voice qualities and

specific short-term conversational functions, such as turn-taking (e.g. Szczepek Reed 2001, 2003, 2011).

This means that VQ can establish continuity between several utterances produced by the same speaker if a more long-term form of VQ is at hand, or create variability between utterances if the speaker's use of a particular VQ was motivated by short-term communicative needs. Within these limitations, VQ can provide identity-specific information.

2.2.1.3 The lay listener's perspective – Organic identifiers

Lay listeners only have limited access to the discriminatory power of organic identifiers, especially if they rely on a memory imprint of the voice and perform an ad hoc rather than a systematic comparison. This is the case for earwitnesses. While the analysis of voice quality is predominantly categorical and performed on an auditory basis even by experts, an acoustic analysis of speech frequencies is not available to lay listeners. It is therefore necessary to ask in what form this information is perceptually available to an earwitness. The auditory correlate of f0 is called "pitch" (Roach, 2009: 120). Pitch is known to play an important role in lay listeners' judgments of voice similarity (Baumann & Belin, 2010; Nolan et al., 2011; Remez et al., 1997; Sørensen, 2012). While f0 measurements are continuous, pitch is usually perceived and described in a categorical way, i.e. in terms of an imagined scale ranging from "low" to "high". Instrumentally measurable f0 differences, especially when recalled from memory, may not be perceived by the listener (Roach, 2009: 120), resulting in a covert contrast between two samples. Moreover, the individual listener's idea of a pitch scale is only a "pseudo-spatial" representation of the measurable parameter f0, since individual listeners will conceptualise this scale differently (Roach, 2009: 119). It is therefore not surprising that a voice perception study by Tompkinson and Watt (2018) only found marginal correlations between lay listeners' assessment of pitch and measured f0 values, when exposed to unfamiliar voices. Interestingly, a few individual listeners were – against the general trend – capable of making consistently accurate judgements of relative pitch, which is a first indication that listeners may differ in their individual voice processing capabilities.

Different observations have been made regarding lay listeners' estimations of pitch ranges. Honorof and Whalen (2005) found that listeners were capable of estimating whether speakers were using a high or low pitch relative to their overall pitch range when listening to unfamiliar speakers. Similarly, Zhang et al. (2022) showed that lay listeners were relatively proficient at estimating an unfamiliar speaker's speaking f0-range from brief samples of their

voice (in this case ‘f0 range’ did not refer to the entire phonation range, as the falsetto register was excluded). This skill may come to use when listeners are trying to establish continuity between same-speaker utterances across different speaking styles.

While it can be assumed that experts and lay listeners share an understanding of the concept “pitch”, untrained listeners may have differing concepts regarding the remainder of the discussed organic identifiers. It is, for example, a trend in psychological and psychoacoustic studies to describe the organic properties of voices by means of the dichotomy pitch and “timbre” (loudness may be included as a third property) (e.g. Aglieri et al., 2017; Bruckert et al., 2010; Chartrand & Belin, 2006; Humble et al., 2022; Jenkins et al., 2021). Timbre roughly translates to the “colouring” of the voice (based on the German term “Klangfarbe” that von Helmholtz [1875] used when he first introduced the concept of a tone quality that is different from pitch and loudness). If the perceptual properties of voice are reduced to these two dimensions alone, timbre must necessarily be a combination of the perceptual correlates of several independently measurable organic identifiers, including harmonics, the shape of the vocal tract in terms of LTFDs, phonatory settings, and supralaryngeal voice quality. Although this dichotomy is widely used in psychological studies, it is not well understood whether the dimensional reduction adequately represents lay listeners’ perception. Phonetic research has, for instance, shown that lay listeners are capable of accurate ad hoc descriptions of certain voice qualities, especially laryngeal voice quality, and use descriptive labels that match those used by expert listeners (Watt & Burns, 2012). With regard to the colouring caused by resonant frequencies, on the other hand, Nolan concluded that auditory analysis is “inherently insensitive to between-speaker variation caused by differently shaped vocal tracts” (Nolan, 2005: 403), which consequently applies to lay and expert listeners alike. This is the type of variability that e.g. acoustic LTFD measurements can in principle pick up on. It is therefore debatable whether a two-dimensional approach to organic vocal identifiers is fine-grained enough to represent listener perceptions, especially if connections are to be drawn to articulatory causes of these perceptions.

Moreover, the term “timbre” is used in two conflicting ways in the literature (Chartrand & Belin, 2006; Handel & Erickson, 2001): A technical definition by the *American National Standards Institute* defines timbre as the “attribute of auditory sensation which enables a listener to judge that two nonidentical sounds, similarly presented and having the same loudness and pitch, are dissimilar” (ANSI, 1994: 35). This definition acknowledges that the timbre of the same voice is variable and will change with e.g. loudness and pitch (Handel & Erickson,

2001: 121). There is, accordingly, an observable and systematic interaction between pitch and perceived timbre, as was experimentally shown by e.g. Marozeau & de Cheveigné (2007). Moreover, it is an acoustic fact that sounds with a low f0 have more overtones than sounds with a higher f0 from the same source (Siedenburg & McAdams, 2017: 3).

The second meaning of timbre can be defined as "an invariant quality based on perceivable transformations across pitch and/or loudness that is assumed to underlie the ability to identify one instrument or voice" (Handel & Erickson, 2001: 121). According to Chartrand & Belin (2006: 164) it is this second definition of timbre that ought to be employed by psychoacoustic studies. The idea is that listeners can establish continuity between utterances of the same source based on an invariant tone quality that is unique to the source. This notion of timbre must, however, be actively constructed in the mind of the listener (Siedenburg & McAdams, 2017: 3), given the high degree of variability of its articulatory constituents. To this effect, individual researchers have argued that the term is as unscientific as the term "appearance" in the context of visual sensations, as it can mean different things to different people (Martin, 1999: 43). Particular caution is advised in a forensic context, where it is commonly accepted that the result of an individualisation task can only be an opinion and not a statement of fact (cf. Section 2.1). On the other hand, qualitative identity of two samples can be objectively described by means of timbral qualia without claiming logical continuity between them. It is therefore advisable to differentiate between "timbre as a quality" and "timbre as a contributor to source identity" (Siedenburg & McAdams, 2017: 2), and to acknowledge that only the former type that is objectively observable in (lay) speaker individualisation tasks.

The above explanations conceptualise timbre from a linguistic point of view and outline the complex relationship between perceived timbre and its articulatory and acoustic correlates. It can, however, be observed that the term is used more liberally in psychological research, where it is also used to describe properties of voice itself. This blurs the line between voice production and perception, which may lead to some difficulties when trying to integrate psychological and linguistic findings on voice perception. For instance, a recent psychological study by Humble et al. that investigated between-listener differences in voice recognition refers to pitch and timbre as "anatomical inevitabilities" (2022: 2). A similar study by Aglieri et al. (2017: 108) uses the terms "pitch" and "timbre" to describe the quality of the stimuli from the experimenter's point of view. Both examples imply an underlying paradigm that follows strict physical realism, which is commonly referred to as the "ecological approach to perception" in

psychology. It regards perception as the passive picking-up of already existing environmental information (Clarke, 2005: 17) by the listener. Note that this approach does not completely disregard listener effects. It e.g. accepts the perceiver's need to "tune" their perceptual system to the assumed invariant properties of the source (Clarke, 2005: 25).

Physical realism does have a place in the description of timbre. When describing the sound of musical instruments, for instance, sound quality is largely influenced by the physical properties of the instrument, e.g. the materials it is made of (Clarke, 2005: 18). Note, however, that this type of distinction is mostly categorical and helps differentiate between types of instruments rather than between individual instruments (i.e. tokens). Nolan rightly points out that Laver's concept of voice quality assumes that from the production point of view, the majority of speakers have 'equivalent' vocal apparatuses (Nolan, 2005: 391) in that they are made up of the same components and materials. It follows that human speech sounds are formed by the same category of "instrument". While it is of course uncontested that human vocal apparatuses are not identical (Nolan, 2005: 390), a significant proportion of the individual sound quality is not the consequence of anatomy, but of the speaker's habits of using the vocal apparatus (Nolan, 2005: 389). As outlined in Section 2.2.1.2, the resulting within-speaker differences in voice quality can be of a short-term or long-term nature. A helpful biological analogy is provided by Siedenburg & McAdams, who observe that "a single type of sound-producing object or sound-synthesis algorithm may give rise to a timbral genus that can encompass various timbral species" (2017: 3). The adoption of an ecological approach to timbre would imply that the connection between timbral species and genus is inherent to the signal itself and does not require interpretation on the part of the listener. This raises the general question whether an ecological approach to timbre is conducive to studies that presuppose that listeners differ in their voice processing capabilities and aim to find these differences.

2.2.2 Cultural vocal identifiers

While the previously discussed identifiers can be traced back to organic differences between speakers, further features of individual speech are the result of cultural influences. They are the result of the speaker's various social affiliations. The most prominent conceptualisation of the relationship between an individual and society is Tajfel's *Social Identity Theory* (*SIT*, Tajfel, 1981), which defines "social identity" as "the part of the self-concept that derives from group membership" (Hogg & Vaughan, 2002: 401). This concept of identity is different from the identity relations outlined in Section 2.1, which are all related to personal identity, i.e. the identity relations between and within individuals. Note that social identity is not to be understood as the identity of groups themselves, as *sui generis* entities (Simon & Klandermans, 2001: 320).[3] Instead, social identity is created as individuals consider themselves to be part of a particular group (Assmann, 1992: 132). The individual's sense of belonging to a group is referred to as "groupness" in SIT (Beinhoff, 2013: 19). Since groupness may give rise to the individual's self-categorisation[4], social identity is a flexible construct because it allows the individual to be part of multiple groups, e.g. a nation, a region, a profession, a football club, etc. (McEntee-Atalianis, 2019: 18), as well as to change these affiliations over time. At the same time, certain group memberships are the product of socialisation and are less flexible; this includes e.g. gender and (social) age (Jessen, 2012: 177). In sum, groups are the 'sources' of an individual's social identity (Simon & Klandermans, 2001: 320).

This sociological understanding of source can further inform the concept of 'source' in a forensic scenario as it was outlined in Figure 2. It can be argued that, since the personal identity of the perpetrator has multiple societal sources itself, a QS and KS can be from the same source in a wider meaning of the word, even if perpetrator and suspect are not numerically identical, in that both samples may be traced back to the same community. This broader concept of source is the prerequisite for the observations made in speaker profiling tasks, which are based on the assumption that observable properties of an individual's speech are indexical of its social sources (Foulkes et al., 2019: 93). In a comparison case, an increasing number of shared social sources between QS and KS will increase the degree of qualitative identity between the samples.

[3] Some researchers who follow the Durkheimian school of thought refer to the identity of groups as "collective identity" or "group identity", while other researchers use these terms as synonyms of social identity (Simon & Klandermans, 2001).

[4] "Self-categorisation theory" is sometimes regarded as a separate theory, but will be regarded as an extension of SIT for the purpose of this discussion. For a detailed discussion see (Simon & Klandermans, 2001: 323).

2.2.2.1 Sources of social identity

Language is considered a key determinant of social identity (Jenkins & Setter, 2005: 5). On a general level, for instance, individuals become native speakers of a language by means of socialisation. Similarly, socialisation leads to the individual's membership in more regional language communities. This regional background is indexed by the speaker's accent and dialect. The term "accent" refers to "those features of pronunciation which identify where a person is from, regionally or socially" (Crystal, 2008: 3), while a dialect is a "regionally or socially distinctive variety of language, identified by a particular set of words and grammatical structures" (Crystal, 2008: 142). While this definition of dialect encompasses social influences, the term "sociolect" may be used when a variety is spoken by members of a socio-economic class, professional group, age group, or other social group, independent of geographical area (Wolfram, 2004).

Hughes, Trudgill and Watt (2012: 3) observe that many people, including some researchers, do not make a clear distinction between accent and dialect/sociolect. The distinction is, however, sensible in a forensic context as over the course of the twentieth century and into the present British dialects have undergone a process of levelling, meaning that differences between dialects have decreased (Kerswill, 2003). Especially younger speakers may speak with a regional accent, while their dialect is not distinguishable from *Standard English* (SE), with the possible addition of sociolectal lexis (jargon or slang). Pronunciation is therefore particularly indexical of social identity (Jenkins & Setter, 2005: 5) and – at least for native speakers – "the link between accent and social identity is well established" (Beinhoff, 2013).

Phoneticians performing a forensic speaker comparison employ a combination of auditory and acoustic techniques to assess differences in the quality of speech sounds between two samples (Jessen, 2012: 40). Phonetic differences in the allophonic realisation of phonemes, i.e. the meaning-distinguishing sounds of a language, are to a great extent attributable to accent. Vowel formant values change during the production of different speech sounds as the speaker's modification of the articulators change the resonating properties of the vocal tract. Acoustic measurements can thus help determine the exact quality of a speaker's vowels. (In theory, this also applies to consonants that are vocoids). In this connection, the first (F1) and second formant (F2) are particularly important as they allow for plotting the vowel realisations on a two-dimensional vowel space, which is in turn indicative of the speaker's tongue position during speech production. More specifically, the measurable values correlate with the position of the tongue's highest point on the mid-sagittal plane, in terms of the open-close axis (F1) and the

front-back axis (F2). A similar description of vowel quality can be performed on an auditory basis alone by a phonetician who has been adequately trained in impressionistic phonetics. The latter option may be more reliable when recordings are mismatched for imprint fidelity, as this has an impact on acoustically measured formants and other perceptual elements. The relationship between the third formant (F3) and sound production are less well understood, but correlations with lip-rounding have been observed (Jessen, 2012: 107).

Accents also exhibit allophonic variation in terms of consonant production. A non-exhaustive list of forensically relevant consonant variation across British English accents includes:

- The voice onset time (VOT) of plosives (i.e. the time difference between the release of a plosive and voicing onset of the following vowel).
- The omission of consonant phonemes (e.g. yod-dropping and h-dropping)
- Glottaling/glottalisation of plosives
- Rhoticity
- Consonant epenthesis
- Velarisation of /l/ in syllable codas.

Several features may be added when non-native speech is investigated, e.g. aspiration and VOT differences in the production of plosives. If individual features or a combination of these features coincides across observed samples, it increases their qualitative identity and increases the likelihood of a common source. Note that the comparison of cultural variation might be inhibited when the compared utterances are mismatched for speaking style, e.g. when spontaneous speech is juxtaposed with read speech. This is because read speech usually adheres to the grammar, syntax, and lexis of SE, thus delimiting cultural variation to accent. At the same time, read speech demonstrates less phonetic variability than spontaneous speech.

2.2.2.2 Idiolect

Jessen refers to the discussed cultural identifiers as "idiolectal" speaker characteristics (2012: 176-200), and defines "idiolect" as a variety that characterises an individual speaker rather than a group of speakers (Jessen, 2012: 176). As a concept, however, idiolect is avoided in the present categorisation for several reasons.

Jessen himself points out that the term must be carefully defined and used with caution, as there are broader and narrower definitions of idiolect to be found in the literature (2012:

176ff.). He makes a case for employing a narrow definition of the term because wide definitions, especially by Hammarström (1980), typically suggest that idiolect encompasses the vast majority of a speaker's characteristics, including organic and learned/acquired features alike. He infers that this broad definition of the concept as the sum of all speaker characteristics tacitly insinuates that a speaker can be fully individualised by means of his or her idiolect, which is practically impossible. Jessen does, however, endorse the concept if it is delimited to a combination of variety-dependent and -independent phonetic and linguistic features.

Other authors, most notably Nolan (1991, 2005), are generally sceptical of the concept of idiolect and its applicability in forensic speaker comparison cases. While he accepts the theoretical possibility that "no two humans pronounce everything alike" (Nolan, 1991: 489), his criticism is rooted in the absence of large-scale empirical findings that would corroborate this claim. Moreover, he doubts that the general paucity of speech material available in typical speaker comparison cases, combined with the amount of free variation within a speaker's speech, allows for classifying the specific idiolectal source, especially when the speakers are part of the same homogeneous accent group (Nolan, 1991: 489). This view is in line with the claim made earlier in this thesis, that the uniqueness of an utterance does not make the source (i.e. a numerical identity) uniquely identifiable. In other words, the observed idiolect in a QS, which is necessarily a combination of several cultural identifiers, cannot be extrapolated for comparison with a KS, which was uttered at a later point in time and in a different context.

The reasons for excluding the term "idiolect" from the present categorisation are of a more theoretical nature and are a direct consequence of defining cultural variation by means of Social Identity Theory SIT, which regards groups as the sources of social identity. On the one hand, the concept of idiolect would translate to a group with one member. This results in circular reasoning in that according to the principle of (self-)categorisation, the speaker could only identify as him or herself, which makes idiolect a property of personal identity rather than social identity. On the other hand, SIT allows for a flexible notion of social identity as speakers can change group affiliations over time. If the term "idiolect" was used, it would therefore have to be discussed to what extent individuals might have multiple idiolects rather than one idiolect within a given time frame and context, for instance the time frame and context of a criminal investigation. The speaker might never be in a fixed position, even relative to their own vocal identity.

This does, however, not mean that SIT is completely incompatible with ideas of speaker individuality in the sense of uniqueness. Jessen, for instance, makes the compelling point that

a speaker's biography has an impact on the way in which varietal features are combined in his or her speech. In terms of SIT this relates to the number of group memberships. While such combinations cannot lead to complete individualisation, they can drastically increase the speaker discriminatory power of such features when found across samples (Jessen, 2012: 177). This is particularly true for uncommon combinations of cultural features. Such observations have been applied in real casework. For example, Baldwin and French (1990: 67-68) report on a case in which the QS exhibited an uncommon combination of London accent features as well as foreign features that were later identified as Cypriot Greek.

2.2.2.3 The lay listener's perspective – Cultural identifiers

The lay listener's understanding of cultural identifiers may differ from that of the expert phonetician. It is, for example, not to be expected that a lay listener will make a distinction between accent and dialect/sociolect, given that many experts do not make this distinction (Hughes et al., 2012: 3). At the same time, SIT allows for characterising both the listener and the speaker. Tajfel hypothesises that groupness will allow an individual to segment his "social environment into his group and others" (Tajfel, 1978: 76). It can therefore be assumed that in lay speaker identification cases, the constellation of perpetrator and listener in terms of shared group membership defines the sort of information that the listener can provide.

The lay listener should therefore be able to determine categorically whether the heard speaker is a member of the same social community or not. Empirical findings from Atkinson (2015) support this hypothesis. In his perception study, listeners from the northeast of England were significantly better at recognising northeast accents than listeners from other parts of the UK. Interestingly, another listener group, who were not from the northeast but had familiarity with northeast accents, performed worse than listeners from the northeast but better than listeners who were generally unfamiliar with northeast accents. A similar in-group listener advantage has also been identified for non-native accents. For instance, Shen and Watt (2015) found that L1 English listeners were less accurate than L1 Mandarin listeners when tested for their ability to tell apart Mandarin-, Japanese-, and Korean-accented English. Depending on the size and diversity of a language community, this sort of judgement may have a huge impact on determining the a priori likelihood of compared samples being produced by the same speaker.

Accordingly, forensic phoneticians who have narrowed down the regional accent heard in a sample often consult speakers of that variety to see if they identify the accent as their own (Baldwin & French, 1990: 65). This suggests that social identity equips lay in-group members

with a form of intuitive accent-recognition expertise that is different from the expertise that trained phoneticians can provide.

Empirical findings further suggest that in addition to superior accent-recognition skills, in-group lay listeners also exhibit superior speaker recognition skills than out-group-lay listeners. The so-called "other-accent effect" (Myers, 2001) that is deemed responsible for the weaker individualisation skills of out-group listeners, has been investigated in phonetic and psychological studies. Braun et al. (2018) conducted a study in which lay listeners were asked to pick a previously heard voice from a line-up. Three different line-ups were constructed, each comprising stimuli from one of three different regional accents from the northeast of England. Listeners missed the target voice significantly less often when listener and line-up were matched for accent. Stevenage et al. (2012) obtained similar results in a line-up task with listener groups from Southampton and Glasgow.

The other-accent effect might also apply to a lay listener's ability to describe accent features. Tompkinson and Watt (2018) conducted a questionnaire study that revealed a continuum on which the degree of familiarity with an accent correlated with the accuracy of the listener's accent description. However, the authors only discussed their results with regard to accent familiarity rather than group membership.

Finally, SIT allows for the prediction that a lay listener's judgement of speaker identity may be inseparable from a value judgement regarding the speaker's personality. Hogg and Vaughan (2002: 574) claim, for instance, that accent influences the way in which individuals are evaluated. This may partially be based on the listener's stereotypes about members of the group with which the speaker is associated. "Stereotyping" is a technical term in this context which describes the phenomenon that an individual group member is "assigned all the characteristics perceived to define their group" (Beinhoff, 2013: 21). As a result, a general tendency can be observed to homogenise members of the same group, especially when they are out-group members from the listener's point of view. "These individuals become perceptually interchangeable" (Beinhoff, 2013: 21). Therefore, the constellation of the questioned speaker's, the known speaker's, and the listener's social identities may have a bearing on the accuracy with which a lay listener can perform an individualisation task.

2.2.3 Habitual vocal identifiers

A last category of identifiers comprises features that are acquired, i.e. learned by the speaker; not as a result of socialisation but rather as a result of idiosyncratic behaviour. This type of identifier is therefore largely variety-independent.

2.2.3.1 Disfluencies and intonation

Jessen (2012: 38) observes that commonly analysed habitual identifiers include speech tempo, pausing behaviour, and f0 variability. Following McDougall and Duckworth's definition of disfluency as "any phenomenon originated by the speaker which changes the flow of the speaker's utterance" (2018: 206), both speech tempo and pausing behaviour are subsumed under this term. As an umbrella term, "disfluency" is a useful concept since the various metrics available for analysing fluency are intertwined in some types of measurements. A commonly used metric for speech tempo, for instance, is a speaker's 'articulation rate' (AR), which is the mean number of linguistic units produced per measure of time – usually syllables per second – in a given stretch of fluent speech (Jessen, 2007: 51). In contrast, the metric 'speech rate' (SR), which is also a ratio of linguistic units per measure of time, is based on the total length of a given utterance, i.e. including pauses and disfluencies (Jessen, 2007: 51). Pausing behaviour therefore has an impact on speech tempo when the latter is defined as SR, but not when it is defined as AR.

From a speaker comparison point of view, most researchers have concluded that AR has a higher speaker discriminatory power than SR. In a meta-study, Jessen (2007: 51) argued that the combination of fluency behaviour and articulation rate in a single metric led to greater within-speaker variability because the same speaker may be more or less fluent when using different speaking styles, e.g. when read speech is compared to spontaneous speech. Early empirical studies have confirmed for German (Henze, 1953) and English (Goldman-Eisler, 1968) that AR is less variable than SR across the utterances produced by the same individual. Künzel (1997) confirmed in an experiment with particular focus on speaking style that the ARs of his German participants remained relatively stable across read speech and two different types of spontaneous speech, while the speakers' SRs varied between styles. A speaker comparison performed by an expert may therefore benefit from a separate analysis of tempo and other fluency phenomena (Jessen, 2012: 133).

To this end, Hughes et al. (2016) showed that the acoustic characteristics of filled pauses, such as "uh" or "um", are to an extent speaker-specific and can potentially help discriminate

speakers. A broader approach was taken by McDougall and Duckworth (2017) who proposed a 'Taxonomy of Fluency Features for Forensic Analysis' (TOFFA) that also considers unfilled pauses, repetitions, prolongations, and interruptions. In their study, TOFFA was used to generate a combined fluency metric for 20 speakers of Standard Southern British English (SSBE), that expressed the total number of disfluencies per reference unit of 100 syllables. At the same time, a fluency profile was created for each speaker which showed the proportions of the different types of disfluencies outlined by TOFFA. It was observed that speakers with similar overall fluency scores could have very different fluency profiles (McDougall & Duckworth, 2017: 24f.). A follow-up study (McDougall & Duckworth, 2018) showed that the TOFFA profiles of individual speakers were also relatively stable across two different types of spontaneous speech. Those TOFFA features that are based on a quantitative analysis of duration rather than spectral features, are eo ipso less affected by the fidelity of the recording. These identifiers therefore address the common problem that samples available for a forensic speaker comparison are often mismatched for recording quality and/or speaking style.

Besides disfluencies, intonation patterns can function as habitual identifiers. Intonation is a suprasegmental phenomenon, which means that it is observable across multiple individual segments, i.e. on the phrase or utterance level. While there is no universally accepted definition of intonation amongst phoneticians (Vaissière, 2005: 238), two main types of definitions can be identified: A narrow approach that restricts intonation to the "ensemble of pitch variations in the course of an utterance" (t'Hart et al., 1990: 10) and a broader approach that includes other prosodic properties, such as loudness, segmental length and quality (S. Baumann & Grice, 2006). In a forensic context, the former definition, i.e. f0 variability, is more prevalent. Whereas the discriminatory power of mean f0 measurements has already been discussed as an organic identifier (cf. Section 2.2.1), f0 variability is to a greater extent a result of the speaker's phonatory habits, and reflects the speaker's use of the f0-range that is theoretically available to them from an organic point of view. This variability can be expressed in form of a variation coefficient, as described by e.g. (Kraayeveld, 1997), which is the quotient of a speaker's f0 standard deviation and mean f0, commonly multiplied by 100 [$(SD_{f0}/M_{f0})*100$]. This relative judgement in the form of a percentage is more informative than absolute measurements of f0-standard deviation, because experimental studies have shown the standard deviation to be positively correlated with mean f0 (e.g. Jessen et al., 2005). The variation coefficient thus avoids redundancy in the analysis in cases where mean f0 has already been considered as an organic identifier. It is essential that individual metrics be uncorrelated for an accurate

estimation of speaker discriminatory power if multiple features are to be combined to reach an overall opinion (Rose 2002).

Intonation is, however, a difficult object of study, being functionally complex despite its formal simplicity. In many stress-accent languages, for instance, intonation is used by a speaker to convey pragmatic meaning. Speakers may use intonation to highlight and structure information contained in an utterance or use intonation as a paralinguistic means of conveying their mood or emotions (for a detailed discussion cf. Schäfer, 2017). Consequently, intonation is context-dependent, and its analysis requires the recordings to be matched for speaking style.

2.2.3.2 The lay listener's perspective – Habitual identifiers

Jessen hypothesises that lay listeners may not be able to assess disfluencies and speech tempo separately (Jessen, 2012: 133). The perceptual integration of these two separately measurable identifiers may put the lay listener at a disadvantage when it comes to a systematic comparison of voice samples.

The discussed empirical findings suggest that disfluencies and speech rate might be a more reliable identifier for untrained ears than intonation. This is above all the case because intonation is markedly more variable across speaking styles than disfluencies, reducing a listener's chance to compare samples that are mismatched for style, which is common in a forensic scenario. Evidence from voice memory corroborates this ranking of habitual identifiers. An experiment showed that the impression of pitch is highly susceptible to memory distortions, while the impression of speech rate is not (Mullennix et al., 2009). This might indicate that these aspects of voice are perceived and stored separately.

2.2.4 Challenges of this classification

The categorisation of identifiers into organic, cultural, and habitual identifiers helps demonstrate the multidimensionality of the features that make up a speaker's identity. This clear-cut distinction is, however, predominantly an academic exercise and cannot necessarily be made in practice. This is because many of the observable identifiers discussed above have multiple causes, despite being, in principle, primarily attributable to one of these three categories. Organic features like measured f0 and voice quality do, for example, have a biological foundation, while their use by a speaker is also impacted by group membership (idiolectal) and preference (habitual) (Foulkes, 2020). Moreover, the example of formants has already shown that the same feature can be influenced by both anatomy, i.e. the resonant

properties of a speaker's vocal tract, and the modification of the same vocal tract during the production of speech sounds, whose quality differ depending on the spoken variety, i.e. a cultural feature. There is furthermore tentative evidence that formants may have a habitual dimension as well. Lindblom (1963), for instance, showed in an early study that formants tend to be more centralised during faster stretches of speech. A more recent study by McDougall (2004) on the realisations of the Australian English phoneme /aɪ/ under different conditions, however, found no significant effect of SR on the formant frequencies of the diphthong.

Further empirical findings demonstrate the absence of a clear separation between the three overarching categories. A study by Henton and Bladon (1988) established that the use of creaky voice varied according to the sex and accent of British English speakers. This shows that a particular voice quality, which has been classified as a predominantly organic feature above, can function as a sociophonetic marker of group membership, i.e. a cultural feature. Similar interactions have been described for habitual and cultural features. A recent study by Leemann (2016) revealed that in Swiss German AR is indicative of the speaker's sex and regional accent. The latter phenomenon was observed for the major regional varieties as well as on a more local level.

In general, the various interactions between categories of identifiers are not well investigated. In this connection, studies that control for two of the three categories are of particular interest. This is the case for studies on monozygotic twins who grew up in the same environment, controlling both organic and cultural identifiers. This was the premise for a study by Nolan and Oh (1996) who analysed differences in the realisations of the liquids /r/ and /l/ in stressed syllables, produced by three pairs of monozygotic twins from England. They observed that in terms of formant values, the differences were greater between pairs than within pairs. Nonetheless, the speakers within each pair could be told apart on the basis of their formant distributions. The authors concluded that twins can make different "use of the leeway allowed […] by the phonological system of their language" (Nolan & Oh, 1996: 49). A study by Loakes and McDougall (2010) compared productions of voiceless plosives by three pairs of monozygotic and one pair of dizygotic twins who spoke Australian English. It was observed that frication was relatively stable across productions of individual speakers, while differences could be observed between speakers, even if comparisons are made within twin pairs. While such findings help discern organic, cultural, and habitual influences on vocal identifiers, more large-scale approaches would be needed.

2.3 Evaluating observations of qualitative identity

The described identifiers provide a basis for making individual observations about the quality of an utterance. In this connection, an identifier can take on a particular value through measurement or qualitative assessment. These values can be continuous or categorical. For example, the identifier *mean f0* can be assigned a value on a continuous scale in the unit Hertz by measurement. As shown in the previous section, such values can be similar or dissimilar for samples produced by different speakers and for samples produced by the same speaker. The present section aims to determine ways in which conclusions about speaker identity can be drawn from these observations of similarity or dissimilarity between samples.

2.3.1 Beyond similarity

As outlined in Section 2.1, similarity, i.e. the commonality of properties, establishes qualitative identity between two observed objects. Similarity and qualitative identity are therefore used synonymously in this thesis. At the same time, it was established in Section 2.1 that qualitative identity – and eo ipso similarity – does not allow the observer to deduce source identity. Consequently, the reasoning that leads to a conclusion in an individualisation task must also include criteria beyond similarity.

2.3.1.1 Typicality

Works that describe the process of speaker comparison from the expert witness' point of view (e.g. Jessen, 2012; Rose, 2002) commonly introduce the "typicality" of an observation as an additional criterion. According to Jessen, the underlying idea is that the values that most identifiers can take are not uniform for a given population but follow a particular distribution (Jessen, 2012: 41).[5] Consequently, values in the mid-range of said distribution are more prevalent in the population and thus constitute a more typical find. Values at the perimeter of the distribution are less typical. In other words, such observations are less likely to be obtained by chance if a random sample of the relevant population is taken (Rose, 2002: 307).

For example, a comparison of two samples might reveal that in each sample postvocalic instances of /r/ are not realised, i.e. they feature 'non-rhotic' speech. This observation of similarity is typical for a case set in the Southeast of the United Kingdom, where the absence

[5] While Jessen specifically speaks of a normal distribution, it is not certain whether this can be said for all identifiers.

of rhoticity is a common accent feature. If, however, the relevant population of the case is North American, the chance of randomly obtaining unrelated samples of non-rhotic speech is much lower, as most North American accents are rhotic. In this case, the samples are "similar in ways which set them apart from the rest of the relevant population" (Nolan, 2005: 400). The assessment of typicality therefore increases the evidentiary value of qualitative identity by setting it in relation to the relevant population.

While an expert witness cannot in most circumstances prove that two samples were produced by the same source, they can assess the likelihood of this being the case by considering both the similarity and typicality of their observations. Their conclusion is a statement of opinion that can function as evidence in a legal investigation. This point will be pick up in detail in Section 3.2.

2.3.1.2 Prototypicality

It has already been argued in this thesis that a lay listener's concept of vocal identifiers might differ from an expert's concepts, which means that lay listeners will have different means of assessing similarity. It is therefore equally necessary to describe typicality from the lay listener's point of view. This aspect is especially important as the literature on earwitnesses typically does not make a clear distinction between a lay listener's ability to make observations and their ability to draw conclusions from these observations. The ability to distinguish between similarity and typicality is often associated with expert listeners, as these concepts are discussed in pertinent resources for forensic phoneticians. It is often overlooked that the distinction between similarity and typicality originates in cognitive psychology where it was first conceptualised in Lance Rips' seminal paper "Similarity, typicality, and categorisation" (1989). While Jessen (2012: 40) refers to Rose (2002) when introducing the concepts of similarity and typicality, Rose himself does not refer to any other work when introducing these terms (2002: 306 – 310). Revisiting Rips' original definition of the terms bears great potential for earwitness research as it is already following a cognitive approach.

Rips differentiates between similarity and typicality as different approaches to categorisation, which is the act of deciding "whether an object belongs to a category" (Rips, 1989: 21). When assessing similarity in an earwitness case, the category of interest would be the overall vocal identity of the perpetrator, i.e. the category 'utterances produced by the perpetrator'. The object that is to be categorised is the KS, whose continuity with this category is disputed. In the so-called "resemblance approach" to categorisation (Rips, 1989: 22), the

observer assesses the similarity between the object in question and known category members (Rips, 1989: 21). This approach is of limited use in a typical earwitness scenario (and other types of speaker comparison) as the only known category member is the QS, i.e. the utterance overheard at the crime scene. An exception would of course be a case in which the witness is familiar with the perpetrator and therefore familiar with other category members.

Another approach to categorisation is based on the "typicality of the instance with respect to the category", which Rips defines as "the similarity between instance and prototype" (Rips, 1989: 25). "The more similar an instance is to the prototype, the more typical it is of the category, and the more likely it will be classified as a category member" (Rips, 1982: 25).

Similarity thus accounts for typicality in Rips' approach (Rips, 1989: 25), which is also true for Rose's take on the topic. While for Rose typicality is the similarity of the combined QS and KS to the reference population (Rose, 2002: 306), Rips makes a *prototype* the point of reference to which instances – in our context speech samples – must be similar. The word "prototype" is here to be interpreted along the lines of Rosch's original idea of the concept, i.e. as an object that is perceptually salient in the formation of a category (Rosch, 1973: 330, 348). Correspondingly, *prototypicality* is the "function of how similar an instance is to other category members and how dissimilar it is to members of contrast categories." (Rips, 1989: 26).

In a speaker comparison case, a typicality assessment is to answer questions about an individual in relation to a group of speakers. Consequently, if Rips' definition of typicality is applied to earwitness testimony, the category of interest must now be a group of speakers rather than the vocal identity of the perpetrator. This is because the question is not how prototypical an utterance is for the perpetrator but how prototypical an utterance is for a member of a speech community that is relevant to the case. The prototype is in this case the functional equivalent of the reference population in a speaker comparison performed by an expert.

The central question is how prototypes are formed in the mind of the listener. A reference sample is carefully selected by an expert with the aim of being representative of the relevant population, and the identifiers can be assigned values for the reference population based on measurement or observation. On the other hand, the formation of prototypes might be more subjective and tied to the individual listener's experience. If prototypes are perceptually salient members of a category, the prototype for a speaker of a given accent may be based on specific speakers of that accent which are known to the listener. Moreover, prototypes might be more or less detailed depending on a listener's familiarity with an accent or potentially if speaker and listener share group membership. A related question would be whether different listeners use

different categories to begin with. A speaker who is familiar with an accent may, for instance, have distinct exemplars in mind for different subtypes of the accent. The idea of multiple references voices in the mind of the listener can be formalised as an exemplar-based model of lay voice perception. Exemplar-based models have, however, been less influential than models which assume that listeners form a single general voice prototype as point of reference.

The latter type of model originated in the realm of face recognition. Valentine and Bruce (1986) suggested that individuals extract a general prototypical face from previously encountered faces. Newly learned faces are then memorised in form of the transformations that are required to map that specific face onto the prototype. The same idea was later translated to the vocal domain (Belin et al., 2011: 716 f.). Note that this concept will be picked up again in more detail in Chapter 4 in relation to some influential voice processing models.

2.3.1.3 Distinctiveness

The prototype model of voice perception gave rise to the idea of an abstract multidimensional space, in which listeners store voices. Voice space paradigms have been used in notable recent perception studies (cf. Andics et al., 2010; Bruckert et al., 2010; Latinus et al., 2013; Lavner et al., 2001; Schweinberger & Zäske, 2018). The findings of these studies will be discussed in Chapter 4, which deals with the psychological foundations of voice perception.

> "Within the voice space, perceptually similar voices are thought to be stored in close neighbourhood, while dissimilar voices would be represented further apart. The more a given voice deviates from the prototype in terms of its acoustic parameters, the more distinctive it should be on a perceptual level, and the stronger its activation of voice-sensitive brain areas." (Schweinberger & Zäske, 2018: 544)

The above quote introduces the term "distinctiveness" as an attribute of voices that expresses the level of difference from the prototype. If the prototype is an average representation of voice, this implies that distinctiveness is the perceptual correlate of typicality. While such a distinction between typicality and distinctiveness would certainly be helpful, it has not been a convention in the literature. Several studies that make use of the voice space paradigm refer to stimuli that are perceptually similar to the assumed prototype as 'typical' (e.g. Andics et al., 2010; Mullennix et al., 2011), whereas other studies refer to these voices as being 'less distinctive' (e.g. Bruckert et al., 2010; Latinus et al., 2013; Schweinberger & Zäske, 2018) or use both naming conventions interchangeably (e.g. Mullennix et al., 2011). Moreover, perception studies

outside of the prototype paradigm may use further labels, such as ‘more common’ or ‘less common’ (Sørensen, 2012) or paraphrase the concept (Foulkes & Barron, 2000: 189-194).

The absence of a uniform terminology may cause problems when interrelating results from studies that define the distinctiveness of their stimuli by means of acoustic measurements and studies that define distinctiveness by means of listener ratings. Individual studies do indeed suggest that there is a high correlation between the measured acoustic typicality of a stimulus and its perceived distinctiveness, which would make a combined approach to typicality and distinctiveness permissible. Latinus et al. (2013), for instance, placed stimuli in a three-dimensional voice space according to their average f0, formant distribution (FD), and harmonics-to-noise ratio (HNR); z-transformed by speaker sex. They found a significant Spearman correlation (r_s [CI 95%] = .73, $p < .001$) between the participants’ distinctiveness ratings and the stimuli’s Euclidean distance to the centre of the acoustic voice space. The exact size and homogeneity of their listener group was, however, not specified, making it difficult to assess how representative their results are, and whether similar results would be obtained with a differently defined voice space. In sum, it is generally still not clear how acoustic information is stored in vocal prototypes (Belin et al., 2011: 716).

The perceived distinctiveness of voice stimuli is a central question in earwitness research. A key principle is that no stimulus should stand out in a voice parade to a non-witness, to ensure that all voices are equally likely chosen at random (Broeders & Rietveld, 1995: 33-36). Moreover, information about the distinctiveness of the perpetrator’s voice, which could be obtained by interviewing the witness, is helpful for assessing the witness’s credibility. This is because earwitness research has shown that voice distinctiveness correlates with individualisation accuracy. Mullennix et al. (2011) presented participants with a series of highly distinctive and less distinctive voice recordings. The distinctiveness of the stimuli was determined by means of listener ratings from a previous experiment, although information about the number and characteristics of the listeners was not provided. Participants were under the assumption that they were participating in a vowel categorisation experiment and reinvited for a surprise voice individualisation test in the form of a line-up after one week. The results indicate that non-distinctive foils have a high likelihood of being mistaken for a non-distinctive target.

Similar conclusions were drawn by Foulkes and Barron (2000), as well as by Sørensen (2012). They demonstrated, for familiar and unfamiliar listeners respectively, that speakers with an unusually high or low f0 are easier to recognise than speakers with a typical

f0. Sørensen also found out that the effect extends to voice memory, rather than just to immediate recognition. Note that both studies determined the difficulty of their stimuli by means of acoustic measurements, and are therefore assessing the effect of typicality rather than its perceptual correlate. Establishing a correlation between earwitness accuracy and stimulus typicality (rather than stimulus distinctiveness) might, however, be the more relevant research goal as it would eliminate the need for listeners to rate the distinctiveness of foil voices, ultimately making the procedure of compiling a VP more objective. A novel method for finding foil speakers in an acceptable typicality range was proposed by Gerlach et al. (2023) who used ASR software to identify 'voice twins' on the basis of acoustic measurements.

McDougall (2021: 51) stresses that it is a key task for future earwitness research to find out "what makes a voice more distinctive and/or memorable to listeners". If the terminology established above is used, this translates to establishing a clear relationship between typicality and distinctiveness. It could be added that the prototype hypothesis would be a helpful research paradigm for future studies on voice distinctiveness. To this date, only a few studies on the topic discuss their findings before the background of a theoretical model, be that exemplar-based or prototype-based; a notable exception being the discussed study by Mullennix et al (2011), which employs the idea of a general prototype. The prototype hypothesis allows for making helpful predictions about participant behaviour. Early studies on faces have, for example, established that more distinctive stimuli are recognised faster than less distinctive stimuli (Valentine, 1991; Valentine & Bruce, 1986). (This does not mean that they are categorised faster as a face. In fact, the opposite is the case (Valentine, 1991). The same increased processing speed is true for more familiar stimuli, albeit for unrelated reasons (Valentine & Bruce, 1986: 525; a detailed account of familiar vs unfamiliar voice recognition is given in Chapter 4). However, linguistic studies of voice distinctiveness either rarely take reaction time (RT) measurements, or do not explain why RT measurements were considered as a variable.

The latter is true for a study by San Segundo et al. (2016) in which Spanish and English lay listeners judged the similarity of short speech stimuli (3s) produced by Spanish speaking monozygotic twins. The purpose of the study was to investigate whether lay listeners rely on the holistic perception of VQ when judging the similarity of acoustically similar stimuli. Five male twin pairs were chosen from a corpus of twin speech with the aim of achieving a high degree of similarity within and between pairs. To this end, pairs were selected according to their age, mean f0, and a similarity coefficient, which expressed the difference between the speakers

in a pair, based on the results of a simplified VPA. Twenty English and twenty Spanish listeners were presented with the task to rate the similarity of 90 different-speaker pairs on a 5-point scale, i.e. each speaker was paired with every other speaker, including his twin. Listeners did not know that some of the pairs involved twins. The main statistical analysis consisted of trial-based ordinal mixed effects modelling, which also considered RT as an explanatory variable, albeit without providing a hypothesis or a rationale for including it. No RT effect was found for the English listener group, whereas Spanish listeners were faster in their responses, i.e. the more dissimilar they deemed a voice pair. This effect is not discussed by the authors. It is, however, noteworthy that the effect only applied to the native listeners, reducing the likelihood of the effect generally applying to similarity. Consequently, voice distinctiveness could be taken into consideration as a confounding factor in the experimental setup. Faster RTs might have been caused by one voice in the pair being less prototypical than the other voice, i.e. less distinctive. It might be for this reason that the effect did not apply to the non-native listeners, who may have different voice prototypes as reference points. Although the stimuli of the experiment were mainly restricted to VQ cues to speaker identity, VQ can function as sociophonetic markers (as discussed in Section 2.2.4). This makes it plausible for different listener groups to differ in their prototypical representation of VQ. Valentine stresses that distinctiveness effects can be "explained by the role of knowledge of the population" (Valentine, 1991: 165). Interestingly, when San Segundo et al analysed both groups together the RT effect was confirmed for non-twin pairs, while it was reversed for twin pairs, i.e. those were more likely to be rated as similar if a judgement was fast. This can potentially be traced back to the aforementioned familiarity effect, leading to a speaker's voice appearing to be familiar as the extremely similar voice of his twin was presented in the same trial. In sum, it can be assumed that listeners had different strategies for twin and non-twin pairs, whereby distinctiveness played a greater role in perception as familiarity decreased.

2.3.1.4 Noteworthiness

While typicality and distinctiveness are concepts that are regularly discussed in the literature, another criterion may be relevant when evaluating qualitative identity for the purpose of earwitness individualisation. The criterion takes inspiration from the following observation by Baldwin and French:

> "It is certainly true that human beings vary considerably in their auditory attention and memory, so that a speech event, for example, which would pass one person by, leaving no trace in his auditory memory and exciting no reaction, would strike another as noteworthy in some way, perhaps in the matter of accent, and he would retain an accurate auditory impression of that speech event." (Baldwin & French, 1990)

While evidence for between-listener variability will be discussed in chapter 4, the notion of "noteworthiness" that Baldwin and French evoke might have broader implications for earwitness testimony. Distinctive voices have already been described as more "memorable" than less distinctive voices in the previous section. The question raised here is whether a notion of noteworthiness can be separated from the notion of distinctiveness if different meanings of the word "memorable" are distinguished. On the one hand, the word "memorable" can be used in the meaning of "easily remembered/able to be remembered" (OED, 2023). It is this meaning of the word that applies to the discussion of distinctiveness in the literature, where it is generally accepted that distinctive stimuli are "better remembered" than less distinctive ones (Valentine & Bruce, 1986: 525).

A second meaning of the word "memorable" is "worthy of being remembered" (OED, 2023). It is this meaning that Baldwin and French seem to have in mind when they say that a speech event 'strikes a listener as noteworthy'. In terms of identity relations in a lay speaker comparison scenario, distinctiveness seems to influence the quality of a voice's imprint in the listener's mind, or the decay of the imprint, respectively. Noteworthiness on the other hand would play a role in whether an imprint is created in the first place.

It goes without saying that such a notion is only helpful if it can be shown to be sufficiently independent from distinctiveness. In Section 2.3.1.3, Schweinberger and Zäske (2018: 544) were already cited with the claim that more distinctive voices lead to greater activation of voice-sensitive areas in the listener's brain. This suggests that the saliency of distinctive voices leads to a certain alertness of the listener, which potentially means that distinctiveness and noteworthiness are inseparable. To my knowledge, however, there is no existing research that addresses this question systematically, i.e. research that investigates whether voices that are deemed equally distinctive are equally likely to alert a listener.

What may cause a listener to be alert, apart from distinctiveness, is a matter of speculation at this point. Psychological voice perception research suggests, however, that listeners make assumptions regarding the speaker's character and valence on the basis of voice features (Lavan et al., 2021: 282). The character of the speaker is assumed to be judged against three trait dimensions: trustworthiness, dominance, and attractiveness (McAleer et al., 2014). Mileva and Lavan (2023) could show that first impressions of a speaker's character are formed rapidly in the listener's mind. In their experiment, listeners were exposed to 100 voice recordings of varying length (50, 100, 200, 400, and 800 ms) and asked to provide character ratings. It was observed that stable trait judgements had formed after 400 ms. Research would be needed to establish whether certain judgements alert the listener for various reasons, e.g. speakers who are perceived as untrustworthy or attractive, and therefore lead to better imprints of the voice. If this is the case, noteworthy voices may not necessarily be distinctive. A study by Bruckert et al. (2010), for instance, showed that typical voices are perceived as being more attractive than untypical voices.

Moreover, noteworthiness may be listener-dependent. While trait impressions are usually similar across listeners in the long run, it has been shown that initial trait impressions, which are particularly relevant in an earwitness scenario, are variable and not accurate (Lavan et al., 2021: 282). Like distinctiveness, and unlike typicality, noteworthiness would therefore depend on the specific constellation of listener and speaker.

2.3.2 Speech as representation of the speaker

Section 2.3.1 showed how conclusions about speaker identity can be formed by judging observations of vocal similarity between utterances against a population or against an abstract mental representation thereof. This type of reasoning is mainly informed by between-speaker variability. It is equally important to judge the similarity of utterances against the totality of the assumed source's speech; in other words, to assess what sets the utterance apart within the speaker's vocal identity.

While there is ample reference material available for the first task, an earwitness case is usually characterised by a paucity of reference material for the latter. Nonetheless, it can be analysed whether certain phenomena that may apply to an earwitness case can lead to systematic changes to vocal identifiers, which result in an utterance being less representative of the speaker. Four such phenomena will be discussed in the following subsections, considering within-speaker

variability as well as imprint fidelity. Such observations can help assess the plausibility of source identity when differences are observed between the compared utterances.

2.3.2.1 Contemporaneity

The comparability of utterances produced by the same speaker may be greatly reduced if a great amount of time has passed between their production. This is because the source may have altered markedly in the meantime. While voices are subject to short-term changes, for example due to changes in the speaker's physical and psychological constitution, these changes are hard to predict. Long-term changes due to ageing are more systematic, and therefore produce more predictable patterns across speakers.

Of course, the general question is to be raised whether earwitness evidence is admissible in cases where identification procedures can only take place after a long delay. This is a legal question. Contemporaneity played a role, for instance, in the infamous Hauptmann case (cf. Section 3.4.1). When Lindbergh identified Hauptmann as the perpetrator, he heard the voice at the District Attorney's office 29 months after the alleged first exposure to the voice at the crime scene (Solan & Tiersma, 2003: 373). Many works that have discussed this delay focus predominantly on the effect that it could have had on Lindbergh's memory. At the same time, however, a delay of almost two and a half years may have had an impact on Hauptmann's speech as well.

Literature on contemporaneity often make a distinction between alterations in a speaker's voice due to particular events or pathologies and "'normal' aging", caused by "the continuing progression of time" (Bowie, 2011: 29). The idea of normal ageing is potentially interesting for the assessment of earwitness testimony as it implies that there are certain regularities to the ageing of voices. This requires the lay listener to be aware of such regularities. In this connection, a study by Braun (1996) concluded that lay listeners are capable of estimating a speaker's age by means of holistic analysis with an average deviation of 6 years. While this does not answer the question whether listeners can extrapolate former or future versions of an unfamiliar speaker's voice, it provides limited evidence for listeners' awareness of ageing effects.

Several studies have described the effects of ageing from an acoustic point of view. Watt (2010: 78) observes that the anatomical structure of the vocal tract is "essentially fixed in early adulthood", except for foreign objects, such as braces or piercings. In its context, this statement is referring to anatomical changes related to growth. This information does not, however, help

explain within-speaker variation in most criminal cases, as Jessen (2012: 25) observes that the most relevant population in forensic cases is between 18 and 60 years old. Physical alterations to voices in this age range are the result of a "generalized loss of muscle tone, ossification of laryngeal cartilages, and hormonal changes" (Beck, 2010: 178). Most of these changes result in differences in mean f0 (Beck, 2010: 178). Several authors show that for female speakers, on average, f0 steadily declines from early childhood until the age of about 50 years. It then plateaus and may rise again after the age of 80 years (Baken & Orlikoff, 2000: 173-176; Jessen, 2012: 25; Schötz, 2006: 107). The relationship between f0 and age is less clear for male speakers (Beck, 2010: 178). A study by Rhodes (2012: 270) indicates that the mean f0 of men decreases between early adulthood and the age of about 40 years, with a particularly noticeable drop after the age of about 35. According to Jessen (2012: 25), it may rise again after the age of 50 and even more noticeably after the age of 70 or 80.

Irrespective of speaker sex, speech rate increases until the age of 30 and then decreases continuously (Jessen, 2012: 25-26). Rhodes' (2012) study, which is a longitudinal investigation of eight individuals' speech between ages 21 and 49 in seven-year intervals, further revealed that ageing affects vowel formants. He found that the formant values of monophthongs generally decreased over the course of the study's observation window. F1 frequencies decreased to the greatest extent; by 8.5% on average. This effect could be demonstrated for all eight speakers and was stronger for front vowels than for back vowels. Rhodes hypothesised that this phenomenon is caused by reduced flexibility of the temporomandibular joint with age (Rhodes, 2012: 271). A smaller frequency reduction (average 3.7%) was found for F2 measurements of the same vowel phonemes, followed by F3 with the smallest amount of reduction (Rhodes, 2012: 271). The effect of this phenomenon on a lay listener is unclear. Although the observed reduction in formant frequencies is supposedly of biological origin, its auditory manifestation would be in the form of changed vowel quality. This might be interpreted as a change in accent by a listener, and therefore as a cultural rather than a biological feature. At any rate, the findings suggest that ageing may lead to noticeable within-speaker variability and that this variability is somewhat systematic. It must be stressed though, that the described changes are tendencies and subject to individual variability.

2.3.2.2 Short-term contexts

Particular situations can elicit greater than usual within-speaker variability on a short-term basis. There is a multitude of contexts in which a speaker may produce less representative utterances. The following list of examples is not exhaustive and is not arranged in a particular order.

In 1911 the French otolaryngologist Etienne Lombard (1869-1920) published a paper entitled "Le signe de l'élévation de la voix", in which he reported on the observation that speakers unconsciously increase their vocal effort in loud environments (Brumm & Zollinger, 2011). The so-called *Lombard effect* was later found to apply not only to speech produced in noisy environments but also to other specific situations, e.g. to speech produced in telephone conversations (Hirson et al., 1994). This could be explained by more recent experimental findings, which suggest that Lombard speech is not a general response to ambient noise, but sensitive to the frequency composition of the noise. Stowe & Golob (2013) observed that speakers did not produce Lombard speech when competing with noise from which speech-like frequencies were removed.

The Lombard setting is louder than the speaker's modal setting. Specifically, "intensity increases about 0.38 dB for every 1.0 dB increase in noise level above 55 dB sound pressure level" (Stowe & Golob, 2013: 640). Additionally, it is characterised by a significant increase in mean f0 (Jessen et al., 2005: 177). In read speech, the Lombard setting also leads to greater variation coefficients, which cannot be observed to the same extent for spontaneous speech (Jessen et al., 2005: 203). While expert phoneticians are aware of this effect, the potential impact the on earwitness testimony is difficult to assess.

Deviations from the speaker's usual physical or psychological state can also lead to temporary vocal changes. Causes for such short-term variation include psychological stress (Kirchhübel et al., 2011), intoxication (Braun, 1995; Klingholz et al., 1988; Künzel et al., 1997), illness (Braun, 1995), tiredness or exhaustion (Whitmore & Fisher, 1996), as well as several states of emotional agitation (Rostolland, 1982). The effects of these conditions are hard to predict and are not well investigated. The exact acoustic parameters affected by psychological stress are for instance not known (Kirchhübel et al., 2011). A detailed discussion would exceed the scope of this investigation. In short, some conditions may elicit extreme within-speaker variability. Even low levels of alcohol intoxication have been shown to increase f0 standard deviation by up to 100% (Klingholz et al., 1988). Similarly, shouting is known to drastically differ from a speaker's normal speech, including a large rise in mean f0 (Blatchford & Foulkes,

2006). At the same time, different voices are assumed to become more similar to each other in the shouting register; most likely due to a reduction of f0 variability between speakers (Rostolland, 1982: 123).

In sum, it is important to keep in mind that what is true for the whole is not necessarily true for the particular (Meuwly, 2006: 206). Short windows of observation may reveal differences that are not representative of what is the norm for a speaker. Consequently, the conditions described above may have a confounding effect on individualisation tasks if the compared utterances are mismatched for these conditions. This has implications for earwitness testimony as the speech of a perpetrator is very likely to be affected by one or a combination of these factors during the criminal event, while recordings used in identification parades are usually obtained under controlled conditions.

2.3.2.3 Disguises

An important question to consider in a forensic context is whether the perpetrator deliberately disguised his or her voice, so that the speech produced during the criminal event is less representative of their vocal identity. It is self-evident that the within-speaker variation caused by disguised speech may be different from, or greater than, undeliberate within-speaker variation, making it difficult to establish that a disguised QS and a KS from the same source. Several studies confirm that disguise reduces both discrimination and recognition scores for familiar as well as for unfamiliar voices (Clifford, 1980; Hirson & Duckworth, 1993; Hollien et al., 1982; Reich, 1981; Reich & Duke, 1979).

It is hard to find reliable recent statistics on the proportion of criminal cases in which vocal disguises play a role. Data from Gfroerer (1994), reported in Masthoff (1996: 160-161), indicate that 52% of the *German Federal Criminal Police*'s voice comparison cases between 1989 and 1994 involved some form of vocal disguise, with the specific percentage for blackmail cases being even higher at 69%. On the other hand, Künzel estimated for the same institution that 15 to 20% of their annual cases in between 1980 and 2000 involved vocal disguises, which is much lower than Gfroerer's estimate while comprising the same period (Künzel, 2000: 149). Künzel's lower figure is in line with Braun's (2006) findings for private forensic voice experts in Germany. She analysed 175 cases that were conducted by 4 different experts and found that 23% of the cases featured disguised speech. Moreover, Braun could establish that voice disguises were particularly prevalent when perpetrators could expect to be recorded, due to the context of the crime (e.g. kidnapping) or when they were familiar to the victim. For the cases

of the private laboratory *JP French Associates* in the UK, significantly lower proportions of 2.5% (Clark & Foulkes, 2007: 198) and 1% (Kim, 2008: 47) were estimated for the years 2007 and 2008, respectively. It is, however, likely that voice disguises will become more common in future criminal cases, as electronic forms of disguise become more readily available (Clark & Foulkes, 2007: 196).

Two types of disguise can be differentiated, based on the perpetrator's intentions. The differences can be illustrated with the help of Figure 2: The first type is a disguise in the narrow sense, whose primary aim is the production of utterances that have a reduced qualitative identity with their source. From a strategic point of view, speakers may want to change the values of vocal identifiers, so that they are untypical for them. The speaker thus tries to hide his or her identity. The second type of disguise is mimicry, whose primary aim is to produce utterances with high qualitative identity to a different source. The speaker's strategy is to produce values for vocal identifiers that are typical for a specific different speaker. The speaker thus tries to assume a different identity. A reduced qualitative identity of the disguised utterances with the actual source may be a by-product. In other words, disguises in a narrow sense aim to hide existent continuity of the utterance with the source, while mimicry aims to suggest a non-existent continuity with an alternative source.

The distinction is important as the types of disguises cater to the expectations in legal proceedings in different ways. A disguise in the narrow sense is intended by the speaker to provide evidence for the Hd, i.e. that QS and KS are produced by different sources, rather than for the Hp that the source is identical. Mimicry, on the other hand, aims to support the Hd by discrediting the Hp's premise that the most likely suspect was chosen, i.e. by offering an alternative Hp.

Expectations are particularly important in earwitness cases, and have an impact on the success of mimicry. According to Nolan (1982), speakers may only be capable of imitating a few vocal identifiers of a target voice at a time. Additionally, an investigation by Zetterholm (2003) shows that imitators are likely to significantly overshoot or undershoot characteristics of the target voice. Nonetheless, imperfect mimicry may still be successful depending on the listener's expectations. Kreiman and Sidtis (2011: 247) give the example of the late actor David Niven, whose dialogue in the film *Curse of the Pink Panther* was overdubbed by mimic Rich Little due to Niven's battle with amyotrophic lateral sclerosis. Audiences were generally unaware of the voiceover, as the heard voice sufficiently matched their expectations. To provide a less anecdotal example, research by Schlichting and Sullivan (1997) investigated mimicry in

the specific context of voice line-ups. Swedish listeners were presented with different versions of line-ups and asked to identify the stimulus that was produced by their then Prime Minister, Carl Bildt. Along with foil voices, some of the line-ups contained a voice sample by Bildt as well as a sample produced by a mimic. In another version of the experiment, listeners were presented with a target-absent line-up that only contained the imitator's sample and the foil voices. It was observed that most listeners correctly identified Bildt when his voice was present but mistook the imitator for Bildt in the target-absent setup. Earwitnesses may therefore be especially susceptible to mimicry if they were primed by the context of the criminal event. More recent research by González Hautamäki et al (2015) showed that even automatic speaker verification systems were affected by mimicry in a similar way as primed human listeners.

Several taxonomies of disguises have been suggested in the literature, e.g. by Rodman (1998) and Neuhauser (2012: 13), that classify disguises based on their articulatory origin. In short, vocal disguises can in theory have as many dimensions as there are vocal identifiers or combinations of identifiers. Research suggests, however, that in actual cases, disguises in the narrow sense are above all manipulations of laryngeal VQ (Masthoff, 1996). In two thirds of the cases that featured disguise in Braun's (2006) study, speakers attempted to alter their phonation, most commonly by raising their mean f0. According to Künzel (2000), on the other hand, male and female speakers employ different strategies when disguising phonatory settings. He showed in an experiment that men tend to lower their f0 when asked to disguise their voice, while women tend to raise it. To this end, speakers may employ VQs beyond the modal setting, i.e. creak or falsetto, to create the auditory impression of a lower or higher average pitch, respectively. Surprisingly, these strategies are in opposition to the speakers' physical capabilities, as it was shown in the same experiment that women are capable of greater relative lowering and men capable of greater relative raising of their mean f0.

Disguise by means of VQ alterations is known to have a "devastating" effect on the performance of earwitnesses (Bull & Clifford, 1999: 121). This is particularly problematic as individuals can deliberately alter VQ to a great extent (Kreiman & Sidtis, 2011: 157). Experiments have shown that structurally simple disguises can be highly effective for hiding a speaker's identity. In a perception experiment by Wagner and Köster (1999) 20 lay listeners participated in a speaker naming task, identifying target voices from normal and falsetto samples. The recognition rate was 97% for normal speech but only 4% for falsetto samples.

Whispering is considered another effective form of voice quality disguise (Hollien et al., 1982; Künzel, 2000; Masthoff, 1996). Yarmey et al. (2001) exposed two groups of lay listeners

to speech samples of varying lengths, produced by familiar and unfamiliar speakers. One group listened to whispered versions of the stimuli, while the second group listened to versions produced in the speakers' modal voice. The whispered condition had a severe impact on recognition accuracy. However, increased sample duration and familiarity with the speaker led to a general improvement of accuracy across conditions. Smith et al (2017) tested these conclusions with a different experimental setup. Instead of exposing two experimental groups to different conditions, they exposed a single group to both the whispered and the control condition. Listeners and speakers were recruited from a pre-exiting social network of eleven women. Stimuli were of varying lengths and featured network members as well as foils. In line with Yarmey's findings, identification accuracy was above chance (64%), but the lowest for the shortest speech samples in the whispered condition.

While the disguises applied by the speakers in the above examples are systematic from an articulatory point of view, their effectiveness can be explained by the fact that the resulting change in auditory impression is not predictable. Whisper will allow the listener to make assumptions about the speaker's sex, because size and shape of the vocal tract remain unaffected (Smith et al., 2017: 7). A speaker may therefore recognise the resonant properties of a familiar vocal tract, while it is not possible to 'reverse-engineer' an unfamiliar speaker's normal voice from a whispered or falsetto utterance.

Speakers may employ more complex disguises that extend to cultural properties of voice; for instance, by manipulating accent features. Neuhauser (2011, 2012) investigated the ability of different listener groups to tell apart authentic and imitated French accents in German. From an articulatory point of view, she observed that German speakers who tried to imitate a French accent were generally inconsistent in their productions and significantly exaggerated some of the foreign accent's features. It is because of these inconsistencies that imitations did not fool most expert listeners. At the same time, the imitations included great phonetic detail, and picked up on several features of authentic French accents, including reductions in the voice onset time of plosives, h-dropping, and the absence of word-final syllable reductions. Lay listeners were generally not good at telling apart authentic and fake accents. Interestingly, lay listeners tended to rate the imitated accents as being more representative of a French accent than the authentic accents. This might be explained by the fact that imitators and listeners belonged to the same language community and therefore also shared conceptual stereotypes (Beinhoff, 2013: 21) of the imitated language.

Accent imitations are blurring the line between disguise in a narrow sense and mimicry as they could be interpreted as mimicry of a speaker group as opposed to an individual speaker. If the criterion of speaker intention is applied to tell these types apart, a categorisation would depend on whether the accent imitation was primarily produced to hide the speaker's identity or to also have an impact on the direction of the investigation.

2.3.2.4 Fidelity of the imprint

As discussed earlier, the imprint condition is unavoidable for all forms of speaker comparison. These imprints have a certain fidelity in terms of how true they are to the original. Speech material can therefore be less representative of the source due to low imprint fidelity. In the specific case of earwitness testimony, the imprint formed at the crime scene (QS) is a memory imprint, while the imprints provided in e.g. a VP (KS) are recordings. The QS is of greater interest in this connection as the KS is usually a reasonably high fidelity digital recording.

The imprinting process during the criminal event may involve several filters that may alter the vocal information before it is imprinted in the listener's mind. For instance, the perpetrator's voice may have been heard over the telephone, which has become an increasingly common scenario in earwitness cases (Künzel, 1994: 136). Likewise, telephone transmission plays an important role in voice comparisons conducted by experts, in which recordings exist for all compared utterances. Künzel (1994: 136) reported that more than 95% of the BKA's voice comparison cases involved telephone transmission. The landline bandwidth in most European countries is delimited to frequencies between 340 and 3700 Hz (McDougall et al., 2015). Consequently, fricatives and affricates are most affected by this limitation since most of their energy is concentrated above this range. Additionally, background noise may be present in the on the line (Kreiman & Sidtis, 2011: 250). Note that the perception of f0 is not affected since as this information is deduced from harmonic spacing (Kreiman & Sidtis, 2011: 250, Fn 6). Observed f0 may still be higher for unrelated reasons, e.g. the Lombard effect. F1 values are shifted upwards by up to 13% in landline transmission (Künzel, 2001). Byrne and Foulkes (2004) observed an even greater raising of F1 in mobile phone transmissions, by an average of 29%, but up to 60%. F2 values are, on average, largely unaffected in either form of transmission.

Early studies suggest that voice recognition is significantly impacted by the effect of telephone transmission due to the known impact of high-pass filtering (Compton, 1963; Rathborn et al., 1981). More modern studies could not confirm this (Kerstholt et al., 2006; Perfect et al., 2002), which Kreiman and Sidtis attribute to improvements in telephone

technology (Kreiman & Sidtis, 2011: 250). Modern standards based on "voice over IP" technologies have the potential to produce bandwidth ranges of 50 – 14000 Hz (superwideband) or 20 – 20000 Hz (fullband) (Cox et al., 2009: 106). On the other hand, an experiment by Fenn et al. (2011) showed that listeners generally did not notice when interlocutors were exchanged halfway through a telephone conversation, indicating that identity perception over the telephone is indeed limited.

Further filters may distort the perpetrator's voice, e.g. garments worn over the face, which is a relevant consideration in masked robberies. Fecher (2014) showed that unfamiliar speaker recognition was compromised when speakers wore facewear. It is hypothesised that this effect is mainly due to energy absorption characteristics of the facewear material, as well as from the facewear's impact on initiation (breathing), and articulation (exterior pressure to the mouth/cheeks) (Fecher, 2014: 245). Speaker-specific information related to plosive bursts may be particularly affected by the wearing of certain garments, impacting speech recognition as well as speaker recognition (Fecher, 2014; Fecher & Watt, 2013). The former phenomenon may contribute to the latter in that a greater proportion of the listener's cognitive capacity may be required to decode what is being said, leaving less capacity for speaker identity processing.

2.4 Conclusion: The problems of speaker individualisation

Speech is a multifaceted phenomenon shaped by a speaker's physical traits, socialisation, and idiosyncrasies. Voices are therefore high-dimensional, requiring numerous descriptors to characterise them fully.

This chapter presented a summative framework to represent vocal identity, using individual utterances as the basic unit of analysis. A core assumption was that each utterance has unique qualities distinguishing it even from other utterances by the same speaker. Thus, earwitnesses face a difficult task: determining if two utterances sound different due to distinct speakers or the same speaker's variability. This challenge grows when utterances are compared out of context, as with a suspect's voice from an interview and a perpetrator's voice during a crime. Establishing continuity between these decontextualized "snapshots" proves challenging. Moreover, vocal variability itself is complex, occurring at organic, cultural, and habitual levels.

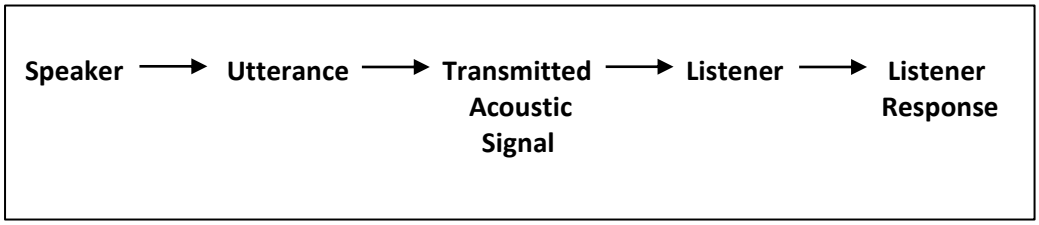
Several common-sense beliefs about individualisation by earwitnesses could be formulated based on the relationship between trace and source. Most importantly, the task should be easier the more familiar the speaker is to the listener, and the longer the compared

utterances are. A familiar listener may know that an utterance is plausibly produced by a particular speaker although the utterance is not very representative of that speaker, and although the utterance to which it is compared is substantially dissimilar. At the same time, a listener who was exposed to a longer utterance will have been exposed to more within-speaker variability. If a broader definition of source is applied, i.e. in the form of social groups, listeners may have an advantage if they share group membership with the source(s).

3. The earwitness as source of information

The previous chapter dealt with the first three links of the "speech chain" (Figure 1), which are all related to the production and transmission of speech. It was demonstrated that speech samples from three categories of speakers are relevant in the context of speaker individualisation: the perpetrator, the suspect (who may be identical with the perpetrator), and members of relevant reference populations. It was also shown that the main source of information, the individual utterance, can be described by means of a multitude of interdependent and context-dependent dimensions. Individual utterances are therefore only revealing a specific proportion of a speaker's vocal identity. Their preservation in the form of recordings or memory imprints introduces further variability.

Figure 3: The "speech chain" revisited (Kreiman, 1997: 87)



The present chapter focuses on the last two links of the speech chain, i.e. on the earwitness who has heard the voice of a perpetrator and might be asked to give individualisation evidence about that voice during legal proceedings. Depending on the jurisdiction, the witness may have been presented with further voices from the other two categories of speakers, i.e. the suspect and the reference population, within the framework of a standardised procedure, such as a VP. In such procedures, the earwitness is a conduit for the questioned sample, i.e. the perpetrator's voice, bearing its imprint. At the same time, the witness is the individual performing the comparison with other voices for the purpose of speaker individualisation.

It is the aim of this chapter to characterise earwitness testimony as a source of information in legal proceedings. This entails the expectations of the legal system, procedures used for eliciting testimony, the variables that have a bearing on the quality of the testimony, and potential ways of improving individualisation evidence.

Note that for pragmatic reasons, the discussion of legal aspects is restricted to UK jurisdictions, while, occasionally, examples from other Common Law (CL) jurisdictions are provided.

3.1 The earwitness's task – terminological challenges

The term "identification" is widely used in the linguistic, legal, as well as psychological literature, albeit referring to a wide array of concepts within and between disciplines. The lack of clear and consistent terminology has been criticised by several authors (e.g. Plante-Hébert et al., 2021; Ramon & Gobbini, 2018). Tasks commonly referred to as "identification" are diverse, and span various phenomena that could more precisely be referred to as e.g. "individualisation", "recognition", "discrimination", or "diarisation", depending on the task at hand. In the following section, a clear terminology will be established for the study of (ear)witness testimony.

It has already been established in Chapter 2 that the concept which most authors have in mind when using the term "identification" is individualisation. Identification is a concept in forensic sciences as well, but it refers to the act of establishing that something is one and the same numerical identity (Kirk, 1963: 236). Identification is, for instance, required to show that the chain of evidence was not broken, by establishing that an object found at the crime scene is the same object that was later investigated by a forensic scientist or presented in court (Meuwly, 2006: 207). Individualisation on the other hand is the act of linking an object to its individual source; i.e. establishing source identity. In the present thesis, the term "individualisation" is used exclusively to refer to this concept. Note, however, that most theoretical claims about individualisation tasks made in this section are derived from reference works in which the term "identification" is used by the author(s); this applies to (Hammersley & Read, 1996; Lavan et al., 2019; Plante-Hébert et al., 2021; Ramon & Gobbini, 2018; Rose, 2002; Schweinberger & Zäske, 2018).

Large parts of the discussion in Chapter 2 juxtaposed the two tasks within FSS that involve a comparison of speech samples from the perpetrator and a suspect, i.e. earwitness testimony and forensic speech comparisons performed by experts. In the past, various terms have been used to refer to the latter task, which all contain the term "identification" (in the meaning of individualisation), e.g. "speaker identification", "speech identification", or "voice identification". These terms are not used anymore because the identity of the source cannot be established with absolute certainty by means of a methodology that is to a large extent based on inductive reasoning (cf. Section 2.1.2). Rose rightly points out that the actual task is one of discrimination, i.e. the act of differentiating between two samples (Rose, 2002: 9). Individualisation is only a secondary result (Rose, 2002: 9). This secondary result is inferred when the observations made in the discrimination task are analysed before the background of

two hypotheses about the numerical identity of the source (the prosecution hypothesis and the defence hypothesis).

Experts as well as earwitnesses perform a type of comparison in their respective roles. However, the conditions of these comparisons differ. An expert performing a forensic speech comparison is provided with speech material from the perpetrator and the suspect, and can compare them at the same time. A witness, on the other hand, has heard the perpetrator's voice in the past and compares the memory of that voice to a newly presented recording of the suspect's voice. The earwitness's task therefore additionally entails an aspect of "recognition", which is the act of establishing that a voice has been heard before (Ramon et al., 2019: 468). Consequently, there has been a prior personal experience with the perpetrator's voice in a different context, potentially involving some sort of interaction with that voice, e.g. engagement in a conversation (Hammersley & Read, 1985). However, in most cases, voice recognition by the witness is not automatically followed by individualisation. This is because the witness cannot necessarily link the heard voice(s) to a particular source identity, as this would require personal familiarity with the speaker or further information about the speaker, such as being able to name the speaker or having seen his or her face. Individualisation therefore requires the same subprocesses as recognition, and additionally depends on the ability to provide semantic information about the source (Ramon et al., 2019: 468).

In most earwitness cases, the witness performs a recognition task with insufficient sematic information to individualise the speaker. However, this recognition task is context-specific in that the witness is not asked whether they recognise the voice which produced the KS in general, but whether they recognise the voice from a specific context, the crime scene. It goes without saying that this contextual delimitation may cause inferences whenever the KS sounds familiar to the witness for reasons unrelated to the criminal event.

Several authors have pointed out that there is a need to differentiate between individualisation and recognition in witness research (e.g. Hammersley & Read, 1996; Plante-Hébert et al., 2021; Ramon et al., 2019; Ramon & Gobbini, 2018; Schweinberger & Zäske, 2018). It can be demonstrated experimentally that these tasks are of a different nature and that recognition in fact precedes individualisation. Listeners may, for instance, recognise a personally familiar voice, without being able to retrieve identity-specific semantic information, or only being able to do so after a delay (Hanley et al., 1998; Hanley & Turner, 2000). This phenomenon is colloquially known as 'not being able to place someone'. Interestingly,

experiments suggest that such familiar-only experiences are more prevalent for voices than for faces (Hanley & Turner, 2000).

In summary, a hierarchy can be described in which individualisation is dependent on recognition, which is in turn dependent on discrimination (Figure 4). While individualisation is the aim of forensic investigations, the frame conditions of earwitness cases do not necessarily allow the witness to individualise the source, so that in most cases the task can be characterised as 'context-specific voice recognition'. Individualisation evidence provided by expert witnesses, who were not exposed to the QS during the criminal event, is technically a mere discrimination of samples, whereby a likelihood of source identity is inferred before the background of explanatory hypotheses.

Figure 4: A hierarchy of different voice processing tasks

Task:	**Description:**	**Performed by:**
Individualisation	Providing semantic information that links the sample to a specific source identity	Some earwitnesses, esp. earwitnesses familiar with the perpetrator
↓ *Depends on*		
Recognition	Establishing that a voice was heard before (potentially in a particular context)	Earwitnesses
↓ *Depends on*		
Discrimination	Assessing similarities and differences between samples by means of comparison with the aim of inferring the number of sources	Earwitnesses and expert witnesses

For the sake of completeness, some other terms will be addressed that can be found in the literature on earwitnesses. The term "voice processing" is generally used to refer to all kinds of cognitive responses to voices and is thus an umbrella term for the aforementioned concepts. A common but confusing term found in the literature is "matching", which is above all used in

psychometric tests, such as the *Bangor Voice Matching Test* (Mühl et al., 2018). The term is usually referring to a task in which participants are exposed to multiple stimuli, either simultaneously or in very short succession, and asked to assess whether the stimuli stem from the same source. Usually, this sort of task involves unfamiliar stimuli (Schweinberger & Zäske, 2018: 539). It is thus about an immediate response, i.e. a discrimination task or a form of recognition that only requires minimal memorisation effort. At the same time, it has already been shown in Section 2.1 that the term "match" is highly controversial in a forensic context. It is, however, widely used for this sort of task in the literature.

Another potentially ambiguous term is "voice memorisation" or, conversely, "voice recollection/recall". This concept can, for example, be found in the names of the *Glasgow Voice Memory Test* (Aglieri et al., 2017) and the *Jena Voice Learning and Memory Test* (Humble et al., 2022), which are both recognition tasks, while the JVMT taps into long-term memory to a greater extent. Nonetheless, all recognition tasks and individualisation tasks involve memory to some extent. It must therefore always be defined what sort of memory is addressed by a voice memory test. Moreover, it should be clear that when source identity is of interest, the task of 'voice recollection/recall' may apply to both recognition and individualisation tasks.

3.2 Lay and expert listeners

3.2.1 Types of testimony

Earwitnesses testify in court when no recording of the QS is available (McDougall, 2021: 33). In Common Law (CL) jurisdictions, witness testimony is subject to the 'opinion rule', which states that witnesses should report what they have observed with one of their five senses and not draw conclusions from these observations themselves (Robertson et al., 2016: 161f.). Inferences from the witness's observations are considered an 'opinion' and are within purview of the trier of fact. Consequently, the information provided by a witness is supposed to have the character of 'facts' or 'raw data' (Robertson et al., 2016: 162).

Considering the discussion about the nature of vocal traces in Chapter 2, it is, however, apparent that the information an earwitness can provide will never have a factual character, even if the witness intends to merely report on observations. For one thing, it was shown that voices can be described in an exceedingly large number of dimensions, which means that the observed number and combination of dimensions will have an impact on the observable similarity between samples. While an observation about two voices sharing the same pitch may be factually correct, substantial differences between the voices may emerge when more

dimensions are considered. Moreover, a witness must verbalise his or her observations to share them with the court. It is thus important to consider whether the witness has command of adequate vocabulary for this intent and whether a similar vocabulary is shared by a majority of phonetically untrained members of the public.

Whenever a recording of the QS is available, witness testimony is not required because the trier of fact can directly draw conclusions from their own observations of the trace (Robertson et al., 2016: 162). The described limitations of a lay individual's ability to observe and describe voices do not, however, only apply to earwitnesses, but also to judges and jury members, who do not usually possess linguistic expertise either. In these cases, one of the parties can call on an 'expert witness' to provide evidence, when the skill to assess the facts is assumed to be greater with the expert than with the trier of fact (Basu et al., 2022: 1; Robertson et al., 2016: 162). Crucially, expert evidence is exempt from the opinion rule (Robertson et al., 2016: 161). This means that while a witness is only allowed to state an observation *E*, the expert witness is allowed to discuss the probability of their observation in the light of the prosecution hypothesis and the defence hypothesis, i.e. $P(E|Hp)$, and $P(E|Hd)$. Robertson et al. (2016: 162) rightly point out that de facto, the expert will inevitably express $P(E|Hp, K)$ and $P(E|Hd, K)$, where *K* denotes the expert's knowledge and experience.

The VP procedure does not unambiguously belong to either category of testimony because it involves earwitnesses as well as experts.[6] It is therefore important to characterise the sort of testimony elicited by the procedure in more detail. To this end, a scenario may be considered in which the individualisation evidence provided by an earwitness was not elicited by means of a VP. For instance, an earwitness may hear the voice of a suspect during a trial or a police investigation and claim to recognise the voice from the crime scene. It is unequivocal that such evidence – if considered by the court – would be a case of earwitness testimony and not expert witness testimony. In cases where VPs are conducted, the expert who constructed the parade will give expert evidence about the function of the parade, in addition to the outcome of the parade (Nolan, 2003). This measure may be interpreted as an effort to comply with the opinion rule, as the earwitness's recognition of the voice is already a conclusion, whereas the expert may express an opinion about how this conclusion should be treated.

The VP is a structured framework for the elicitation of individualisation evidence in the form of a test. It contains samples from the suspect as well as distractor voices which are

[6] Note that the Home Office guidelines that outline a procedure for VPs in England and Wales (Home Office, 2003: n.p.), specifically refer to the person compiling the parade as an "expert witness".

selected by an expert. Consequently, the factor *K* introduced in the formula above (the expert's knowledge and experience) also plays a role in VP evidence. What sets the VP apart from unguided individualisation evidence is that the outcome of a VP provides the trier of fact with a likelihood based on the ratio of KS to foil voices in the parade. This means that the individualisation evidence which is produced when the witness selects the KS is a less likely scenario than the witness failing the parade, assuming that the witness is merely guessing. In contrast to an expert opinion, however, this likelihood is only an *a priori* likelihood, i.e. the prior odds are stacked in favour of the Hd as there are several foils and only one recording from the suspect (KS). An expert opinion, on the other hand, provides a probability that is based on prior odds as well as the expert's assessment of the evidence's strength, i.e. a 'posterior probability' (Rose, 2002). In other words, if the earwitness picks the suspect's voice in a VP, the trier of fact only knows that this was the less likely scenario to begin with, while they cannot assess the likelihood of the witness's general voice recognition capabilities or the witness's ability to derive correct conclusions from observations. Moreover, the judge or jury must rely on the expert's ability to correctly manipulate the prior odds by choosing adequate foil voices.

In summary, the VP is a way of eliciting individualisation evidence from earwitnesses, using an expert's knowledge to create a valid test for the witness's ability to individualise the suspect's voice. If the parade is valid (and not biased) it maximises the likelihood of an individualisation being the product of the witness's ability (Cutler & Wells, n.d.: 115). Validity is ideally ensured by the involvement of an expert. While the outcome is subject to a likelihood, this form of prior odds manipulation must not be confused with the posterior probability that characterises an expert witness's opinion.

3.2.2 Conditions and categorical differences

Since expert testimony is the preferred option when a recorded QS is available (Künzel, 1994: 137), earwitnesses and expert witnesses are often compared regarding the quality of their testimony (e.g. Elaad et al., 1998; Schiller & Köster, 1998; Shirt, 1984). Such a comparison is, however, of limited informative value given that both types of testimony are born of different circumstances. There is no real-world scenario in which a witness's expert status is the only dividing factor between lay and expert testimony, while all other factors are equal. In the absence of a recorded QS, the account of a lay listener who witnessed the crime, and their participation in a VP, may well be the only viable way of obtaining individualisation evidence. The main differences between expert witnesses and earwitnesses are summarised in Table 1. Individual aspects will be discussed in the following subsections.

Table 1: Comparison of earwitness and expert witness testimony

	Expert witness (phonetician)	Earwitness (lay listener)
Conditions:		
- Witness to the criminal event	no	yes
- Familiarity with perpetrator	no	potentially
- QS	recorded	memorised
- KS	recorded	recorded
- Potential outcome	discrimination	recognition or individualisation
Observations:		
- Trained	yes	no (but not necessarily inexperienced!)
- Instrumental measurements	yes	no
- Analysed identifiers/ dimensions	acoustic and auditory overt and covert	auditory overt
- Quantitative assessment	continuous and discrete	discrete
- Qualitative assessment	instrumental and impressionistic	impressionistic
- Analytical reasoning	featural	holistic?
Objective reliability:		
- Method	observable, explained	'black box'
- Reasoning	transparent	'black box'
- Expertise	professional reputation/ recognised methodology	ad hoc expert
- Weight of evidence	posterior probability	prior odds of VP and self-assessed certainty
- Communication	professional vocabulary	explanatory gap?

It was already shown in Section 3.1 that an earwitness who is familiar with the perpetrator is the type of witness most likely to individualise the perpetrator, while most other earwitnesses are limited to recognition, and expert witnesses limited to a discrimination of the available samples. Nonetheless, expert witnesses may be generally perceived as the more credible type of witness. Note that the legal understanding of 'witness credibility' is not restricted to a witness's truthfulness, but also entails the "objective reliability of the witness", i.e. the "ability to observe or remember facts and events" about which they give evidence (*Thornton v Northern Ireland Housing Executive* [2010]: §12).

In the legal systems of the UK, the admissibility of expert witness testimony is strictly regulated by Part 19 of the UK *Criminal Procedure Rules* (CrimPR; The Criminal Procedure Rules, 2020), which lay out the rules for criminal proceedings in magistrates' courts, the Crown Court, the Court of Appeal, and certain High Court cases. The guidelines state that the expert's report must:

- explain the expert's qualification [CrimPR 19.4, a]
- provide information about the consulted reference literature [CrimPR 19.4, b]
- lay out the facts based on which an opinion was formed [CrimPR 19.4, c]

These rules extend to further persons who may have been consulted by the expert or contributed to the expert's testimony [CrimPR 19.4, e]. Moreover, an expert can only give opinion within their established area of expertise [CrimPR 19.4, 1, a, ii]. Similar guidelines are in place in other CL jurisdictions. In the US, for instance, the so-called Daubert standard applies to expert testimony (Robertson et al., 2016: 169f.), for which *Daubert v Merrell Dow Pharmaceuticals Inc* (1993) set the precedent.[7] The five main criteria of the standard are (Robertson et al., 2016: 171):

- whether the theory or technique that the expert uses can be, and has been, tested
- whether the technique has been published or subjected to peer review
- whether actual or potential error rates have been considered
- whether the technique is widely accepted within the relevant scientific community
- whether standards and procedures are maintained.

[7] The Daubert standard has largely replaced the *Frye* standard (Frye v United States (1923)), which stated that the method used by an expert witness "must be sufficiently established to have gained general acceptance in the particular field in which it belongs".

Note that these criteria are not a definite checklist, but rather a point of reference for the trier of fact to help assess the reliability of the expert testimony. None of the criteria is therefore "determinative of admissibility" (Robertson et al., 2016: 171).

In either system there is consequently a need for demonstrating the 'objective reliability' of expert testimony, although the "Daubert standard" pays greater attention to the method or technique used by the expert, while the CrimPR pay greater attention to the reputation of the expert witness and the transparency of their line of argumentation, irrespective of the used method. The last point, i.e. transparency, is in line with the opinion rule (cf. Section 3.2.1) according to which inferences from observed facts should ideally be made by the trier of fact. A trier of fact should be able to reconstruct every step in the expert's line of argumentation (Künzel, 1994: 137). Consequently, the expert's approach is necessarily based on a componential or featural approach (Jessen, 2012: 46) in which the different dimensions of the samples (discussed in Chapter 2) are analysed separately and judged against a reference population.

In contrast, an earwitness's objective reliability is hard to assess. Every lay listener does, of course, have a certain experience in the discrimination and recognition of voices, as this is a task performed by listeners on a daily basis. They are, however, not experts in this domain by virtue of reputation or training, nor can they demonstrate that their conclusions are based on a transparent and/or established methodology, whose reliability – or fallibility for that matter – can be assessed. Thus, expert testimony entails a type of 'diagnostic information' for the trier of fact, which lay witness evidence does not possess. The lack of diagnostic information is partially due to substantial gaps in the research literature (Cutler & Wells, 2009: 106). According to Cutler & Wells (2009: 106), the missing diagnostic information for lay individualisation testimony entails three types of information:

- base-rate information on the accuracy of (ear)witnesses in general[8]
- knowledge of the differences in accuracy between individual (ear)witnesses
- specific parameters of the encounter with the perpetrator.

The first and second point on this list are addressed by the experiments conducted in the present thesis, which therefore help define the 'black box' that currently is the individual witness's objective reliability.

[8] Cutler & Wells (2009) discuss eyewitness testimony, but their point applies to other types of lay testimony as well.

3.2.3 The effect of training on ability

The empirical part of this thesis will not make the exact type of reasoning employed by earwitnesses more transparent, i.e. earwitnesses will remain a 'black box' in that internal processes are not observed, and neither will participants be asked to describe their reasoning. However, diagnostic information will be gathered empirically by observing the conditional accuracy of lay listener judgements in scenarios where the identity of compared speakers is known to the experimenter. In other words, the product of recognition will be observed rather than the underlying processes. If such diagnostic information can be established by means of testing, information about individual earwitness ability will become more transparent to the trier of fact, and the testimony more or less credible depending on the performance of the witness in the test. Hence, the trier of fact will be provided with information that helps determine the weight of the evidence.

While the internal processes of lay speaker individualisation are not actively investigated in this thesis, it is important nonetheless to form hypotheses about the nature of lay voice processing as opposed to professional voice processing, i.e. about the general relationship between inherent ability and phonetic training. Better understanding of this connection has implications for expert as well as lay testimony. The question of interest is whether ability (particularly accuracy) changes as a result of training, and whether lay listeners are therefore at a substantial disadvantage compared to the expert. The competing hypothesis would be that the expert's advantage is not a result of training but rather of advantageous circumstances.

3.2.3.1 Level of detail

These advantageous circumstances comprise a controlled and repeatable exposure to the QS. Moreover, the expert has access to special equipment allowing them to perform acoustic measurements in addition to the auditory analysis (cf. Table 1). Earwitnesses, on the other hand, are limited to their sensory impressions. This means that an expert's quantitative reasoning is in part based on continuous measurements, while an earwitness can only make discrete quantitative observations. For instance, an expert witness may analyse the compared samples' f0 by means of measured Hertz values on a continuous scale, while a lay listener will assess pitch by means of discrete levels, such as 'high', 'medium', and 'low'. Note that some features of voice will also be analysed by means of discrete quantification by experts. This applies above all to features that are difficult to measure and are primarily assessed by means of auditory analysis, such as the assessment of voice quality features in the creation of a vocal profile. It

further applies to situations where the technical quality of materials is too poor to allow confident measurement.

In further support of the hypothesis that experts have an advantage at individualisation tasks by virtue of training, some authors speculate that the phonetically trained listener will pick up on more dimensions than lay listeners on an auditory basis (Rose, 2002: 310). This presupposes that vocal identifiers have different degrees of saliency, whereby only a few 'overt' dimensions are accessible to the lay ear, while an expert with a structural knowledge of speech can pick up on additional, 'covert' identifiers (cf. Table 1).

3.2.3.2 Abstraction

There is also limited evidence that phonetically trained listeners can infer some characteristics of an original speech signal when being confronted with a low-quality imprint. Lawrence et al. (2009), for instance, observed that phonetically trained listeners were able to compensate for the 'telephone effect' (Künzel, 2001; cf. Section 2.3.2.4) when judging the quality of the vowels /i/, /u/, and /æ/ in speech samples produced by young male speakers of SSBE, sourced from DyViS (Nolan et al., 2009). Stimuli were presented in two different conditions, i.e. in studio quality or alternatively through a landline bandwidth filter. It was found that the individual phoneticians' placement of a speaker's productions on the vowel quadrilateral did not differ substantially across conditions, indicating that the listeners compensated for the raised F1 values resulting from the telephone effect. The experiment did not, however, juxtapose lay and expert listeners' perception as the setup of the experiment requires a working understanding of the vowel quadrilateral on the part of the listener. It is therefore not clear whether the ability to compensate for the telephone effect is a result of training. Moreover, Lawrence et al. observed that irrespective of presentation condition, vowel quality judgements differed significantly between experts. This raises the question whether these differences are the result of different types of training and experience or whether the same type of training elicits different kinds of expertise across individuals. If training has a different effect on the auditory-perceptual skills of different experts, this further raises the question whether the lack of diagnostic information about the individual's capacity is a problem that applies to experts as well as to earwitnesses.

3.2.3.3 Perceptual processes

Apart from the number and accessibility of observable dimensions, it is of interest to a trier of fact to what extent a lay listener's perceptual process differs from an expert's feature-based reasoning, which is characteristic of the auditory-acoustic approach (Jessen, 2012: 44). After all, a feature-based approach renders evidence more transparent to a trier of fact and allows for a reconstruction of individual steps in the analysis. However, it has already been shown in Chapter 2 that the dimensions in which speech is perceived and described are likely to differ between lay and expert listeners, at least when the latter group is acting in their official capacity as expert witnesses. In this connection, the perceptual dimensions which are used in some disciplines to describe voices from a lay listener's point of view, such as 'timbre', were shown to be meta-dimensions, in that they comprise several independently describable and measurable dimensions. This implies that researchers in some fields assume that lay voice perception operates on a holistic basis, which applies above all to the field of psychology, where these vocal dimensions are widely used.

A similar assumption can be found in the forensic phonetic literature. Jessen (2012: 44), for instance, refers to the expert's deconstruction of speech into individual identifiers as an "analytical" procedure and juxtaposes this type of processing with holistic processing. He identifies the ability to apply analytical processing in both the acoustic and the auditory domain to be a direct consequence of (phonetic) training. He further characterises the analytical approach as a scientific approach, based on the premise that the division of natural phenomena into individually observable and interpretable features is a key aspect of scientific reasoning. Thus, a picture emerges in Jessen's line of argumentation, in which a holistic approach to voice processing is necessarily a lay approach in that it is not scientific and not the consequence of training. His main point is that a holistic approach is therefore not suitable for use by a forensic voice expert. At the same time, his characterisation of the processing that is actually employed by lay listeners is inconsistent. On the one hand, he suggests that the general human capability to recognise other individuals by their voice is based on holistic processing and rooted in biology (Jessen, 2012: 45). On the other hand, he points out that lay listeners necessarily employ a form of featural analysis when they not only answer the question whether two samples were produced by the same speaker, but can also answer the question why this is the case (Jessen, 2012: 46). This is because the provided reasons must be based on vocal features of some sort. Lay listeners might thus be capable of featural voice processing when prompted to do so; potentially also being able to isolate individual vocal features retrospectively from a memory impression that was originally formed holistically.

A prominent paradigm for explaining the processes underlying lay voice processing that goes beyond holistic perception is derived from 'Gestalt psychology' (Wertheimer, 1925). Gestalt psychology is a branch of cognitive psychology and therefore assumes that sensory impressions such as "seeing, hearing, remembering are all acts of construction" (Neisser, 1967: 10). This constructivist understanding of cognitive processes opposes the ideas of direct realism, which consider the perceiver to be the mere recipient of the information conveyed by perceived objects. The 'whole' is an important concept in this paradigm. It is assumed to be the object of perception and to consist of individual parts (Guberman, 2017: 7). The central hypothesis is that an organised whole is perceived to be more than (or different from) the sum of its parts (Guberman, 2017: 2).

Nolan suggests that lay speaker individualisation may indeed work on a Gestalt basis, and in fact require Gestalt reasoning, since the "overall character" of a voice may not become apparent when individual components are analysed separately (Nolan, 2005: 400). He draws an analogy to the perception of faces and notes that similarities or differences between faces become instantly apparent when entire faces are juxtaposed, while a comparison of individual features may not be conclusive. This argumentation foreshadows core ideas of the *auditory face model* (Belin et al., 2004), which presumes several similarities between voice and face processing and will be discussed in more detail in Chapter 4.

Recent adaptations of the Gestalt concept in voice research are, however, only loosely based on Wertheimer's original definition, resulting in an incoherent use of the concept. Kreiman & Sidtis (2011: 158), for instance, define Gestalt recognition as "apperception as a whole". In this case, Gestalt recognition would not be different from the kind of holistic processing described in Jessen (2012), i.e. the combined perception of all vocal features as a single entity. Based on this equation of Gestalt and holism, they create a voice perception framework which they summarise as an "interplay between 'featural' processing and 'Gestalt' pattern recognition" (Kreiman & Sidtis, 2011: 158). The same framework was also adopted in forensic phonetic studies (e.g. by Cambier-Langeveld et al. (2014), who explored ways of including holistic processes in forensic speech comparisons). Yet, 'perception' is defined as "a process of reconciliation of the interpretations of the whole and of the parts" (Guberman, 2017: 7) in Wertheimer's original Gestalt theory. The framework suggested in Kreiman & Sidtis (2011) is therefore de facto very similar to Wertheimer's (1925) idea of Gestalt processing, which he summarised as follows:

> "There are contexts in which it is not what happens in the whole that is derived from the nature of the individual pieces and how they come together, but rather the other way around, where - in the concise case - what happens to a part of this whole is determined by the internal structural laws of its entirety" (Wertheimer, 1925: 42).[9]

This shows that, rather than being reduced to holism, Gestalt processing takes the whole as a starting point, and assumes that features (parts) are perceived differently when embedded in the whole. Now, the question for any type of voice comparison is whether the voice as an organised whole creates a framework in which its individual parts are perceived differently than when assessed in isolation. If this is the case, it may have implications for lay and expert listeners, even if they are intending to isolate features from the entirety of an observed voice.

A similar question, which – to my knowledge – has not been addressed in detail in the research literature, is related to the level of observation at which Gestalt processing of voices takes place. The question is: What is to be regarded as 'the whole' when voices are the perceived objects? Most authors who mention 'Gestalt' in relation to speaker individualisation tacitly assume the whole to be the material at hand. In the case of a forensic voice comparison or an earwitness case, this is the observed/memorised sample material (i.e. QS(s) and KS(s)). In turn, the respective parts of the whole would be the individual vocal features. Such an interpretation is plausible if a voice is assumed to be a pattern and if it is further assumed that this pattern can be found in all of a speaker's utterances. Speaker individualisation would in this case be the act of recognising the same Gestalt in different samples. The individual's whole vocal Gestalt would thus be observable through each individual utterance.

However, a conceptual framework was introduced in Section 2.1.2 for the purpose of voice individualisations, in which voices are described as the sum of individual utterances, rather than a pattern found in individual utterances. Such a summative conceptual framework stresses the diversity of vocal identities, which comprise a plethora of utterances produced in different contexts, at different points in time, and via different modes of transmission. Diversity is characteristic of forensic conditions, in which the compared samples are usually analysed outside of their original contexts. If this framework is applied, Gestalt processing would be required to occur on multiple levels. While, at first, the individual compared utterances are

[9] Translated by the author of this thesis from German: „Es gibt Zusammenhänge, bei denen nicht, was im Ganzen geschieht, sich daraus herleitet, wie die einzelnen Stücke sind und sich zusammensetzen, sondern umgekehrt, wo - im prägnanten Fall - sich das, was an einem Teil dieses Ganzen geschieht, bestimmt von inneren Strukturgesetzen dieses seines Ganzen." (ibid.)

entities consisting of smaller parts, the utterances themselves are parts of a more abstract 'whole', defined as the sum of a speaker's utterances. This whole was termed 'vocal identity' in the framework and cannot be observed. Consequently, the listener may not be able to apprehend 'the whole' vocal identity, but only arrive at an approximation by means of exposure to a growing number of utterances. This means that speaker individualisation would – following Guberman's (2017: 7) summary of the Gestalt perception process – require the listener to reconcile interpretations of individual observed utterances (parts) with interpretations of an unobservable/inferred vocal identity (whole).

If Gestalt processing is applied to speaker individualisation, listeners would consequently have to fathom a Gestalt that plausibly reconciles the QS and the KS. It is possible that expert listeners have an advantage at this task as their experience would help them understand the possible transformations of the same vocal Gestalt across different situations. For instance, listening to two qualitatively different samples of the same speaker outside of the original context may lead to a situation in which the listener cannot establish continuity between both samples. An expert, who was trained on mock cases (with a known solution) may on the other hand be able to imagine a 'missing link' between the samples.

Cambier-Langeveld et al. (2014) suggest an interesting method for including Gestalt processing in forensic voice comparisons. The idea is that the expert, who has not had any exposure to the material that is to be compared, is provided with a line-up that was prepared by a colleague. In terms of material, the QS and KS are dissected into 'snippets'. Foil snippets from further speakers are added as well, following similar criteria that would be used in a VP (Cambier-Langeveld et al., 2014: 21). The 'blind' expert essentially performs a sorting task in which they try to assess the number of speaker identities across the recordings by grouping them. Similar sorting tasks are commonplace in voice processing research (e.g. Johnson et al., 2020; Lavan et al., 2022). The expert then explains why the recordings were grouped in this particular manner, the crucial question being whether the QS and KS are attributed to the same identity. Cambier-Langeveld et al. (2014) use the concepts of 'Gestalt processing' and 'holistic processing' synonymously. They consequently purport the main advantage of their procedure to be the expert's ability to analyse both features and holistic impressions; they assume the latter to above all play a role in voice quality processing. At the same time, the procedure allows for Gestalt processing on the level of observation described above, where utterances are the parts of an inferred whole. The expert using this method must ask the question whether the utterances can be plausibly interpreted as parts of a common whole.

In sum, this section showed that experts necessarily rely on featural analyses since their analysis needs to be transparent and reproducible. Lay listeners may to a greater extent employ holistic perception, although it is difficult to assess whether particular ways of questioning may allow for eliciting feature-based reasoning from e.g. earwitnesses. The term 'Gestalt' is often detached from its original meaning in the research literature. Theoretical observations indicate that unfamiliar voice individualisations by lay and expert listeners may benefit from Gestalt processing in the original sense, which is not restricted to holistic processing.

3.2.3.4 Experimental studies

While it has been shown that the frame conditions of earwitness and expert witness testimony differ, several experimenters have attempted to put both types of witnesses in an identical situation, with the aim of assessing whether individualisation accuracy is a consequence of phonetic training.

Shirt (1984) conducted a seminal study to this end. It was a direct response to a motion passed at the 1980 Colloquium of British Academic Phoneticians, which demanded that phoneticians demonstrate their ability to individualise speakers before acting as expert witnesses in court (Shirt, 1984: 101). In essence, this motion is asking for the establishment of base rate information on individual experts to help a trier of fact assess the likely accuracy of their expertise. This plea is similar to the aim of this thesis, which is to find ways of establishing such base rate information on individual earwitnesses (cf. Section 3.2.2). Shirt's study was therefore created with a specific forensic application in mind.

The stimuli for the study were sourced from a pool of 4s-long recordings collected from 74 speakers. The number of stimuli elicited per speaker was not reported, but it is likely that at least two stimuli per speaker were collected to allow for same-speaker comparisons without repeating stimuli. Participants were subjected to 13 tests in different formats, including six closed individualisation tasks, two open pairing tests, two closed pairing tests, and three discrimination tests. In the closed individualisation tasks, participants were presented with a set of six stimuli as well as a seventh stimulus that was produced by one of the speakers who produced the initial stimulus set. In each of the pairing tests, ten stimuli were presented which had to be grouped into pairs by speaker identity. While in the closed pairing tests the number of different speaker identities was known to the participants, participants were not given this information in the open format tests. Lastly, each of the three discrimination tasks required the listeners to determine whether two stimuli were produced by the same speaker or not. When

this was not the case, the speakers were related to each other, the assumption being that family members possess similar sounding voices. Shirt did not, however, confirm this assumption by means of analysis. In general, the expected difficulty of voice pairs was not defined on a phonetic basis.

Results from 20 phoneticians and 20 lay listeners who completed the described test battery were presented in Shirt's report. In summary, she found phoneticians to perform marginally better than lay listeners, with an average accuracy of 53% (range 38 – 76) for the entire test battery, compared to 46% (range 19 – 76) for the untrained listeners (Shirt, 1984: 102). Watt (2010: 81) criticises Shirt's study for a lack of 'phonetic realism', arguing that the stimulus length of four seconds is not representative of typical forensic cases. Moreover, he argues that some of the 'mistakes' made by phoneticians in Shirt's study "were made for valid phonetic reasons" (Watt 2010: 81); the source of this information is unclear as Shirt does not report on her participants' reasoning.

On a more fundamental level, it may, however, be criticised that the conditions created in Shirt's study lack validity in terms of her aim to evaluate the impact of phonetic training on individualisation accuracy. It should be noted in this context that Shirt characterised the work as a pilot study (Shirt, 1984: 101). First, the results of 20 experts were included in the report, while 26 experts completed the study. The results of the six experts with the lowest accuracy were discarded for unknown reasons. If results are discarded in this sort of setup, it would be more conducive to use the amount/quality of phonetic training as a criterion (independent variable), rather than the achieved accuracy in the test (dependent variable). At the same time, all lay listener responses were considered. Second, Shirt observed that those experts who created a copy of the stimuli, and who were thus capable of comparing stimuli side by side, achieved the highest accuracy scores. Since the experts' results had been analysed before the untrained listeners' judgements were elicited, all lay participants were given the option of a side-by-side comparison, based on the phenomenon observed in the expert group (Shirt, 1984: 102). As a result, the lay listener group used a homogeneous approach, which was recommended to them, while the expert group was heterogeneous with regard to the used analysis techniques. It was also not reported whether all experts were aware that copying the files was an option. In summary, given the number of confounding variables and changes in the setup between groups, the accuracy averages of the experimental groups must be compared with caution. If, however, the groups are analysed separately, the greater accuracy of those experts who were able to directly compare samples indicates that the frame conditions of the

comparison may outweigh training-based effects (although this effect was not quantified). This impression is corroborated by the fact that individual lay listeners who were given the same opportunity outperformed some of the experts. Since the main advantage of side-by-side comparisons is a limited need for using working memory, memorisation might have a great impact on individualisation accuracy, even if samples are memorised intentionally and for a brief period only.

A similar design was chosen by Elaad et al (1998) who subjected three different listener groups to an individualisation study: One group of 15 blind lay listeners, a group of three phonetically trained listeners with forensic casework experience, as well as a control group of 18 sighted lay listeners. The experiment specifically investigated the role of working memory in speaker individualisations by making a distinction between long-term and short-term working memory. The experimental groups were chosen as it was hypothesised that the short-term working memory of listeners trained for voice comparisons is extended by long-term working memory, which means that some aspects of a memorised voice are stored in long-term memory but can quickly be retrieved if specific cues in the short-term memory are activated (Ericsson & Kintsch, 1995). This ability to extend the working memory load is assumed to be specific to areas of cognitive demand which are regularly used (Ericsson & Kintsch, 1995). For this reason, Elaad et al. (1998: 75) hypothesised that listeners who were born blind cannot access this skill to a greater extent than sighted listeners, unless they were specifically trained to compare voices with the aim of speaker individualisation. Experts who were trained to perform systematic voice comparisons were on the other hand hypothesised to outperform both lay listener groups. The assumed reason is that lay listeners are required to process both the first voice sample heard and the sample to which it is compared in short term memory, causing interferences.

Speech samples were collected from 69 male speakers (average age 25.9), who were randomly assigned to the 'perpetrator' (N = 16) or the 'innocent' condition (N = 53). Speakers in the perpetrator role committed a staged crime in which they stole a varying amount of money. They then were then asked to stage a phone call to an accomplice in which they reported on the crime's execution. In between one and seven days after the staged crime, all speakers were invited to collect KSs for the creation of VPs. The 'perpetrators' were offered a higher remuneration if their identity was not found out by the investigators. They were also consistently reminded of the importance of not being found out, which resulted in four of the perpetrators attempting to change their voice.

Seventeen VPs were created, one for each 'guilty' speaker, and one target-absent parade. The number of foil voices varied between one and five. All listener groups completed all parades (in multiple sessions if required) and did not know whether they were target-present or target-absent parades. Experts and lay listeners were put in the same situation in that experts were not allowed to use acoustic analyses.

The findings confirmed the authors' hypotheses as both lay listener groups individualised the guilty speakers' voices with an accuracy of 53%, while experts had an accuracy of 77%. Moreover, experts identified innocent speakers with an accuracy of 91%, while the blind and sighted lay listeners had an accuracy of 79% and 80%, respectively. The authors' conclusion that the experts' greater success is rooted in their harnessing of long-term working memory capacity may, however, exceed the available evidence. For one thing, only three experts took part in the experiment, compared to a total of 33 lay listeners. Moreover, the experimenters introduced voice disguise as an additional independent variable, which was not, however, considered in the analysis. It may therefore be the case that experts performed better because they were less frequently fooled by disguises (assumed that their training allows them to detect mismatched speaking styles). Since there were four known cases of 'guilty' speakers disguising their voice, such an effect would apply to a quarter of the 16 culprit-present parades.

In the same year, Schiller & Köster (1998) conducted a similar, but more balanced study with 27 German listeners, ten of whom being phoneticians and the remaining 17 participants lay listeners. Six speakers were recorded while reading aloud a ca. 1-minute-long paragraph. Subsequently, three four- to eight-second-long utterances were extracted from each speaker's recording and rerecorded under telephone landline conditions. Each of the six recordings obtained per speaker were then re-recorded three times, so that a total corpus of 108 speech samples was created (18 passages per speaker). The article does not state whether the measure of 're-recording' is in this case referring to a lossy process, resulting in declining imprint fidelity or whether the re-recorded samples were identical copies of the original utterances, and thus resulting in several indistinguishable stimuli.

One speaker was chosen as target. All listeners were familiarised with the target speaker's voice by listening to the entire passage five times, being aware that an individualisation task was to follow. Participants then listened to all 108 utterances in randomised order and decided for each utterance whether it was produced by the target speaker or not. On average, the expert group correctly identified the target voice in 98% of the occasions in which it was presented, whereas the equivalent score of the lay listener group was 92%. While both these hit rates are

high, the expert group's advantage over the lay listener group was statistically significant. There was, however, no significant difference in false alarm rates. Experts incorrectly made a same-speaker judgement in 1% of the cases in which a foil voice was presented, and lay listeners in 2%. The authors' assumption was that phoneticians are better suited for speaker individualisation testimony than lay listeners and they reached the conclusion that this is the case. Note, however, that although equal conditions were created for both participant groups, some of these conditions are exclusive to earwitness testimony. For instance, all participants had to memorise the target voice and compare the subsequent recordings to the memory imprint. Consequently, there was no option of comparing samples side by side, nor was there a possibility to listen to stimuli twice. The conditions are therefore not conducive to the authors' research question, since the consultation of a phonetically trained expert witness is dependent on the existence of a recorded QS. At the same time, the findings are relevant for the present thesis since the results suggest, in line with Elaad et al.'s (1998) findings, that phonetic training may result in an advantage when systematically comparing voices, even when the task is based on a memory imprint. Note, however, that some experts performed at ceiling level and most lay listeners achieved an accuracy above 90%, indicating that the experiment's difficulty was not well calibrated. In general, no attempt was made to describe or control the difficulty of the used stimuli.

In summary, existing research suggests that the conditions under which voices are compared have great impact on individualisation accuracy. This is especially true for the existence of a recorded QS, allowing the individual who compares the voices to compare samples directly rather than from memory. Nonetheless, there is some evidence to suggest that phoneticians have an advantage in memory-based tasks as well, even when not allowed to make acoustic measurements (which is their second unique advantage). A possible, but not well-investigated explanation is that listeners trained at voice comparisons might be able to unlock long-term working memory capacities.

3.3 Voice parade procedures

3.3.1 Background

A VP presents an earwitness with several voices, one of them being the KS and the other voices being foils. The purpose of this measure is a prior odds manipulation in favour of the defence hypothesis (cf. Section 3.2.1). This general principle is inspired by visual identity parades ('line-ups' in American English), which were likely invented by police forces in the 19th century, probably in response to demands of the judiciary for sounder individualisation evidence (Devlin Committee Report, 1976: 3). Criminologists of the 19th century were aware of the voice's capacity for individualisation (e.g. Bertillon, 1893). Nonetheless, the habit of conducting dedicated parades, in which the voice is the only identifying characteristic came into being much later. For instance, the Devlin Report, which evaluated the reliability of eyewitness testimony in the United Kingdom between 1908 and 1972, refers to one criminal case from 1972 and two cases from 1969 in which the suspect and foils in a visual line-up were additionally asked to produce speech samples (Devlin Committee Report, 1976: 67ff.). This hybrid approach was deemed necessary by the investigators because the perpetrators in these cases were either wearing forms of facial concealment or not seen clearly by the witness(es). The voice was therefore regarded as the "principal factor" for individualisation in these cases (Devlin Committee Report, 1976: 67). Consequently, there is evidence that in the early 1970s there had still not been a dedicated procedure in place for vocal individualisations, in spite of the judiciary's awareness of the voice's potential for individualisation.

Nonetheless, some good practice guidance[10] had been established for including voices in a visual line-up. For example, if the suspect produced an utterance, all other participants in the parade had to produce an utterance as well. Moreover, utterances were not included in the parade when the suspect's voice would stand out from the other participants' voices. In one case discussed in the Devlin report, witnesses reported that the perpetrator had a foreign accent. In the subsequent line-up, the suspect, being the only participant in the parade with a foreign accent, was not asked to produce an utterance because investigators feared that the witness might pick him based on the presence of a foreign accent alone (Devlin Committee Report, 1976: 68).

[10] In the discussion of a particular case, the Devlin Report explains that some of the measures regarding the inclusion of voices in a parade were in line with the "the current practice", implying that there is a norm for this sort of procedure. It is, however, not stated whether this norm is based on a particular document or rather born out of habit (Devlin Committee Report, 1976: 68).

These early guidelines are in line with the two central objectives of VPs described by Atkinson (2015: 57): First, the procedure should enable earwitnesses to make accurate identifications, meaning the speaker most likely to be chosen is the perpetrator (if present in the parade). Second, a fair test should not be so easy, that the suspect[11] stands out too clearly among the foils, nor so difficult, that the suspect's voice is indistinguishable from the foils. Atkinson points out that achieving both an accurate identification and a fair test can be mutually exclusive. Upon closer examination, these two objectives can be formulated as a single rule: The presented voices should be equally plausible options for a general listener, but not for the witness (provided that the perpetrator is present in the parade).

Nolan (1983) observes that early dedicated VPs did not follow this rule. When he was consulted by a police force, they asked him to listen to two voice line-ups that they had created. Despite lacking any prior knowledge of the suspects or the cases, he accurately identified both target voices. This is because the foil samples used in the VP were constructed from read speech, whereas spontaneous speech from police interviews was used for the suspect samples. Consequently, any listener, not just the witness, could easily pinpoint the suspect.

It should be mentioned that there is a third, and often overlooked objective of VPs, which equally applies to visual line-ups; the endeavour to avoid a direct confrontation of witness and suspect (Devlin Committee Report, 1976: 152). The VP is a device that prevents the witness from making a rushed decision, which is likely to happen in a direct encounter with the suspect. The individualisation evidence obtained in a direct encounter would therefore be of limited value in a trial (Devlin Committee Report, 1976: 152). It is assumed that the parade will eliminate uncertain witnesses, i.e. those who are not certain enough to make a judgement after having listened to all samples. This objective is, however, based on two premises that first have to be proven true: First, that there is a correlation between certainty and accuracy, and second, that uncertain witnesses actually refrain from making a judgement. Both these premises will be discussed in Chapter 4 in the context of recently conducted voice recognition tests (BVMT, GVMT, JVLMT). It may be foreshadowed at this point that, so far, there is no clear empirical evidence to underpin either of these premises.

[11] Changed from Atkinson's original rule, which says "perpetrator" (sic!).

3.3.2 Relevant legislation

Guidelines on the construction of dedicated VPs began to emerge in the 1990s, including, most notably, proposals by Broeders & Rietveld (1995), Hollien (1996), and Broeders & van Amelsvoort (1999; 2001). This research laid the groundwork for the so-called McFarlane guidelines, which were developed collaboratively by Detective Sergeant John McFarlane from the Metropolitan Police and Prof. Francis Nolan from the University of Cambridge (Nolan, 2003). Their collaboration stemmed from joint work on a case presented at the Central Criminal Court in 2002 (McDougall, 2021: 34). Soon thereafter, the Home Office released a circular entitled "Advice on the Use of Voice Identification Parades" (Home Office, 2003), which is based on the McFarlane guidelines and lays out a recommended procedure for VPs conducted in England and Wales (McDougall, 2021: 34).

The United Kingdom is geographically divided into three distinct CL jurisdictions: English and Welsh law, Scots law, and Northern Ireland law. Identification procedures under English and Welsh law are regulated in detail by Code D of the *Police and Criminal Evidence Act (PACE,* Home Office, 2017), which was originally passed by Parliament in 1984 (Emson, 2004: 332). This was a period in which dedicated VPs began to emerge, albeit in absence of a regulatory framework. Consequently, early versions of PACE Code D did not consider exclusively aural parades (Robson, 2018: 223). Police forces were not (and still are not) obliged to hold a VP, but several cases (e.g. *R v Hersey* [1998]) set a precedent for the admissibility of VP evidence (for a more detailed discussion, cf. Emson, 2004: 332). Eventually, Code D was updated in 2005 to explicitly mention VPs (Robson, 2018: 223). This and all subsequent versions have referred the reader to the Home Office guidelines (i.e. the McFarlane guidelines) for the specifics of such a procedure (Home Office, 2017: 4f.). Although VPs are explicitly mentioned in PACE, it is not clear whether a failure to conduct a VP results in a breach of PACE (Robson, 2018: 223).

The 2015 version of the Northern Ireland equivalent of PACE Code D, i.e. Code D of *The Police and Criminal Evidence (Northern Ireland) Order 1989*, mentions VPs but does not refer to any guidelines for their creation (Robson, 2018: 228). On the other hand, the *Lord Advocate's Guidelines on the Conduct of Visual Identification Procedures,* which regulate identification procedures in Scots law, suggest that dedicated VPs be conducted where appropriate and also outline a procedure (Robson, 2018: 227). The guidelines establish the general admissibility of earwitness testimony by referring to the case of *Lees v Roy* [1990] (Crown Office and Procurator Fiscal Service, 2007: Appendix G, n.p.). The suggested procedure is a live parade, during which

the suspect and volunteer foil speakers, who are "chosen for voice and accent similarity", produce utterances behind a screen (Crown Office and Procurator Fiscal Service, 2007: Appendix G, n.p.). The number of foils is not defined, nor is a curation process outlined. Suspects may listen to the parade multiple times and are not allowed to comment on the parade before having heard all speakers (Crown Office and Procurator Fiscal Service, 2007: Appendix G, n.p.). Witnesses are also reminded that the perpetrator may not be present in the parade. If no identification is made, the suspect is to be asked whether any of the speakers "sounds like the person" (Crown Office and Procurator Fiscal Service, 2007: Appendix G, n.p.). From an objective point of view, the conduciveness of this question is doubtful, given that one objective of a VP is to eliminate uncertain witnesses (cf. Section 3.3.1). In March 2021, the Crime Strategy Department of the Scottish Police published a national guidance paper on identification parades reiterating this procedure (Police Scotland, 2021). Note, however, that the Scottish guidelines for visual line-ups explicitly permit asking the participants to speak (Robson, 2018: 227); similar to the hybrid line-ups outlined in the Devlin Report (cf. Section 3.3.1). The result would be a parade with a focus on the participants' voices, in which the foils were not selected for their vocal characteristics but for their visual appearance (Robson, 2018: 227), i.e. potentially violating the rule that no speaker should stand out.

3.3.3 The Home Office guidelines and their application

As demonstrated, the current state of VPs in the UK is heterogeneous. The present research will focus on the Home Office guidelines for England and Wales (Home Office, 2003), which – out of the discussed procedures – have to the greatest extent benefitted from linguistic insight. Not all aspects of these guidelines were, however, determined based on linguistic research. It was rather an aim at the time to closely follow the well-established procedure for visual parades, in order to mitigate the risk of defence attorneys challenging VPs (Pautz et al., 2023: 2, note 1). There is consequently a growing number of studies on those parameters that are not yet based on linguistic or psychological research (e.g. McDougall, 2013b, 2013a, 2021; Nolan et al., 2011; Pautz et al., 2023; Smith et al., 2019, 2020, 2022; Smith & Baguley, 2014). This research will be discussed in detail in Section 3.4.2.1, which deals with the variables in earwitness cases that can be controlled by the judiciary.

In terms of general rules, the Home Office procedure does not allow live line-ups, which means that all parades must be compiled from recorded samples (Home Office, 2003: Pt. 7). Moreover, it is advised that these samples be collected within four to six weeks after the crime

to account for degradation of the witness's memory imprint of the QS (Home Office, 2003: Pt. 10). In total, the police officer in charge is asked to supply the expert compiling the parade with a sample from a police interview with the suspect (Home Office, 2003: Pt. 8) as well as 20 samples "from persons of similar age and ethnic, regional and social background as the suspect" (Home Office, 2003: Pt. 9). These recordings may be taken from police interviews from unconnected cases (Home Office, 2003: Pt. 9).

It is the expert's task to then select eight out of the twenty recordings that are suitable as foils (Home Office, 2003: Pt. 13) and to create the VP stimuli from the provided material. The nine stimuli should consist of ca. one-minute-long recordings, featuring fragments of speech and/or continuous speech (Home Office, 2003: Pt. 13). It is not specified whether the recordings should be treated to mimic telephone bandwidth in cases where the witness was exposed to the perpetrator's voice over the telephone, a measure undertaken by the German Federal Police when constructing voice line-ups (Künzel, 1994: 137). It must further be guaranteed that the propositional content of the stimuli does not inadvertently individualise the speakers as the suspect or a foil (Home Office, 2003: Pt. 15). To ensure procedural fairness, three versions of the parade should be created that are made up of the same stimuli but in different randomised orders (Home Office, 2003: Pt. 14), with the defence lawyer selecting the version to be used (Home Office, 2003: Pt. 24).

The expert's core task is thus the selection of foil speakers who provide "a fair example for comparison against the suspect", considering "accent, inflection, pitch, tone and speed of the speech" (Home Office, 2003: Pt. 15). Alternatively, the expert may advise against conducting a VP at this stage, in cases where the suspect's voice is so unusual that it would stand out even to an impartial listener (McDougall, 2021: 34). Rather than suggesting a clear procedure for the selection of foils, the guidelines recommend assessing the validity of the foil selection by means of conducting test parades with lay mock witnesses (Home Office, 2003: Pt. 16). The mock witness, who is given some background information about the case, should "try and pick out the suspect" (Home Office, 2003: Pt. 16). If the parade is fair, the witness will not be able to do so, or only be able to do so with a random chance of success. The number of test parades is not specified by the guidelines, nor is clear advice given on how to interpret the results. A desirable outcome would certainly be a situation in which none of the mock witnesses make a judgement, indicating that no voice stands out. It is, however, not specified whether the selection of a voice would necessarily lead to an exclusion of that voice from the parade.

The test should ensure that the suspect's sample does not stand out in terms of content (Home Office, 2003: Pt. 17 i) or in terms of the "manner of speech" (Home Office, 2003: Pt. 17 ii). In practice, modern applications of the guidelines address these two criteria in separate tests. For instance, following a procedure outlined by McDougall (2021: 35), the foils, which are chosen by the expert on phonetic grounds, subsequently undergo a perceptual distance test (de Jong et al., 2015; McDougall, 2013a). During this test, impartial lay listeners provide pairwise similarity ratings of the VP stimuli on a nine-point scale. With the help of the multidimensional scaling method, the obtained similarity ratings can be expressed as a visual arrangement of the stimuli in a two-dimensional space, allowing for the detection of outliers (McDougall, 2021: 35). Recent research has also explored possibilities of using automatic speaker recognition (ASR) systems in these processes (Gerlach et al., 2020, 2023).

To detect samples whose propositional content may induce bias, a second test is conducted, in which lay listeners who were familiarised with details of the case listen to all VP stimuli. For each sample they assess on a nine-point scale whether it is likely that the speaker was interviewed about the crime in question. Samples with an extreme average are rating are removed (McDougall, 2021: 35).

During the actual parade, the witness must be explicitly instructed that the voice of the perpetrator[12] may or may not be present in the samples played (Home Office, 2003: Pt. 25). The witness is allowed to listen to all the voices before making a judgment, and the parade can be repeated as many times as the witness desires (Home Office, 2003: Pt. 25). While the Home Office guidelines refer to VCR technology, modern VPs are typically presented as PowerPoint presentations with embedded sound files (McDougall, 2021: 35).

To prevent bias, the officer conducting the parade should only be informed of the speakers' identities after the parade is completed (Home Office, 2003: Pt. 27). The aim of this requirement appears to be the creation of a double-blind procedure, in which neither the witness nor the identification officer knows which sample was produced by the suspect. Psychological studies on visual line-ups (Fitzgerald et al., 2015; Kreiman & Sidtis, 2011; Wixted & Wells, 2017: 15) suggest that a double-blind setup should be the standard for ensuring impartiality and accuracy in voice identification parades. However, it should be noted that following the guidelines will not lead to a double-blind setup under all circumstances, as both the officer

[12] The guidelines say „suspect" (sic!). This is likely to be a mistake as the suspect's voice will always be present in the parade. It is more plausible that the warning would point out that the voice which the witness heard at the crime scene, i.e. the perpetrator's voice may not be present, which is the case when the suspect is innocent.

compiling the material for the parade and the officer conducting the parade are referred to as "the identification officer" (Home Office, 2003: Pts. 6 & 27), which in theory allows the same person to perform both tasks.

In sum, it can be observed that current approaches to VPs conducted in England and Wales benefit from relatively comprehensive (albeit non-compulsory) guidelines. Gaps in these guidelines must be interpreted by individual experts and have led to the development of now established procedures for foil selection and parade validation. The guidelines establish safeguards against a biased line-up. The expert's role comprises a series of suitability checks, entailing the qualitative suitability of the audio material for a VP, the suitability of the suspect's voice for a parade, and the suitability of potential foil speakers for a parade. There is, however, no suitability check in place for the witness. This aspect may be crucial given the central rule that the suspect's voice should stand out to the witness, but not to other listeners. For this rule to apply, it is a necessary prerequisite that the witness possesses a certain level of voice processing skills and is generally capable of identifying (unfamiliar) individuals by their voice.

3.4 Variables in earwitness cases

3.4.1 The Hauptmann case and seminal earwitness research

In 1932, the infant son of Charles Lindbergh, pilot of the first non-stop flight across the Atlantic Ocean, was kidnapped from the Lindbergh's family home (Fisher, 1987). The subsequent legal proceedings (*The State vs Hauptmann* [1935]) against the prime suspect Bruno Richard Hauptmann, a German immigrant, became one of the most infamous modern-day trials in which individualisation evidence was accepted from an earwitness. The witness was Lindbergh himself who heard the perpetrator's voice when he and his friend Dr John Condon delivered ransom to the kidnapper (Kennedy, 1985: 266 ff.). The delivery took place at a Bronx cemetery at night-time (Kennedy, 1985: 266 f.). While waiting for the kidnapper in their car, Lindbergh and Condon heard a male foreign accented voice shouting the words "Hey, doktor! Over here, over here", from a distance of approximately seventy to one hundred yards (Solan & Tiersma, 200: 373). In spite of the ransom payment, Lindbergh's son was not returned, and found dead one month later (Fisher, 1987). In September 1934, 29 months after hearing the voice at the cemetery, Lindbergh acknowledged in front of a grand jury that it "would be very difficult to sit here and say that I could pick a man by that voice" (Fisher, 1987: 248). Yet, he was called into the district attorney's office the following morning, where he met Hauptmann, who was

asked to produce the same words that Lindbergh heard at the cemetery (Solan & Tiersma, 2003: 373 f.). In trial, Lindbergh then testified that it was Hauptmann's voice which he had heard at the cemetery, and that he recognised the voice when hearing it again at the DA's office (Fisher, 1987: 249). Hauptmann was sentenced to death in 1935 and executed in 1936 (Fisher, 1987).

Lindbergh's lawyer stated that the earwitness evidence played a key role in the conviction: "The minute Lindbergh 'pointed his finger' at Hauptmann, the trial was over" (Solan & Tiersma, 2003: 374). Although the admission of earwitness evidence was in line with legal precedent at the time (McGehee, 1937: 249), it caused controversy and sparked a research interest in the voice recognition skills of lay listeners. It was generally conceded at the time that individuals could be identified by their voice (McGehee, 1937: 250). However, some circumstances had already been defined, in which earwitness evidence should be given less weight (this applies to US jurisdictions), e.g., when the witness is not familiar with the speaker's voice (McGehee, 1937: 250). This applies to Lindbergh, who heard Hauptmann's voice for the first time at the cemetery (Read & Craik, 1995: 6). It was also conceded that some voices may be more distinctive than others (McGehee, 1937: 251). The particular constellation of events in *The State vs Hauptmann* [1935] did, however, draw attention to an array of additional variables that may have a bearing on a lay listener's ability to correctly identify an unfamiliar voice. These are as follows.

First, regarding the elicitation of testimony, it must be asked whether the long delay between the first exposure to the voice and the identification procedure had an impact on the witness' ability to recall the voice from memory. Moreover, it is important whether certain expectations were raised when the testimony was elicited. It is reported, for instance, that the DA asked Lindbergh: "Would you like to see the man who kidnapped your son?" before introducing him to Hauptmann (Solan & Tiersma, 2003: 373), hence priming the witness.

Second, regarding the circumstances of the crime, it is of interest whether particular accents, in this case foreign accents, lead to better memorisation of the voice than others, and whether the speaker's voice was disguised. It is also important to consider that the perpetrator was shouting at the first encounter, while Hauptmann was presumably using his normal voice in the DA's office, which means that the QS and KS were mismatched for speaking style.

Third, regarding the listener, it is of interest whether the emotional stress to which Lindbergh was subjected affected his ability to memorise the perpetrator's voice and whether an intentional effort was made to memorise the voice, as opposed to a merely incidental memorisation.

Some of these variables were addressed in a seminal study by Frances McGehee (1937). She conducted a series of voice memory tests with lay listeners. Participants listened to an unfamiliar speaker reading a short paragraph behind a screen. Different listener groups were then reinvited for an identification procedure after different time intervals. In the identification sessions, five voices were presented to the listeners in the same manner, one of them being the previously heard voice. A relatively high recognition accuracy of between 81% and 83% was observed when participants were reinvited within a week after first exposure to the voice. After two weeks, average accuracy dropped significantly to 69%, decreasing even further to 35% and 13% after three and five months, respectively. Different versions of the experiment were conducted, in which one or more additional independent variables were actively manipulated, including a voice pitch disguise on the part of the speaker, the number of speakers heard during the initial exposure, the accent of the speakers, and the task awareness of the listeners. Time intervals were actively manipulated in all versions of the experiment. A total of 740 listeners took part in the experiments divided into 33 different groups, which differed in conditions.

While the research questions brought forward by McGehee have remained highly relevant, her findings must be interpreted with caution. In spite of her premise that all voices are not equally distinctive, the vast majority of her conclusions were drawn from participants who were exposed to the same voice. Moreover, the listener pool was highly homogeneous regarding demographic background and age, consisting predominantly of graduate students. At the same time, it was not well balanced for sex (75% men). This had a great impact on her findings, as listener sex was not actively controlled as an independent variable in a dedicated version of the experiment. Instead, a combined effect of listener sex was established by comparing the performance of men and women across versions of the experiment. Most importantly, however, many listener groups were subjected to manipulations of more than one independent variable (at times up to 5 different variables), making it impossible to assess the impact of individual independent variables on recognition accuracy, the dependent variable. The fact that the results were not analysed statistically and restricted to averages adds to this problem.

Nonetheless, McGehee's research showed that the accuracy of earwitness testimony is dependent on a plethora of potentially interdependent variables and that the particular constellation of variables can have an immense effect on the evidential value of earwitness testimony.

3.4.2 Classifying variables

In modern earwitness research, the many variables which may have an impact on the accuracy of earwitness testimony are usually subdivided into "system variables" and "estimator variables". The distinction was first made by Wells (1978) in reference to eyewitness testimony but has also been adopted by earwitness researchers. The term "system variables" was chosen because these variables can be controlled by the criminal justice system, specifically the investigator who elicits the testimony, whereas estimator variables are beyond the control of the investigator. In the following sections, research on the most important estimator and system variables is reviewed.

3.4.2.1 System variables

The benefit of exploring system variables is evident, as they can help establish a best practice for the elicitation of witness testimony (Beaudry et al. 2014: 1384). Research on these variables has thus increased after the advent of standardised VP procedures, such as the guidelines brought forward by the UK Home Office (2003).

Several works on system variables have focused on the optimal time span during which a VP can yield robust results (e.g. Kerstholt et al. 2006; Öhman et al. 2013) or the potential benefit of different types of interviews conducted between the exposure to a perpetrator's voice and the VP (Memon & Yarmey 1999), as well as the dangers of providing witnesses with post-event information (McAllister et al. 1988; Smith & Baguley 2014). Smith and Baguley (2014) conducted an incidental memory test in which 72 participants listened to a 37-second-long dialogue between a male and female speaker, both having relatively low-pitched voices. Depending on the test group, the participants were subsequently told that other listeners described either the male speaker or the female speaker as having a high-pitched voice. After engaging in a free verbal recall task in which they described anything they could remember about the target voices and filling in a VQ questionnaire, the participants were subjected to two VPs, one for the male speaker and one for the female speaker. One of the VPs was a target-absent line-up and the other one a target-present line-up. Which line-up contained a target-speaker was determined at random. Participants stated whether they thought that the target speaker was present, and picked a voice from the line-up if they thought that this was the case, additionally stating their confidence on a 10-point scale. Average accuracy was low overall, but significantly higher for target-present lineups. The false post-event information affected ratings of voice pitch but did not impact identification accuracy or confidence. Surprisingly, the level of confidence displayed by the participants did not consistently correlate with accuracy,

implying that a sense of certainty did not necessarily reflect a true identification. On the other hand, the extent of verbal recall proved to be a strong predictor of accuracy.

More recently, studies have focused on the stimuli used in the VP procedure. For instance, several studies have investigated whether the length of the stimuli used in a VP has an impact on the accuracy and confidence of lay listener performances (Kerstholt et al. 2004; McDougall 2021; Smith et al. 2020). Smith et al. (2020) compared VPs using 15s-long stimuli with VPs using 30s-long stimuli. They found that, all other variables being equal, stimulus length did not have an impact on accuracy, and nor did participant accuracy and confidence correlate. Pautz et al. (2023) expanded on this setup by adding VPs with 60s-long stimuli, which is the recommended length for VP stimuli according to the UK Home Office's guidelines (2003). While different stimulus lengths did not lead to significant differences in participant accuracy, participant sensitivity (i.e., the trade-off between hits and false alarms) was comparable for 15s- and 60s-long stimuli, but significantly lower for 30s-long stimuli. The authors drew the tentative conclusion that the shortest stimuli may have the benefit of higher distinctiveness, and the longest stimuli the benefit of revealing more listener-specific information, whereas an intermediate length may prevent the listener from fully benefitting from either effect.

In terms of stimulus quality, Smith et al. (2019) found out that voice discrimination accuracy is highest when the used stimuli are not mismatched for speaking style. The specific investigated speaking styles were free, unscripted speech and read speech. Participants rated several voice pairs. Whereas the first recording in a pair always contained read speech, the second recording was either made up of read speech as well or of spontaneous speech, produced by the same or a different speaker. There were 48 trials per participant. Both accuracy and confidence were highest when speaking style did not change between recordings of a pair. A similar experiment was conducted by Stevenage et al. (2021), who investigated the same speaking styles in a discrimination task. In contrast to Smith et al. (2019), they allowed for both styles to occur in the first as well as the second voice in a pair, resulting in four possible combinations. Performance was significantly better when the speech style was kept constant within a pair. There was, however, no difference in performance whether free or scripted speech was presented first, indicating neither style resulted in perceptual deficits. Confidence was strongly correlated with accuracy in the "same" trials. Overall, speaker discrimination was significantly better for read stimuli than free speech stimuli. These findings are highly relevant as the Home Office guidelines do not specify whether parades should be adapted to the style of speech with which the witness was confronted at the crime scene. Further research on this topic

would be needed, especially experiments that investigate voice recognition rather than discrimination, ideally with a strong memory component. Such experiments could additionally address questions of matching VP stimuli for technical quality, e.g. when the perpetrator was heard over the telephone.

Smith et al. (2020) compared different parade types, and concluded that a sequential mode of stimulus presentation, which means that the witness makes a same/different judgement after each presented voice, is preferable to a serial presentation in which the witness makes only one identification (if possible) after having listened to all voices. A total of 91 participants took part in a VP task, counterbalanced for target-presence or absence. Each voice parade consisted of nine samples, each lasting 15 seconds. It was found that the overall accuracy of the participants was relatively low; however, the sequential procedure yielded better accuracy compared to the serial one. Specifically, the target-present sequential lineups were the only condition that significantly surpassed chance accuracy, achieving a 39% success rate. Both procedures showed high false alarm rates, although the sequential condition had numerically lower false alarm rates compared to the serial condition. The authors concluded that the sequential format might enhance discriminability, potentially by mitigating interference effects that develop when a large number of stimuli must be stored in short-term working memory during the encoding.

3.4.2.2 Estimator variables

Due to the dichotomous nature of Wells' categorisation, "estimator variables" can be negatively defined as those variables that are not system variables. The resulting category is heterogeneous in that it subsumes variables related to the event about which the testimony is being elicited – including the perpetrator – as well as variables related to the person whose testimony is being elicited. Hence, estimator variables are further subdivided here into "event variables" and "witness variables" for the purpose of the following overview. Event variables are particularly difficult to study as criminal events cannot be directly observed by researchers. Conclusions can therefore only be drawn from laboratory settings or studies in which a criminal event is simulated (i.e. the so-called "controlled field setting" cf. Cutler & Wells, 2009: 102).

Event variables are related to the nature of the criminal event and the resulting characteristics of the exposure to a voice at the crime scene. Legge et al. (1984) observed in a laboratory study that longer exposure times to a voice result in higher recognition accuracy. A similar connection between exposure time and accuracy was found for the recognition of voices heard over the telephone (Yarmey 1991).

Other variables dictated by the circumstances of the crime include the number of voices present at the crime scene and the type of interaction between the perpetrator and the witness. Several studies indicate that recognition accuracy decreases as the number of originally encoded voices increases (Carterette & Barnebey, 1975; Legge et al., 1984; McGehee, 1937). In the context of an earwitness situation this translates to the number of perpetrators, or – more generally – the number of people present during the criminal event. Carterette and Barnebey (1975) also explored interactions between the number of voices presented during exposure and other variables. A significant interaction was found between the number of voices and the delay in voice recognition. Participants were exposed to either two, three, four, or eight voices in exposure and the double number of voices in the test phase, the additional voices being foils. The delay between exposure and test was actively manipulated by the experimenters to be either 0, 15, or 45 s long. It was observed that the impact of delay on recognition performance varied depending on the number of voices in the recognition set. Specifically, the effect of delay on voice recognition was pronounced when dealing with a smaller number of voices. In contrast, as the number of voices increased, the influence of delay on recognition performance diminished. Furthermore, the study highlighted the role of cognitive strategies employed by subjects during voice recognition tasks. Subjects reported employing mental imagery and articulatory positioning to 'rehearse' voices, indicating a cognitive process beyond mere auditory recognition. This strategy seemed to be more effective when dealing with a smaller set of voices, suggesting that it becomes less useful or relevant as the number of voices increases. However, the study lacks forensic realism, as the stimuli consisted of a phrase comprising three words with a duration of about one second. It is not plausible to assume a criminal context in which eight voices are heard for one second each producing the same utterance without any overlapping speech, and earwitness evidence being elicited within a minute of the event.

Two studies considered a scenario in which a witness has multiple encounters with the same perpetrator. In this connection, Cook and Wilding (2001) showed that listeners who were exposed to a speaker's voice three times had a significantly higher recognition accuracy in a VP conducted a week after the exposure than listeners who only heard the speaker once. However, stimuli consisted of a single sentence and in the case of multiple exposure, the same stimulus was repeated. A similar experiment by Yarmey and Matthys (1992) yielded different results; they found that recognition accuracy was higher when listeners were exposed to the speaker's voice twice, but not after three exposures. This effect is difficult to evaluate as the

researchers manipulated several independent variables, including the length of the used stimuli and delay after which the VP was held.

Hammersley and Read (1985) showed experimentally that listeners recognised a voice better when they actively conversed with the speaker, rather than passively overheard a conversation between the speaker in question and a third person. A related question is whether the witness is also a victim of the crime, which increases the likelihood of an active conversation with the perpetrator but may at the same time affect recognition by inducing psychological stress or trauma. In a meta-study on eyewitnesses, Deffenbacher et al. (2004) concluded that increased stress results in lower hit rates in line-up procedures. Comparable results for earwitnesses, however, are not available. The type of encounter with the perpetrator further determines whether the voice is the only feature that might facilitate identification at a later point, or whether the witness was additionally exposed to the perpetrator's face, leading to a potential facial overshadowing effect (Tomlin et al. 2017).

In his discussion of estimator variables, Wells (1978: 1550) separates "characteristics of the criminal event" from the "characteristics of the defendant". While such a distinction can be made on logical grounds, it cannot necessarily be made in practice, as both have an impact on the unique properties of the vocal material with which a witness is confronted during the particular event. Characteristics of the perpetrator will therefore be regarded as event variables for the purpose of this study because they are part of the event that generated the vocal material. For example, this vocal material might be characterised by background noise (Smith et al. 2019), which is a characteristic of the crime scene, but also by the emotional state and resulting changed voice quality of the perpetrator (Read & Craik 1995), which is a characteristic of the defendant. It is therefore difficult to draw a clear-cut distinction between these characteristics as they are uniquely combined in a speech sample in each criminal event.

Estimator variables related to the witness differ from event variables in that they characterise the witness, i.e. the individual performing the identification, rather than the voice that is to be identified. Separating the witness variables from other estimator variables is not always possible as they can interact with event variables. Several studies have shown, for instance, that listeners perceive voices as more similar when they do not have a good knowledge of the target language or accent (Fleming et al. 2014; Sherrin 2015; Wester 2012; Winters et al. 2008; Yarmey et al. 2001). Philippon et al. (2007) showed that the same effect led to lower listener accuracy in voice line-ups. In such cases, the characteristics of the witness must be interpreted in conjunction with the characteristics of the perpetrator. Other witness variables

are, however, solely attributable to characteristics of the witness. For example, Bull et al. (1983) found that blind listeners outperformed listeners with normal vision in a voice recognition task.

3.4.3 Consequences of this classification

Wells' classification of variables has had great impact on witness research. It has above all raised awareness for the role of the criminal justice system in the elicitation of witness testimony, and the variables that it can actively manipulate. Wells himself reached the conclusion that system-variable research "may, as a general rule, have greater applied utility for criminal justice than does estimator-variable research" (Wells, 1978: 1555). His work therefore led to a surge in research on system variables, which can be characterised as a paradigm shift:

> Wells' (1978) distinction between system and estimator variables not only revolutionized how researchers and courts alike thought of the role of different variables, but it also (perhaps inadvertently) directed the field away from estimator variables and toward system variable research. This shift in the field led to major developments regarding best-practice identification procedures and provides the empirical support for widespread recommendations regarding the collection of eyewitness evidence. Perhaps the time has come for another shift – one toward a more balanced investigation of estimator and system variables. After all, estimator variables are always present and their impact is unlikely to be entirely eliminated even by best-practice identification procedures. (Beaudry et al., 2014: 1393)

The first claim of Beaudry et al.'s assessment, i.e., an observable paradigm shift towards system variables, certainly applies to eyewitness and earwitness research alike. The development has, however, been slower in earwitness research. This may be due to the fact that dedicated procedures for the elicitation of earwitness testimony came into being about two decades after Wells had developed his classification. Thus, there had not been a procedure in place which could have been fine-tuned by means of system variable research. The number of studies on the improvement of VP procedures continues to grow. A recent example of a coordinated approach to tackle the lack of system variable research is the IVIP ('Improving Voice Identification Procedures') project of the University of Cambridge.[13] Such research has led to significant insights into the construction of VPs and was long overdue, given the plethora of similar investigations into the role of system variables in eyewitness cases. Research on the system variables of VPs may even be of greater exigence than equivalent work on visual line-ups, as

[13] https://www.phonetics.mmll.cam.ac.uk/ivip/overview (Last accessed 24/09/2023).

VP frameworks, most notably the McFarlane guidelines, have borrowed heavily from existing procedures for eyewitnesses (cf. Section3.3.3; Pautz et al., 2023; Smith et al., 2020). It must therefore be critically assessed whether procedures applying to eyewitnesses translate to earwitness testimony to a satisfactory degree.

Beaudry et al.'s second argument, i.e. the plea for a new paradigm shift towards more balanced investigation of estimator and system variables, is in line with the present research project. A central point in this context is their observation that estimator variables "are always present" and have great impact on witness reliability, even though they cannot be controlled. This is because estimator variables describe the criminal event itself, whereas the work of the criminal justice system can only begin after a criminal event has taken place. System variables are therefore necessarily limited to the elicitation of testimony from the witness. It could in fact be said that system variables are not only controlled by the legal system but in fact created by it; the elicitation of testimony from the witness is the premise for their existence. For example, finding the optimal number of foil recordings in a VP is only a sensible measure if a VP is to be held in the first place. The importance of understanding estimator variables is therefore self-evident.

At the same time, a paradigm shift that strives for a better integration of system and estimator variable research might benefit from a more nuanced approach to estimator variables. It has already been demonstrated in Section 3.4.2.2 that estimator variables are a heterogeneous category, which subsumes all variables that are not system variables. It may be due to this subsumption that one of the conclusions drawn by Wells (1978) is generally overlooked. While he is often quoted with the above-cited observation that system variable research may be more viable than estimator variable research, he identified a particular group of estimator variables that may prove more useful than others. Specifically, he refers to those estimator variables which "use the individual as a unit of analysis" rather than the situation (Wells, 1978: 1555). In other words, the role of the individual witness as a variable in the individualisation task. The subdivision of estimator variables into event and variables, as applied in the discussion in Section 3.4.2.2, therefore intends to account for this distinction. Table 2 provides a non-exhaustive classification of witness variables, following this classification (variables highlighted in grey are the focus of the empirical part of this thesis):

Table 2: Classification of earwitness variables

Variable	System variable	Estimator variable	Event variable	Witness variable
Stimulus duration in VP	X			
Number of foils in VP	X			
Mode of stimulus presentation (e.g. sequential vs serial)	X			
Vocal profile of the suspect	X			
Vocal profile of the foils	X			
Vocal profile of the perpetrator		X	X	
Sex/gender of perpetrator		X	X	
Sex/gender of suspect		X	X	
Sex/gender of witness		X		X
Witness' familiarity with the perpetrator		X	X	
Vocal disguise by perpetrator		X	X	
Number of voices present at crime scene		X	X	
Duration of exposure at crime scene		X	X	
Age of the suspect		X	X	
Age of the witness		X		X
Age of the perpetrator		X	X	
Hearing abilities of the witness		X		X
Voice discrimination capabilities		X		X
Voice memory capabilities		X		X
Task awareness of witness during crime		X		X

Recent research into between-listener differences underpins the special role of witness variables. Findings suggest that these variables may to a lesser degree be restricted to estimation. Wells chose the term "estimator variables" because the criminal justice system could "use such knowledge to estimate, post hoc, the likely accuracy of a witness" (Wells 1978: 1548). More specifically, he reckoned that investigators could "plug in" the values of a specific crime "and make a professional estimate regarding how likely it is that the witness(es) could give accurate or inaccurate testimony under such conditions" (Wells 1978: 1548). To this end, expert witnesses may appear in court to make case-specific or general statements regarding earwitness testimony (Wells 1978: 1548). A possible danger of this use of estimator variables is that experts are likely to rely on average performance in empirical data, since the specific witness's capabilities are not known (Wells 1978: 1551). However, recent psychological tests concluded that untrained listeners differ substantially in their voice processing capabilities (Aglieri et al. 2017; Humble et al. 2022; Mühl et al. 2018). The heterogeneity of the category "estimator variables" therefore raises the question of whether an estimation based on averages

is the optimal or only option of assessing these variables, or whether a particular subgroup of estimator variables, namely the witness variables, could be assessed by means of testing. Testing witness variables might help establish credibility in the individual witness and is thus the purpose of this thesis.

3.5 Complementing voice parades

3.5.1 Weaknesses of voice parades

According to the *PACE* Code of Practice D (Home Office 2017) identification procedures have a twofold purpose as they are designed to "test the … witness' ability to identify the suspect" as well as to "provide safeguards against mistaken identification". In VPs, safeguards are provided in the form of several foil voices amongst which the voice of the suspect is presented. A VP can therefore have the outcomes summarised in Table 3.

Table 3: Possible outcomes of a voice parade

	Suspect		**Foil**
	Guilty	**Innocent**	**Innocent**
Selected by witness	Hit	False Alarm	False Alarm
Not selected by witness	Miss	Correct Rejection	Correct Rejection

It can be observed that the VP is a test with two subjects. It is the primary goal of the procedure to gather information about the suspect, i.e. to establish whether the suspect's voice was present at the crime scene. In this function, the witness – within the safeguards of the VP – acts as a diagnostic test that compares the vocal material memorised from the crime scene with the vocal material contained in the parade, with the aim to find a 'match'. This comparative function of the VP is particularly apparent in a sequential procedure, in which a judgement is elicited after each presented voice. If the witness matches the suspect with the memorised voice, the outcome can either be a 'hit' (correct match) in the case that the suspect is the perpetrator, or a 'false alarm', if the suspect is innocent. If the witness does not select the suspect's voice, the outcome can either be a 'correct rejection' (in the case of an innocent suspect) or a 'miss' (in case of a guilty suspect). Since the creator of the VP cannot know whether the suspect was the perpetrator

('culprit present') or not ('culprit absent'), the witness functions as a test to determine whether the presented parade is a culprit-present VP or a culprit-absent VP.

It is the secondary goal of the procedure to assess the witness's credibility. Specifically, the witness him/herself becomes a test subject of the VP due to the presence of foil speakers. Since these foil speakers are known to be innocent, they can elicit the same results as an innocent suspect, i.e. a false alarm if they were wrongly selected, or a correct rejection if the witness did not pick any voice or picked the suspect (Table 3). In the literature, witnesses selecting a foil voice may be described as "failing" the parade (Wixted & Wells 2017: 15), which stresses the VP's function as a test for the witness. This scenario is, however, the only way in which an investigator can detect the failing of a witness. In cases where a witness selects the suspect, or does not make a selection, they consequently 'pass' the parade, while the respective undesirable outcomes of false alarm or a miss cannot be detected.

In summary, in its typical form the VP produces a single data point that is meant to be an assessment of both the suspect and the witness. This one data point can discredit the witness evidence if a foil voice was selected, which may have an impact on the witness's credibility in unrelated questions (i.e. questions not about the criminal event rather than the perpetrator). While witnesses can be observed making incorrect identifications in a VP, there is, however, no way of demonstrating their ability to make accurate ones. The only safeguards against an erroneous identification of the suspect are therefore the relatively low likelihood of coincidentally choosing the suspect (1/9 if the Home Office guidelines are followed). The paucity of data points puts the trier of fact in a difficult situation, as from their perspective the result of the VP is an "estimator variable" (Semmler et al. 2018: 404), i.e. a variable they cannot control. Their interpretation of the VP evidence might further be influenced by the witness's confidence. Mock juror studies have shown jurors to be generally believing of earwitness testimony (McAllister et al. 1993), although perception studies generally report low correlations between listener accuracy and confidence (Kerstholt et al. 2004; Öhman et al. 2011; Sørensen 2012); possible exceptions are discussed by Wixted & Wells (2017).

3.5.2 Potential benefits of a screening test

Given the shortcomings of VPs, the present thesis asks the question whether the VP's primary function of eliciting a judgement regarding the suspect could be improved if the function of assessing the witness's credibility was tested separately. Table 4 juxtaposes the characteristics of the VP procedure with the characteristics of a potential complementary screening test.

Table 4: Comparison of a voice parade and a complementary screening test

	Voice parade	**Complementary screening test**
Function	Obtain evidence	Determine weight of evidence
Aim	Identification of a known sample (e.g. suspect)	Assess individual's general ability to recognise voices
Designed for	Every witness (in theory) Average witness (de facto)	Every witness
Premise	Witnesses have similar abilities	Witnesses have different abilities
Experimenter	Knows the identity of the foil voices (wants to determine whether the suspect is the target)	Knows the identity of all voices (can determine whether the target is present or absent)
Outcome	One judgement (accuracy unknown), self-assessed confidence (reliability unknown)	Many judgements (known accuracy), many confidence measures → known correlation between accuracy and confidence, potentially reaction time
Information on witness	none	Witness behaviour (potential bias towards positive or negative identifications, relationship between accuracy and confidence) Witness capability (e.g. voice recognition skills, voice memory skills)

While it is the main function of the VP to produce evidence of the suspect's presence at the crime scene, the screening test may help determine the weight of the evidence. The VP is a one-size-fits-all approach, because the system variables have been optimised according to average performances of participants in perception experiments. The complementary test, on the other hand, is meant to assess the voice processing capabilities of the individual, i.e. a specific listener rather than the average listener. The advantage of the screening test is that it can consist of multiple trials, which can be controlled by an investigator. This makes it possible to obtain listener judgements on many voice samples, while the correct solution is known to the investigator. The investigator can thus determine the witness's general accuracy. If the test includes confidence measures as well, the investigator will also learn of the accuracy-confidence correlation across the witness's responses. The lack of knowledge about the relationship between confidence and accuracy is a general problem of individualisation procedures (Semmler et al., 2018; Stevenage et al., 2021; Walter, 1992). Since such a test generates many data points, it can provide information on the witness's behaviour, such as a potential bias towards positive or negative judgements. NB that following Wells' classification,

the additional test would also create new system variables, as the investigators can control the make-up of the screening test. For instance, event variables, such as background noise or the medium through which a voice was heard, can be considered when creating the stimuli for a screening test.

Figure 5: Integration of a screening test in the heuristic process

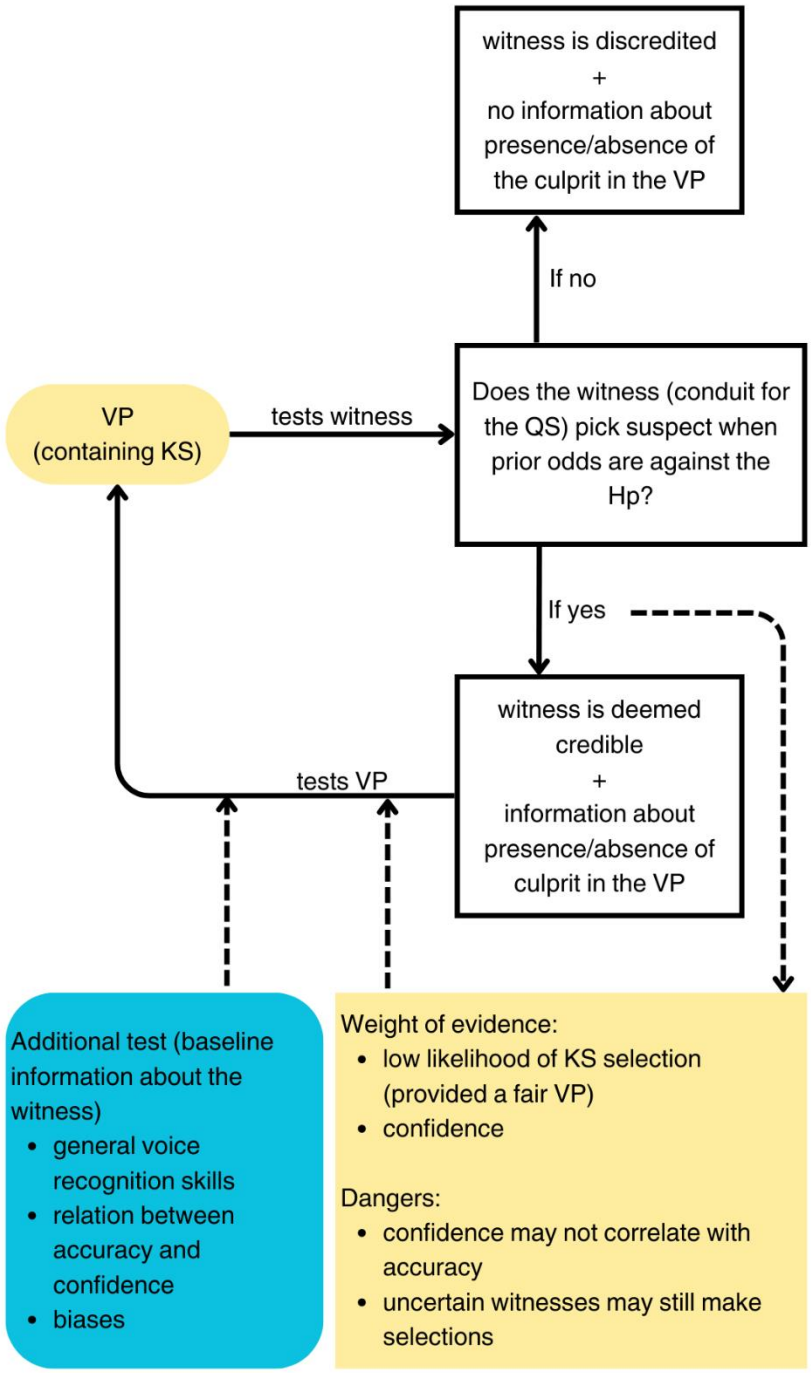


Figure 5 shows how a screening test can be integrated in the existing process for eliciting earwitness testimony. The diagram shows that the twofold aim of a VP leads to circular reasoning in the heuristic process. In a first step, the VP tests the witness' credibility by finding out whether the he or she is able to pick the suspect (KS) from a line-up, when the odds are stacked against a correct individualisation for listeners who have not heard the perpetrator. If the listener is able to do so, he or she has themself functioned as a test to determine whether both the KS and the QS were produced by the suspect. This process is answering a further question of the investigators, i.e. whether the parade was a culprit-present or culprit-absent parade. The result is a closed loop in the heuristic process, in which the witness' assessment of the VP is dependent on the VP's assessment of the witness. A complementary test (blue) will not break this loop. However, it can add further evidence about the witness by establishing base-line information about their general voice processing capabilities and behaviour in individualisation tasks. Crucially, as the diagram shows, this test is independent and only feeding into the heuristic process, rather than being dependent on it.

3.6 Conclusion: The earwitness and the legal system

This chapter has characterised the role of the witness as a source of information in legal proceedings. It was demonstrated that comparisons between expert witnesses and earwitnesses, which are often found in the literature, have to be made with caution, as these types of testimony are borne out of entirely different scenarios. A general inferiority of earwitness testimony does therefore not apply. In fact, earwitnesses are the only types of witnesses who are capable of individualisations, provided that they have semantic information about the perpetrator. Therefore, the credibility of earwitness testimony should be assessed based on the constellation of variables in the particular case. One of these variables is the witness.

To this end, investigating witness variables, a subgroup of estimator variables may prove a fruitful strand of research. A screening test may help to include the insight generated by such research into the process of earwitness evaluation. It would allow the trier of fact to gather baseline information about the witness who cannot be observed making accurate individualisations in a VP. Such a test would furthermore make the outcome of a VP less attackable because one aspect of its double function, i.e. testing the witness, would be assisted by a second test, helping a trier of fact assess the weight of the VP evidence.

4. The psychological bases of voice processing

The previous chapters described the information comprised in the speech signal, as it can in theory be used for the purpose of speaker individualisation (Chapter 2). Moreover, it was laid out how such information is elicited from an earwitness on the context of legal proceedings (Chapter 3), where the witness acts as a (more or less reliable) conduit for the perpetrator's voice, and also as the individual comparing that voice to newly presented stimuli with the aim of individualising the perpetrator.

The witness is an important but not well-understood variable in legal individualisation procedures. Three cognitive witness variables will be investigated in the empirical part of this thesis: individual differences in voice discrimination, individual differences in voice memorisation, as well as individual differences in task awareness. Therefore, the last systematic step in establishing a theoretical background for the conducted experiments is a characterisation of voice perception from a cognitive point of view. Fortunately, questions that are crucial for understanding the role of the listener in forensic speaker individualisation tasks are congruent with research questions asked by psychologists working on voice perception, e.g.:

- What are the differences between familiar and unfamiliar voice processing?
- What are the commonalities and differences between face and voice processing?
- How does voice perception change when listeners are exposed to voices and faces at the same time?
- What predictions can be made about voice recognition after a delay?

Psychological research may thus complement linguistic research by drawing important connections between listener behaviour and the underlying neurological and cognitive principles. In the subsequent sections, a literature review will be conducted, addressing the above-mentioned questions. Where possible, parallels will be drawn between psychological and linguistic findings.

4.1 Familiar and unfamiliar voice processing

4.1.1 Early findings on the neural substrate of voice processing

The modern (psychological) research interest in the neurological bases of voice recognition received a significant boost from the findings of a research group around Diana van Lancker[14], who conducted voice perception studies with brain-injured patients in the 1980s (Van Lancker et al., 1985; Van Lancker & Canter, 1982; Van Lancker & Kreiman, 1986, 1987).

Van Lancker and Canter (1982) examined voice and face recognition in nine right- and 21 left-hemisphere brain-damaged individuals. Their participants were presented with aural and visual stimuli created from famous voices and faces, respectively. In general, it was found that deficits in both face and voice recognition were significantly more common in participants with right-hemisphere damage than in participants with left-hemisphere damage. However, while three of the right-hemisphere damaged patients exhibited deficiencies in both face and voice recognition, two patients showed a deficiency in one of the domains only. There was consequently evidence for the phenomena being neurologically distinct, which led to the coinage “phonagnosia" to describe selective deficits in recognising familiar voices (following right hemisphere damage). The term was chosen in analogy to the term “prosopagnosia”, which had already been an established term to denote face processing deficits (Stevenage, 2018: 165).

Van Lancker and Kreiman (1987) categorised voice processing deficits in greater detail by investigating whether the familiarity with a speaker had an impact on the voice processing capabilities of 45 patients with brain damage. The patients were subjected to two types of tasks. Firstly, a voice recognition task, in which they listened to the voices of famous people and later tried to pick a photograph of the speaker from a four-choice visual display. This task can be classified as an individualisation task according to the terminology established in Chapter 3, since, in addition to recognising the voices, listeners had to link the voice to semantic information about the speaker. In the second task, listeners heard pairs of sentences produced by unfamiliar speakers and were asked to judge if they were spoken by the same or different speakers; i.e. a discrimination task. A control group of 48 neurotypical participants completed the same tasks. Both left- and right-hemisphere damage impaired unfamiliar voice discrimination, while only right-hemisphere damage impaired familiar voice recognition. This effect could be reproduced in a later study (Van Lancker et al., 1989). Most importantly, some

[14] The same author is also cited as Diana Sidtis in this thesis.

patients showed dissociations, with good recognition of familiar voices but poor discrimination of unfamiliar ones, or vice versa, indicating that the abilities have different neural substrates.

Prior to the work of van Lancker and her colleagues, research on human voice recognition had been limited in several ways. Schweinberger and Zäske (2018) identify three main strands of research that had shaped the academic discourse about voice perception before van Lancker's description of phonagnosia. First, early studies focused heavily on applied forensic goals of speaker individualisation, albeit without being strongly grounded in theory (e.g. Abberton & Fourcin, 1978; Hollien et al., 1982; McGehee, 1937; Rathborn et al., 1981; Saslove & Yarmey, 1980). Drawing on the insight from Section 3.1, most of this research can be classified as an investigation of event variables, i.e. the variables that describe the characteristics of the speaker and the circumstances of the exposure of the voice, rather than variables describing the characteristics of the listener (witness variables) or the variables describing how information is elicited from a listener (system variables). Second, developmental studies investigated the voice recognition ability of infants (e.g. DeCasper & Fifer, 1980), but lacked cognitive models. Lastly, acoustic studies aimed to define the physical voice parameters that signal identity (Bricker & Pruzansky, 1966; Pollack et al., 1954; Pollack & Ficks, 1954), i.e. the vocal identifiers discussed in Chapter 2, but failed to find consistent diagnostic cues.

Summing up, the advent of neurological research into speaker identity processing was deeply intertwined with the question of the relationship between speaker and listener. The potential dissociation between familiar voice recognition and unfamiliar voice discrimination is highly relevant for earwitness research, as it has already been shown that earwitnesses who are familiar with the perpetrator are the only type of witness capable of speaker individualisation. These witnesses likely provide a judgement that is neurologically different from witnesses who are not familiar with the perpetrator's voice. However, there is a certain asymmetry in van Lancker's choice of stimuli, in that the voice processing tasks compared by her team did not only differ in the familiarity with the speaker, but also in the inclusion of long-term memory processes, which were addressed by the familiar voice recognition task, but not by the voice discrimination task. This means that on the basis of her research alone, no predictions can be made regarding unfamiliar voice recognition, i.e. an unfamiliar earwitness's task. At the same time, her results indicate that there may be some neurological commonalities between the unfamiliar witness's task and an expert witness's task, as these two types of listeners need to be able to discriminate between unfamiliar voices. While the expert's task ends

with the discrimination of samples, the unfamiliar earwitness additionally requires long-term memory storage.

4.1.2 Distinguishing between different types of familiarity

4.1.2.1 Famous voices

Early psychological studies on voice processing often used a particular type of stimulus to study familiar voice recognition, i.e. "recordings of famous political and entertainment personalities" (Van Lancker & Canter, 1982: 185). While such material is readily available, it has limited implications for forensic applications, since there are not many criminal cases, in which a perpetrator's voice may be familiar to the earwitness on this basis. Moreover, it is difficult to assess whether different listeners have had the same amount of exposure to a famous voice. Therefore, if familiarity with a speaker is a continuum, rather than a dichotomy, it is difficult to select appropriate famous voice stimuli for recognition/individualisation studies. Nonetheless, further research that has picked up this type of stimuli is summarised briefly in this section. Note that most of these studies lacked a control for participants' degree of familiarity with the famous speakers featured in experiments.

Neuner and Schweinberger (2000) examined 36 patients with unilateral brain damage. Three of them showed deficits in recognizing celebrity voices but intact unfamiliar voice discrimination; other patients showed the opposite effect. This result supports the findings of van Lancker and Kreiman (1987) and provides compelling evidence for a double dissociation of unfamiliar voice discrimination and famous voice recognition. Further evidence for such a dissociation is provided by Peretz et al. (1994) who report on two case studies of "amusic" patients (GL and CN) who are unable to recognise melodies as a result of bilateral temporal lobe lesions. Patient GL was unable to recognise previously familiar voices, yet maintained the ability to discriminate between unfamiliar voices. In contrast, patient CN could neither recognise familiar voices nor discriminate between unfamiliar voices. Both patients were able to recognise faces, reconfirming that a deficit in voice recognition does not affect other modalities of person recognition.

With regard to the dissociation of famous voice and famous face recognition, several studies indicate that faces are stronger cues to person identity than voices. In a series of experiments conducted by Hanley et al. (1998), participants were presented with faces or voices of celebrities and asked to name them. Participants showed significantly poorer recognition

accuracy for celebrity voices compared to celebrity faces. The voice recognition task yielded more "familiar only" experiences, where participants could not identify the celebrity but stated that the voice sounded familiar, i.e. a recognition without subsequent individualisation. The study also found that when participants recognised a face as familiar but could not name the person, they were able to remember more details about the person when they heard the voice, as compared to just seeing the face again. This finding may have interesting implications for research on the reliability of witnesses who were exposed to both the voice and the face of a perpetrator. In contrast to the studies discussed in this chapter so far, Hanley et al. (1998) recruited neurotypical participants, making their results more ecologically valid for forensic applications.

Unsurprisingly, studies have found that the accuracy with which famous voices are recognised correlates positively with the duration of the provided voice samples. A study by Schweinberger et al. (1997) exposed participants to famous and unfamiliar voice samples of different durations (0.25 s to 2 s). Participants were asked to indicate if the speaker was famous and, when this was the case, to provide the speaker's name. Voice recognition improved as the sample duration increased, with the biggest gains in the first second. Voice recognition did not always result in individualisation. When participants could not name the speaker, different cues were provided to facilitate access to semantic information about the speaker, such as a second voice sample, the speaker's occupation, or their initials. Initials proved to be the most effective cue to speaker identity. A question that cannot be answered by this setup is whether an increase in stimulus duration necessarily leads to better recognition, or whether stimulus complexity – which usually correlates with duration – is the driving factor.

Njie et al. (2022) combined a famous voice recognition task with an accent recognition task. For this purpose, the authors conducted a voice sorting experiment using recordings of two characters from the TV show "Derry Girls", who speak with a Northern Irish accent. The participant pool consisted of 126 lay listeners from Northern Ireland and England, aged 18 to 40. The participants self-reported their familiarity with the talkers and the Northern Irish accent, and formed four categories accordingly: Participants from Northern Ireland who had watched Derry Girls (N = 32), participants from England who had watched Derry Girls (N = 29), participants from Northern Ireland who had not watched Derry Girls (N = 32), and participants from England who had not watched Derry Girls (N = 33). During the sorting task, the participants were presented with a PowerPoint slide containing 30 numbered boxes, which were movable and played a voice stimulus when participants clicked on them. The two characters

from Derry Girls were represented with 15 recordings each. The participants, who were unaware of the total number of speaker identities, were asked to sort the boxes into clusters based on perceived voice identity. Both talker and accent familiarity improved performance on the voice sorting task. However, the benefits of talker familiarity were larger and more consistent than accent familiarity effects. Moreover, accent familiarity only significantly improved performance for listeners already familiar with the talkers, suggesting a hierarchy where talker familiarity is more fundamental and accent familiarity can provide additional benefits.

In summary, famous voices are convenient stimuli for the study of familiar voice processing and benefit from the key advantage that the same set of speakers may be familiar to unrelated participants. At the same time these voices lack the option of differentiating between different degrees of familiarity. Personally familiar voices may therefore be a more ecologically valid benchmark for real-world voice recognition tasks, such as forensic applications, especially when describing the neurotypical population.

4.1.2.2 Personally familiar voices

A category of familiar voices that is more relevant for forensic research is that of personally familiar voices, also referred to as "familiar-intimate" as opposed to "familiar-famous" voices (van Lancker et al., 1983). It is hard to assess the number of familiar-intimate voices that listeners can store. Since some estimates suggest that the typical size of an individual's social network ranges from 100 to 300 individuals (Hill & Dunbar, 2003; McCarty et al., 2001), Watt (2010: 78) hypothesises that the ability to remember personally familiar voices may extend to at least three digits. On the other hand, Kreiman and Sidtis (2011: 156) stress that psychological studies have not yet found an upper limit of familiar-intimate voice memory.

Voice processing studies have painted a diverse picture of listeners' ability to individualise speakers with whom they are in a close personal relationship. Some authors have, for instance, pointed out that even non-verbal vocalisations, such as laughter, breathing noises, or coughs may provide a familiar listener with enough vocal tract information for a speaker individualisation in some circumstances (Trouvain, 2014: 598; Trouvain & Truong, 2012: 39).

At the same time, listeners may in certain contexts not be capable of individualising highly familiar speakers. An often-cited example for this phenomenon is a one-person experiment, in which the phonetician Peter Ladefoged was the only participant (Ladefoged & Ladefoged, 1980). Ladefoged's wife Jenny presented him with voice stimuli from 29 speakers

with whom he was very familiar, eleven strangers, as well as 13 speakers whose voice he had only heard one or two times. Three types of stimuli were produced by the speakers: A recording of the word "hello", a one-sentence description of a picture, and a 30-second-long description of the same picture in spontaneous speech. Ladefoged listened to all three stimuli produced by each speaker, the aim being the individualisation of the known speakers. Several authors who have reported on the experiment, claimed that Ladefoged had been unable to "recognise" his own mother's voice in the "hello"- and the sentence conditions (Atkinson, 2015: 36; Stevenage, 2018a: 634; Watt, 2010: 79-80). In fact, however, Ladefoged described his mother's voice as that of a "familiar low-pitched woman" in these conditions (Ladefoged & Ladefoged, 1980: 49). Consequently, he *recognised* the voice, i.e. he had a familiar-only experience, while he was not capable of an *individualisation* (cf. Section 3.1). This is a crucial distinction in the context of earwitness testimony, as it has already been shown that VPs elicit a context-specific recognition rather than an ironclad individualisation of the perpetrator. Moreover, Ladefoged was eventually able to establish a semantic link to his mother after listening to the 30-second-long stimulus. Authors often isolate the example of Ladefoged's mother's voice from the rest of the experiment to demonstrate the fallibility of familiar voice processing. However, the most important conclusion to be drawn is that stimulus quality and duration are crucial variables in familiar speaker individualisations. Overall, Ladefoged recognised 31% of the familiar speakers in the "hello" condition, 66% in the single-sentence condition, and 83% from 30-seconds of speech (24/29 speakers). Note that these are individualisations, i.e. Ladefoged establishing the speaker's identities, while the number of correct recognitions, which was not reported, might have been higher (given the example of his mother's voice).

A central challenge of investigating personally familiar voice processing is the recruitment of participants who are personally familiar with the same speakers in a relatively uniform manner. For this reason, pre-existing social networks have been a popular object of research (e.g. Foulkes & Barron, 2000; Rose & Duncan, 1995; Skuk & Schweinberger, 2013; Smith et al., 2017).

Foulkes and Barron (2000) investigated how well a group of ten male university friends, aged 20 – 21, could recognise each other's voices in an open speaker recognition test using telephone speech samples. At the time of the study, the group had known each other well for about 21 months. They were living in three separate but geographically close houses and were socialising with each other regularly. Before that, all ten participants had lived in shared student accommodation for about a year. Most group members had similar southern English accents,

with two outliers who had a Tyneside accent and a non-standard London accent, respectively. One speaker had an occasional stammer. Each group member recorded a scripted message, which comprised 42 words and was rehearsed to sound natural. Two foil speakers were included as well, so that the entire listening material consisted of twelve recordings. Nine members of the group took part in the subsequent listening test, during which all listeners were presented with all twelve recordings, not knowing the ratio of familiar speakers and unfamiliar speakers in the material. The overall success rate for speaker individualisations was 67.8% (61 out of 90 familiar voices). Four listeners made at least one incorrect identification and seven listeners failed to identify at least one voice, with one listener even failing to identify his own. The two foil voices were wrongly attributed to network members three times, which is a relevant finding before a forensic background. The distinctiveness of the voices was a reliable predictor for identifiability: Speakers with average pitch were harder to identify than those with more extreme pitch. Similarly, the two speakers with distinct regional accents were correctly identified by all listeners.

Skuk and Schweinberger (2013) performed a study with a similar existing network of 60 German high school classmates. Forty (20 female, 20 male) of them participated in a voice individualisation test for which the remaining 20 (10 female, 10 male) had produced the voice samples. Each of the speakers recorded four different types of stimuli: a sentence, two vowel-consonant-vowel (VCV) syllables (/aba/ and /igi/), the word “hallo” (hello), and a non-verbal vocalisation (throat clearing sounds). Participants were not only asked to name the speaker, but also to rate the voice’s distinctiveness. The results showed that speakers were best identified from full sentences, followed by syllables, words, and lastly non-verbal sounds. Still, even the non-verbal vocalisations were recognised above chance level. Substantial individual differences in voice recognition ability were found between the students. Male listeners showed an own-gender bias, recognising male voices better than female voices. Female listeners did not show this difference, identifying both genders equally well. Overall, female listeners outperformed males, and male speakers were individualised more often than female speakers. The latter phenomenon may be due to the fact that, on average, male voices were also rated as more distinctive. The study thus validates the two systematic effects that Foulkes and Barron (2000) had observed on a smaller dataset, i.e. correlations between stimulus duration and individualisation accuracy, as well as between voice distinctiveness and individualisation accuracy. While the authors additionally observed an own-gender bias in male participants, there is not enough information provided on the group dynamics to rule out that the better

performance of these listeners with speakers of the same sex is a result of a higher degree of familiarity.

As discussed in Chapter 2, not all utterances produced by a speaker are equally representative of that speaker's vocal identity, in that short-term within-speaker variability may result in marked differences between an individual utterance and the speaker's average. Several studies have therefore tested the robustness of familiar speaker individualisation by addressing within-speaker variability to a larger extent. For instance, Smith et al. (2017) performed a study in which members of a social network (colleagues) were tested for their ability to individualise each other on a vocal basis, while some of the stimuli were produced in whispered speech, i.e. employing a deliberate disguise. The social network consisted of eleven women, including the first author, who worked together at a cosmetics store in York. Six of the women, including the first author, were recorded reading three short texts, each in modal voice and whisper. Additionally, three foil speakers were recorded, including one male speaker. Three of the recorded group members had a Yorkshire accent, while the remaining three participants recorded had accents from different parts of the UK (Leicester, Southwest, and Edinburgh). The foil speakers were from London (male), the Southwest and the Northwest. Based on the recordings, three distinct individualisation tests were constructed, using short (4 syllables) whispered stimuli, short (4 syllables) modal voice stimuli, and long (16 syllables) whispered stimuli, respectively. It is not clear why the authors did not opt for a 2x2 factorial design, i.e. why no test with long modal voice samples was created. Each test set contained samples of all speakers in randomised order. Identification accuracy was, as predicted, lowest with short whispered samples (64%) and highest with short normal samples (93%). Long whispered stimuli elicited an intermediate average accuracy (87%). Note, however, that the average accuracy scores also correlated with the order in which the tests were presented, which was identical for every participant. It is therefore possible that the increasing accuracy scores result from a training effect due to the listeners' increasing familiarity with the voices over the course of the test battery. With regard to listener errors, the study found that listeners had difficulty rejecting the unfamiliar female foil speakers, incorrectly recognising those voices as familiar 26-36% of the time. In contrast, the male foil speaker was almost always correctly identified as an unfamiliar speaker. Overall, the foil samples were more often falsely recognised as familiar voices than the familiar in-group speakers were incorrectly rejected by listeners.

Overall, the research reviewed in this section demonstrates that familiar-intimate voices may be highly recognisable to members of a shared social network. However, recognition

accuracy may vary drastically as a result of within-speaker variability, stimulus length, distinctiveness of the target voice, and the degree of familiarity between speaker and listener. All these factors must be considered in applying findings forensically.

4.1.2.3 Unfamiliar voices

As noted, the category of voice processing that is of highest relevance for earwitness testimony is unfamiliar voice processing, given the number of crimes in which a witness (who may be identical with the victim) and the perpetrator do not know each other. This can, for instance, be the case in a masked robbery or in cases of large-scale communications fraud. Early studies have generally concluded that unfamiliar voice processing is significantly less reliable than familiar voice processing (e.g. Abberton & Fourcin, 1978; Hecker, 1971; Hollien et al., 1982; Pollack et al., 1954). Recent psychological studies tend to agree with this general assessment (e.g. Young et al., 2020: 403, Schweinberger & Zäske 2018: 544)[15], but also show that some of the regularities applying to familiar voice recognition translate to unfamiliar voice recognition and/or discrimination. Above all, voice distinctiveness and stimulus duration/complexity have demonstrated fundamental correlations with recognition accuracy. Several studies indicate that unfamiliar voices are more readily told apart, the more distinctive they are (Latinus et al., 2013; Lavner et al., 2001; Mullennix et al., 2009, 2011; Sørensen, 2012; Yarmey, 1991). Moreover, studies consistently find that stimulus duration correlates positively with recognition accuracy (e.g. Cook & Wilding, 1997; Roebuck & Wilding, 1993; Rose & Duncan, 1995; Schweinberger et al., 1997; Yarmey & Matthys, 1992). A notable exception are the findings by Pautz et al. (2023), which were discussed in Section 3.4.2.1 in the context of system variables for VPs.

In addition to these long-standing research questions, new strands of psychological research have paved the way for a more comprehensive investigation of a listener's behaviour when being confronted with unfamiliar voices, i.e. going beyond questions of speaker identity recognition. For instance, recent studies stress that listeners can extract a rich multivariate set of physical, social, and psychological information from unfamiliar voices, going far beyond identity recognition (e.g. Lavan & McGettigan, 2023). Lavan (2023) exposed 294 British participants to recordings of six female speakers with a Standard Southern British English (SSBE) accent. The stimuli comprised scripted neutral sentences and had an average duration

[15] The same phenomenon can be observed in the processing of faces. Several studies that conducted face sorting experiments found that participants were better at telling together or telling apart photographs of familiar people than photographs of unfamiliar people (Andrews et al., 2015; Jenkins et al., 2011; Zhou & Mondloch, 2016).

of two minutes. In a free description task listeners wrote down any words or short phrases that described the person speaking, after listening to each stimulus. On average, participants listed nine to ten descriptors per voice, which fell into four different categories: physical characteristics (e.g. age, gender, ethnicity), psychological characteristics (e.g. personality traits, emotions), social characteristics (e.g. social class, education level), and stimulus characteristics (e.g. descriptions of vocal features). The highest proportion of descriptors denoted psychological traits (44.3%), followed by physical and stimulus descriptors (15.5% each), and social descriptors (15.1%). The remaining 9.6% comprised miscellaneous descriptors that did not clearly match one of the categories. Although occurring less frequently than psychological descriptors, physical descriptors like age and sex were produced earlier than other characteristics by most participants. In a later article, Lavan and McGettigan (2023) presented a "word cloud" figure based on the descriptors elicited in Lavan's (2023) free description task (cf. Figure 6); the font size correlates with the frequency with which the descriptors were produced.

Figure 6: Free descriptions of unfamiliar female voices (Lavan & McGettigan, 2023: 4)

Overall, the study shows that various person characteristics are inferred rapidly and spontaneously from unfamiliar voices and that different listeners infer similar speaker characteristics. Unfortunately, the study did not directly discuss the accuracy of the elicited descriptors (i.e. to what extent they match the speaker), which would have interesting implications for forensic research – especially in the case of vocal descriptors.

Several studies indicate that listeners can create "profiles" of the kind depicted in Figure 6 when listening to unfamiliar voices (Lavan, 2023; Lavan et al., 2021; Mahrholz et al.,

2018; McAleer et al., 2014; Mileva & Lavan, 2023). These perceptions of person-specific traits may especially apply to unfamiliar voice perception, when listeners form a first impression of these voices (Lavan et al., 2021: 282). This does not, of course, mean that trait perceptions are necessarily completely excluded from familiar voice perception (cf. Section 4.2.3.2). In line with the assumption that trait perceptions coincide with the formation of first impressions, there has been great research interest in the speed with which such trait judgements are made.

McAleer et al. (2014) therefore used short stimuli of the word "hello" in their study on trait perception from voices, the reason being that – as a conversation opener – the word "hello" is typically found in scenarios where listeners encounter a novel voice for the first time. In total, 64 brief "hello" voice samples (32 male, 32 female, average duration 390 ms) were used as stimuli. All speakers were from Scotland and undergraduate students at the University of Glasgow. A total of 320 listeners (117 male, average age 28.5 years) from the same undergraduate population rated the stimuli. In contrast to Lavan's (2023) study, listeners did not engage in an open description task, but rated the speakers' traits on a 9-point-scale along ten pre-defined dimensions: aggressiveness, attractiveness, competence, confidence, dominance, femininity, likeability, masculinity, trustworthiness, and warmth. Each listener only rated one of these dimensions to avoid halo effects[16]; the number of raters per trait ranged from 24 to 36. Overall, the listeners showed high between-rater agreement in judging the ten personality characteristics, with Cronbach's alpha values greater than .88 for all traits. This suggests that trait impressions can be formed rapidly and that different listeners form similar impressions when listening to the same stimulus.

From a sociolinguistic point of view, some clear limitations of the study setup can be identified. On the one hand, speakers and listeners were sourced from the same, highly specific demographic, i.e. Scottish undergraduate students. As discussed in Section 2.2.2, Social Identity Theory (SIT) indicates that these speakers and listeners are likely to self-identify as members of common social groups. The listeners might therefore be biased in that they have positive social stereotypes for in-group speakers to begin with, which may at the same time lead to greater between-rater agreement. On the other hand, the study used very short stimuli. It was shown in Section 2.1 that an individual utterance can only represent a certain proportion of the speaker's vocal identity. While some studies suggest that speakers want to convey more identity-specific information in the opening parts of a conversation (cf. Section 2.2.1.1;

[16] An effect where a positive assessment of one trait leads to an overall positive impression of a person, which in turn affects the perception of other traits.

Bradshaw et al., 2022), a single word stimulus will necessarily contain a highly limited number of cultural and habitual vocal identifiers. A more recent study by Mahrholz et al. (2018) has addressed this shortcoming by eliciting trait judgements from single words as well as sentences; They found moderate to strong positive correlations between ratings of words and sentences for the same speaker, indicating that the rapidly formed initial trait judgement may indeed be stable. However, their study used a similar homogeneous speaker and listener pool of Scottish undergraduate students.

In sum, more research is needed to explore the accuracy, variability and stability of trait judgements from voices. Nonetheless, the perception of speaker characteristics has important implications for unfamiliar voice processing and the reliability of earwitness testimony in particular. The following questions are highly relevant in this connection:

- To what extent are vocal trait ratings influenced by in-group and out-group constellations between speaker and listener?
- Are vocal trait assessments performed by listeners on a regular basis or are they restricted to a laboratory setting, in which they are prompted to do so?
- Do listeners compare vocal trait profiles when comparing voices for speaker identity or is identity processing cognitively and neurologically distinct?
- If the processes are distinct, is trait processing occurring at the cost of identity processing or complementing identity processing?

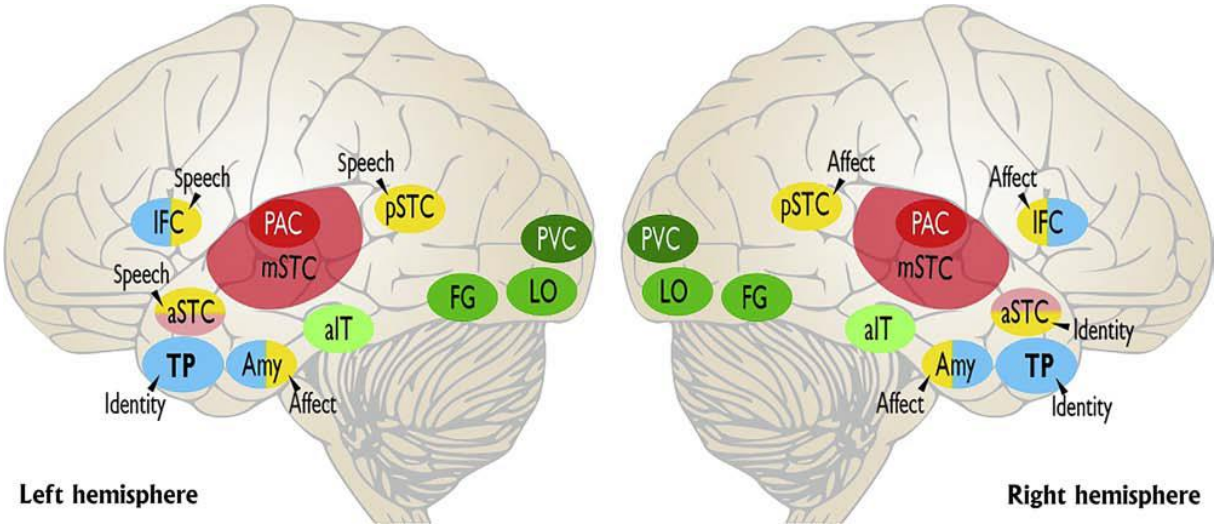
Some of these questions are addressed by a recent voice perception model (Lavan & McGettigan, 2023), which will be discussed in Section 4.2.3.

4.2 Anatomical and neurological foundations of voice processing

4.2.1 Voice-selective brain areas

It was shown in Section 4.1.1 that early psychological research into voice processing capabilities made significant advances by establishing links between certain types of brain damage and specific voice processing tasks. In this connection, very specific processing deficits could be identified, depending on speaker familiarity. More recently, neuroimaging studies have led to a more nuanced understanding of the neural bases of voice processing. These studies have complemented earlier findings with information about the neurotypical population. The schematic illustration below visualises individual brain areas that are relevant for person perception, including both voice and face perception. The assumed functions of these brain areas will be discussed in the remainder of this section.

Figure 7: Voice-selective and face-selective brain areas (Young et al., 2020: 401)



Legend: a/m/pSTC, anterior/mid/posterior superior temporal cortex; aIT, anterior inferior temporal lobe; Amy, amygdala; FG, fusiform gyrus; IFC, inferior frontal cortex; LO, lateral occipital cortex; MFC, medial frontal cortex; PAC, primary auditory cortex; PVC, primary visual cortex; TP, temporal pole.

Using functional magnetic resonance imaging (fMRI), Belin et al. (2000) identified the middle and anterior parts of the superior temporal sulcus (STS) – bilaterally – as voice-selective areas. This means that these cortical regions, which the authors refer to as "Temporal Voice Areas" (TVAs), showed greater neuronal activity when subjects were exposed to vocal sounds than to non-vocal sounds. The STS is the sulcus that separates the superior temporal gyrus from the

middle temporal gyrus within the temporal lobe of the brain. Both the STS and the adjacent superior temporal gyrus make up the superior temporal cortex (STC; Young et al., 2020: 400).

TVAs are regarded as the functional equivalent of the fusiform gyri (bilateral), which had already been established as the main locus of face perception (Schweinberger & Zäske, 2018: 548). While there is clear evidence that the TVAs of both hemispheres contribute to voice processing (Kreiman & Sidtis, 2011: 230), some task-specific distinctions can be made. For instance, some neuroimaging studies suggest that the hemispheres cover different tasks with regard to the type of information that is extracted from a voice. Results from Nakamura et al. (2001), who conducted positron emission tomography scans while subjecting participants to various speech processing tasks, observed that the right STC predominantly responded to questions about speaker identity, while the left STC responded to a greater extent to questions about speech content. Similar findings were obtained in a magnetoencephalography study by Schall et al. (2015). Further evidence for this lateralisation effect is provided by Belin et al. (2002) who observed in an fMRI study that the right STS is also activated when listeners are exposed to vocalisations that do not contain speech sounds, such as laughter, sighs, and cries. While the preceding discussion focuses only on the most influential findings on lateralisation, a more detailed discussion, that also discusses some contradicting findings, is provided by Gainotti (2013).

Studies indicate that there are different centres of voice identity processing within the right STC, depending on the listener's familiarity with the speaker. A first indication that this might be the case was provided by von Kriegstein et al. (2003) who conducted an fMRI study with 24 familiarised voices. This means that the voices were neither familiar-famous nor familiar-intimate to the listeners. Instead, familiarity was achieved through training, i.e. by means of repeated exposure to newly introduced voices. Listeners then performed different tasks that focused on either speaker identity processing or speech content processing. When participants focused on the recognition of these trained-to-familiar speakers, there was increased activation in the right anterior STC, suggesting that this area (particularly the right anterior STS) is involved in familiar voice processing specifically. The results further underpinned left hemisphere dominance when processing speech content. Other studies have shown greater involvement of right posterior STC areas for unfamiliar voices (Belin et al., 2000; Lattner et al., 2005; Warren et al., 2006).

Kriegstein & Giraud (2004) directly compared the neural substrates of familiar-intimate and unfamiliar voice and speech recognition. The results showed that the voice recognition task

activated the anterior and posterior STC to a greater extent than the speech recognition task, with right-hemisphere dominance. Unfamiliar voices resulted in greater posterior STC activity than familiar voices, which elicited greater mid to anterior STC activity. Consequently, the findings led to the hypothesis that increasing familiarity with a voice follows posterior-anterior axis along the (right) STC.

Further studies have since reached similar conclusions (Kriegstein et al., 2005; Latinus et al., 2011; Warren et al., 2006). Perhaps most importantly, an fMRI study that used trained-to-familiar voices observed changes in brain activity as voices became more familiar over time (Belin & Zatorre, 2003). Participants were exposed to unfamiliar voices over five days by means of passive listening. On the fifth day, they underwent an fMRI scan while listening to the now familiarised voices as well as to novel voices. Results showed reduced activation in the right anterior STC in response to the familiarized voices compared to novel voices, providing further support for a posterior-anterior shift. Following a similar approach, Latinus et al. (2011) observed that the mid STC showed some activation irrespective of degree of familiarity with a speaker. Moreover, their results indicate that some areas outside of the TVAs are involved in speaker identity processing, including the right superior temporal pole (TP) in the case of unfamiliar stimuli, and the left temporal pole and bilateral inferior frontal cortex (IFC) in the case of familiar voices. Young et al. (2020) have pointed out that these areas are likely to operate at a post-perceptual level, which means that rather than being involved in the recognition or discrimination of a voice, these areas try to establish links to episodic memory of an encounter or semantic information about the speaker.

In summary, neurological studies have provided evidence to support van Lancker's original assumption that unfamiliar and familiar speaker processing are neurologically distinct.[17] While some areas, like the right mid STC respond to both familiar and unfamiliar speakers' voices, the right anterior STC appears to be the locus of familiar voice recognition, while the left TP and both IFCs may provide additional semantic information that is necessary for individualisation. The right posterior STC on the other hand appears to be predominantly activated by unfamiliar voices. Familiar and unfamiliar earwitnesses may therefore not only differ in the kind of judgement that they can provide (individualisation vs recognition), but to some extent perform neurologically distinct tasks. At the same time, it was shown that

[17] Note that this does not necessarily mean that they are also functionally distinct, as will be discussed Section 4.2.3.2.

familiarity can be established by means of training. This information has several implications for earwitness testimony:

1. It has implications for the number and duration of stimuli in a VP in that a situation must be avoided where listeners achieve a trained familiarity with all voices in the parade, including the foils. This would result in a sensation of speaker familiarity for the wrong reason. A setup mitigating this risk is a sequential parade.
2. For the same reason of creating speaker familiarity for the wrong reason, these findings highlight the danger of exposing a witness to the suspect's voice, e.g. in the manner of a direct confrontation, before a VP is conducted.
3. Differentiation may be possible between different types of unfamiliar earwitnesses, depending on the degree of trained familiarity with the perpetrator's voice established during the crime. The testimony of witnesses who familiarise themselves with the perpetrator's voice to a great extent may neurologically be more similar to the testimony of earwitnesses who are personally familiar with the perpetrator. This familiarity might be achieved by means of exposure duration. More importantly, certain listeners might be more susceptible to speaker identity training, either because they see the need to actively memorise the voice, or because they form a representation more quickly when being exposed to a voice passively. Witness screening may be a way of establishing such a differentiation.
4. Finally, a further differentiation may be possible based on the degree of interaction with the perpetrator, as suggested by Hammersley and Read (1985; discussed in Section 3.4.2.1). The demonstrated interaction advantage may be the result of greater IFC and TP activation when listeners are exposed to the voices again, resulting in an advantage from a neurological point of view.

4.2.2 Voice and face perception

Person recognition comprises more modalities than just hearing. Most importantly, visual processing plays a leading role in identity recognition. In fact, neurological research into face processing predates the equivalent research into voices. Many cognitive models of voice perception – which will be discussed in the subsequent section – are therefore adaptations of earlier models that were originally designed for face processing. There has also been a growing research interest in the possible crossmodalities between face and voice processing. Some necessary information on the neural underpinning of face perception will therefore be

introduced in this section. Since the focus of this thesis is on voice perception, face processing will only be covered briefly. Most of the information in this section is sourced from a recent synopsis by Young et al. (2020), which is also the source of Figure 7. Note that in addition to the already discussed voice-specific brain areas (red), Figure 7 also shows the face-specific brain areas (green). Yellow labels denote areas that are assumed to play a role in both face and voice processing, while blue labels indicate that areas are likely involved in post-perceptual processes.

Young et al. (2020: 401) suggest that incoming signals are first processed in a basic manner in the primary cortices for both modalities, i.e. the primary auditory cortex (PAC) for voices and the primary visual cortex (PVC) for faces. Subsequently, signals are processed in the respective unimodal regions. In the case of voices, these are the TVAs, as described in the previous section. It has already been mentioned that the fusiform gyri are likely to be the functional equivalent of the TVAs for face processing. Other areas that are unimodally responsive to faces are the occipital cortex and the anterior inferior temporal lobe. As is the case for voices, neuroimaging studies indicate that face identity processing is predominantly carried out in the right hemisphere (Gainotti, 2013: 1151). Some distinctions can be made regarding the processing of familiar and unfamiliar faces. A neuroimaging study by Collins et al. (2018) found anterior inferior temporal lobe responses when participants saw familiar faces, while unfamiliar face processing was largely restricted to the visual cortex, including the lateral occipital cortex and fusiform gyrus. In this connection, the posterior visual areas are assumed to be involved in the structural encoding of faces, while the anterior temporal and frontal areas are associated with semantic memory.

The areas involved in post-perceptual processing show significant overlaps for face and voice processing. Specifically, the temporal poles exhibited a sudden increase in activation once sufficient information was available on a familiar face. This suggests the TPs are involved in identity recognition (Collins et al., 2018).

Crossmodal integrations of face and voice processing had already been observed before the availability of neuroimaging methods. For instance, McGurk and Macdonald (1976) described a multisensory illusion based on an incongruence of visual and auditory information. They accidentally discovered that when sound recordings of certain phonemes are presented alongside a video recording of a different phoneme, listeners' perception of the heard sound changes. In most cases, the perceived sound is different from the visually presented sound as well as the auditorily presented sound. For example, when the utterance /ba/ was presented

auditorily alongside a precisely aligned video recording of the utterance /ga/, most listeners perceived /da/. Based on this multisensory illusion, amongst other factors, Calvert et al. (1998), hypothesised that face and voice processing may be cross-modally integrated at the neural level. Their work did, however, make predictions about speech processing, rather than the processing of speaker identity.

Another perception effect that speaks for a multimodal integration of face and voice perception is the so-called "face overshadowing effect". Several studies have shown that voice memorisation ability suffers when participants are presented with the face and the voice of a speaker, as opposed to the voice alone (Cook & Wilding, 2001; McAllister et al., 1993). This applied to setups in which participants were actively asked to memorise a voice as well as to setups in which participants were unaware of the task and thus memorised voices passively (McAllister et al., 1993). Whereas more studies have identified some performance differences when stimuli are presented crossmodally, not all studies agree in the directionality of the effect. For instance, Legge et al. (1984) found that voice recognition was enhanced when voices had been presented alongside the corresponding face during a study phase, as opposed to purely unimodal voice learning in study.

In summary, there is evidence for the multisensory integration of face and voice processing on both a functional and a neurological level, while neurological research also highlights significant dissociations of both modalities.

4.2.3 Cognitive voice processing models

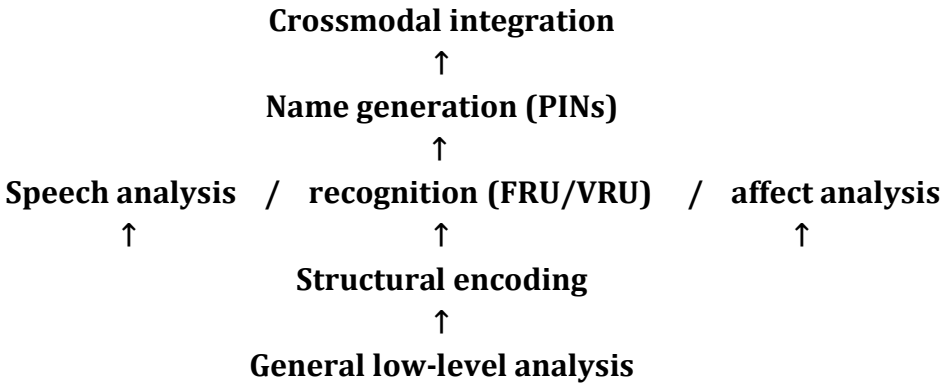
4.2.3.1 'Auditory face' models

Cognitive models for voice recognition have accounted for a cross-modal integration of voice and face processing from the beginning. Belin et al. (2004) proposed a voice recognition model which is a direct adaptation of Bruce and Young's (1986) highly influential face recognition model.

The adaptation of a face recognition model to voices was based on the analogy that in terms of perception the voice can be regarded as an "auditory face" (Belin et al., 2004: 129). This means that similar types of information can be extracted from both faces and voices. In this connection, Bruce and Young's (1986) face recognition model assumes that three broad categories of information can be extracted from faces: identity information, emotional information, and speech information. While the first two terms are to be interpreted literally, the term "speech" here refers to "facial speech", i.e. the visible mouth and face movements

during speaking, which are assumed to play a role in speech perception (as demonstrated by e.g. the McGurk effect). According to their model, the three categories of information are processed in separate routes and later integrated into a unified percept of the face (Bruce & Young, 1986: 312 f.). Similarly, Belin et al. (2004) assume that listeners extract the same broad categories of information from voices, i.e. identity information, emotional information, and speech information; the term "speech" is here referring to the propositional content of what is said by a speaker.

Figure 8: Processing hierarchy in Bruce & Young's face model and Belin et al.'s voice model



Both Bruce and Young's model and the auditory face model can be characterised as hierarchical models because they propose that the processing of faces or voices occurs in a sequence of discrete stages, from simple to complex. Figure 8 shows the individual stages in which voices and faces are analysed according to these models.[18] In a first step, a basic low-level analysis is carried out in the modality-specific primary cortices, i.e. the PVC or the PAC (Belin et al., 2004: 131). Subsequently, the face or voice is structurally encoded, which means that basic physical features are encoded to create a simplified representation, the "structural code".

Then, the three categories of information ((facial)speech, affect, identity) are processed in largely independent parallel pathways. Identity-specific information is processed through 'face recognition units' (FRUs, Bruce and Young, 1986) and 'voice recognition units' (VRUs, Belin et al., 2004), respectively. These units act as structural templates for previously

[18] The figure was created by the author of this thesis and is based on information sourced from Young et al. (2020).

encountered faces/voices against which the incoming structural code is matched. The following key properties of VRUs/FRUs can be identified:

- FRUs/VRUs are structural templates for unique sets of perceptual features (Burton et al., 1990; Yovel & Belin, 2013).
- Each FRU/VRU corresponds to one familiar face/voice (Bruce & Young, 1986; Burton et al., 1990).
- The units are built up gradually through experience with faces/voices (Bruce & Young, 1986; Ellis et al., 1997).
- They are stored in memory as enduring representations for recognition (Burton et al., 1990; Ellis et al., 1997).
- Multiple units are stored to represent all known faces/voices (Burton et al., 1990).
- Matching the input to a unit is done rapidly and automatically (Ellis et al., 1997).

In the next stage, recognised voices/faces are linked to a specific name or identity, i.e. recognition is transformed into individualisation. Identity-specific semantic details are stored in 'person identity nodes' (PINs), which are linked to the individual face/voice recognition units. When an FRU/VRU is activated by perceiving a familiar face/voice, it spreads activation to the associated PIN. Crossmodal integration is achieved as both the FRU and the VRU for a specific person feed into a common PIN. In this sense, PINs are 'supramodal' (Belin et al. 2004: 131).

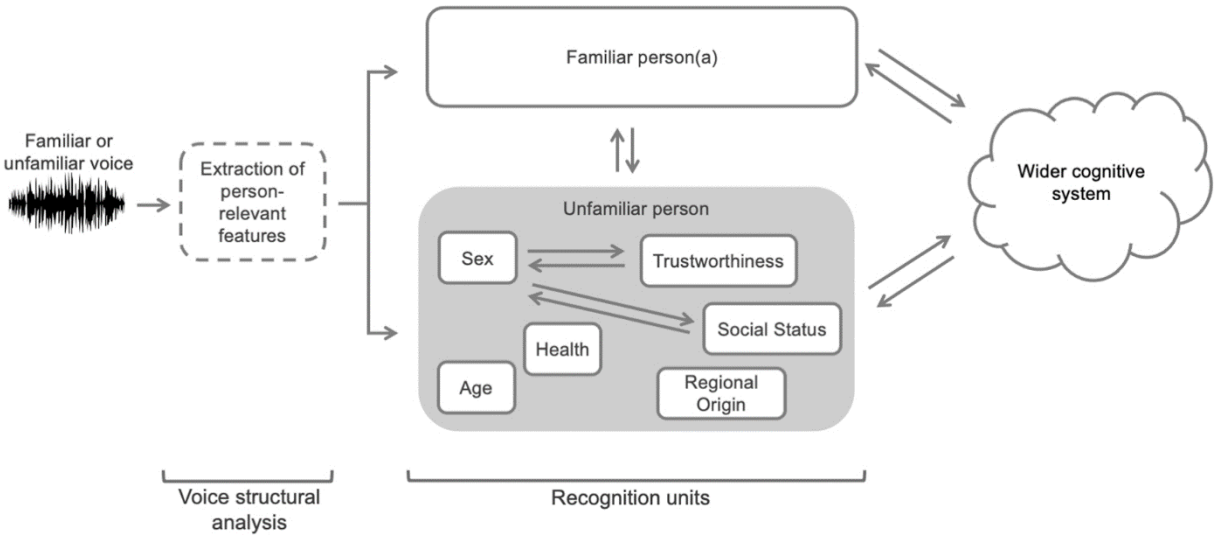
Auditory face models have been the most influential type of voice recognition models in the past twenty years. It must be stressed, however, that both Bruce and Young's (1986) face recognition model and the auditory face model (Belin et al., 2004) are situated in the realm of cognitive theory rather than in neuroscience or biology. Their primary aim was to conceptualise the mental processing involved in face perception at a cognitive level. Several efforts have therefore been made to revise the auditory face model by incorporating new insights from neuroimaging studies (e.g. Belin et al., 2011; Maguinness et al., 2018; Young et al., 2020). Some of these insights were discussed in Section 4.2.1. The most recent revision by Young et al. (2020) therefore agrees with the 'auditory face' metaphor as a functional analogy, in that similar information is extracted from faces and voices. At the same time, they stress that the largely dissociated unimodal pathways in which faces and voices are processed before post-perceptual level speak against a simple equivalence of both processes.

4.2.3.2 Person Perception from Voices (PPV) model

What all adaptations of the auditory face model have in common is that they are primarily concerned with familiar voice processing, i.e. processes that result in the recognition (in VRUs) or individualisation (in PINs) of a speaker. The key advantage of these models is their suitability for explaining familiar voice and face integration, as demonstrated in the previous section. A very recent model by Lavan and McGettigan (2023) deviates from this trend by equally addressing unfamiliar voices. Their model explains voice perception from a unimodal perspective. At the same time, it allows for a more comprehensive description of voice perception by integrating unfamiliar and familiar voice perception. It will be argued in this section that this model is of great value for earwitness research, given the importance of unfamiliar voice recognition in forensic scenarios. This also entails questions of how voices become familiar.

The authors refer to the model as the "Person Perception from Voices" (PPV) model, which stresses that – in contrast to existing models – their model is not limited to speaker recognition, but rather describes how listeners extract a broad range of person-related information from voices. This additional person-specific information has already been discussed in Section 4.1.2.3 of this thesis. It comprises a wealth of personality and physical trait judgements. This multifaceted approach goes beyond the three main categories of information (identity, affect and speech information) considered by the auditory face model and its more recent incarnations.

Figure 9: Schematic representation of the PPV model (Lavan & McGettigan, 2023: 5)



A schematic overview of Lavan & McGettigan's model is provided in Figure 9, which is an unaltered illustration from the authors' original publication. While it is not a rigid hierarchical model like Belin et al.'s model, a rough sequential structure can be identified in the form of two consecutive steps (cf. the horizontal dimension of the diagram). The first step, which is referred to as a "voice structural analysis" is not different from existing models. It describes the translation of the voice into a structural code (cf. Section 4.2.3.1). In fact, the authors do not describe this step themselves and instead refer to the equivalent step in the auditory face model (Lavan & McGettigan, 2023: 6). The next step, which is labelled "recognition units" in the diagram is functionally equivalent to the VRUs of the auditory face model, but cognitively different, as it converges multiple scenarios. The first scenario is a situation in which a voice is immediately recognised as that of a familiar person. In this case, a 'recognition unit' for that specific voice is activated "without any intermediate processing steps" (Lavan & McGettigan, 2023: 5). Crucially, the authors extend this processing route to the recognition of a 'type of person' rather than just a specific individual. These so-called 'personae' are speaker stereotypes; an example provided by the authors is that high-pitched Californian-accented speech with noticeable vocal creak is likely to evoke the "Valley Girl" persona (Lavan & McGettigan, 2023: 5). Interestingly, research suggests that these cognitive stereotypes are not necessarily based on objective phonetic evidence (cf. Dallaston & Docherty, 2019).

Unfamiliar voices are processed via a second route. This comprises the extraction of person-specific traits/features, which may be highly subjective. Bi-directional arrows in Figure 9 indicate likely interactions between individual perceived traits, such as speaker sex having implications for speaker height, as men are on average taller than women (Lavan & McGettigan, 2023: 5). In general, the model is less hierarchically organised within this recognition step, and considers bi-directional effects between the two general described routes of familiar and unfamiliar speaker processing. The authors explicitly state that, although "visually depicted as two routes, these processes are not mutually exclusive to one another (Lavan & McGettigan, 2023: 5)." Moreover, both routes feed into and from the wider cognitive system for the retrieval of connected names, faces and episodic information (Lavan & McGettigan, 2023: 6). The only aspect within the recognition step that may adhere to a hierarchy are possible ordering effects over the course of trait perception. For instance, physical traits like age and sex may be recognised earlier than social/psychological traits (cf. Section 4.1.2.3).

The model does not make a definite suggestion for how recognition is achieved mechanically during the recognition step. However, the authors provide two different theories for how the encoded voices are compared to stored representations of persons, personae, or characteristics (Lavan & McGettigan, 2023: 7f.). The first theory is in line with traditional mechanistic ideas of person perception and considers a prototype-based mechanism (e.g. Lavner et al., 2001; Maguinness et al., 2018; Rips, 1989). This mechanism has already been discussed in Section 2.3.1.2 as a plausible explanation for how listeners cognitively assess the distinctiveness of a voice by perceiving its deviation from a prototypical average voice. Such a deviant feature mechanism is not unlikely, given that many studies have identified strong correlations between voice distinctiveness and recognition accuracy (cf. the discussion in Section 4.1.2). A version of this mechanism is the existence of multiple voice exemplars, e.g. for specific accents, rather than one general prototype (cf. Section 2.3.1.2).

As a second option, the authors suggest that encoded acoustic features of voices may be compared directly to stored representations of persons, personae, or characteristics, i.e. without an intermediate assessment of prototype deviations. While the authors lean toward this mechanism, it must be stressed that there is – as of now – no behavioural evidence available.

As mentioned earlier, the PPV model has important implications for earwitness research. These include the following points:

1. The forensic literature often juxtaposes a forensic voice expert's method for speaker comparisons, which is necessarily a featural (bottom-up) approach, to an earwitness's approach, which is often assumed to be Gestalt-based (top-down) (cf. Section 3.2.3.3). However, the PPV model suggests that lay listeners may also analyse features. Before listeners form a stable representation of a voice that might later lead to more immediate recognition, the most likely route for unfamiliar voice recognition begins with a featural analysis.
2. Moreover, even witnesses familiar with the perpetrator may rely on features in certain contexts. The authors stress that the route to recognition taken by listeners is context-dependent. Brief utterances or utterances of low quality may only make certain features accessible (Lavan & McGettigan, 2023: 6). This is highly relevant for forensic contexts, in which utterances may be short or of low imprint fidelity (cf. Section 2.3.2). In a wider sense, this is where the perception model ties in with the framework of vocal identity introduced in Chapter 2, which demonstrated that individual utterances cannot reveal all aspects of a speaker's vocal identity.

3. The introduction of ‘personae’ to voice perception establishes a link to Social Identity Theory (cf. Section 2.2.2). The ability to recognise personae on a Gestalt or featural basis describes a scenario in which the source is of an utterance is not defined as an individual but as a group of speakers. This can be an advantage if information is to be elicited from a witness for the purpose of profiling. On the other hand, it also bears the danger that in an individualisation or recognition task, the witness wrongly recognises an individual based on a shared persona with perpetrator. This scenario overlaps with problems derived from disguises, i.e. situations in which a perpetrator conceals their persona (cf. Section 2.3.2.3).
4. Finally, the model suggests that discrimination and recognition are highly related tasks, as they are based on a shared mechanism. This allows for the prediction that witnesses who are good at a recognition task can equally be expected to be good at discrimination tasks, which will be considered in the empirical part of this thesis.

4.3 Conclusion: Cognitive and neurological aspects of earwitness testimony

The psychological literature review performed in this chapter has shown that cognitive models of voice perception have been heavily influenced by equivalent research on faces. However, from a neurological point of view, strong evidence can be found that both modalities are processed in largely separate pathways. This leads to the conclusion that earwitness testimony should be treated in its own right rather than rely on adaptations of findings on eyewitnesses (a similar point is made by McDougall, 2021: 38). The discussed PPV model (Lavan & McGettigan, 2023) does therefore provide promising new impulses for earwitness research as it is modality-specific and considers both familiar and unfamiliar voice recognition. The consideration of both ‘persona’- and person recognition, in combination with the possibility of featural processing, allows for predictions of witness behaviour in diverse contexts.

5. Empirical contribution

This chapter provides a background for the empirical part of this thesis. As outlined in the introduction, three perception tests were conducted to describe individual differences in the voice processing capabilities of the general public. Chapter 3 demonstrated that current methods for earwitness testing, such as voice parades (VPs), provide a trier of fact with an a priori likelihood for the correctness of a witness's speaker identity judgement, while there is no baseline information available on the individual witness's capacity to recognise voices. Following the traditional classification of variables in witness cases (Wells, 1978), it is understood that so-called "estimator variables" can be used to estimate a witness's reliability. However, it was argued in Chapter 3 that this approach may be too categorical an assessment of the witness, who is a crucial variable in the individualisation process. After all, it is the witness who compares the suspect's voice (KS) with a voice they heard the crime scene (QS), within the safeguards of the VP. The direct testing of witness-related variables by means of a screening test might therefore allow for a more nuanced assessment of an individual witness's capabilities (cf. Section 3.5).

5.1 Experimental baseline for witness variable testing

The conducted voice perception tests would fall short of the requirements for a fully-fledged screening test for earwitnesses, as they did not emulate the complex event variable combinations that can be found in most criminal cases. Moreover, the point was made in Section 3.5.2 that a screening test for earwitnesses would ideally be tailored to the event variables of the specific case for which it is composed, as by doing so, the test can create new system variables for investigators to control. The current tests can therefore not be regarded as a definitive template for earwitness screening. They rather provide an empirical baseline for future test development.

This baseline is critical as there are – to my knowledge – no existing voice processing tests that exhibit ecological validity for forensic applications, as will be reviewed in Section 5.3. Consequently, the current tests provide an impulse for further research into witness variable testing, by presenting listeners with tasks that are deemed "easy" compared to the potential complexity of a real-world earwitness scenario. An implicational hierarchy is assumed, in that finding significant between-listener differences under these simplified conditions is a strong indication for even greater variability under more complex real-world conditions. Further

research that systematically increases task complexity will be essential to fully mapping out this hierarchy. The current tests thus represent a starting point, while considerable work remains to develop maximally valid earwitness screening.

What sets the conditions in these baseline tests apart from some crime scenes is the high audio quality of the speech material and the matching conditions between exposure and test. That is, the technical quality of materials was identical in the first exposure to the voice and the second exposure, after which a judgement was elicited. In this connection, Smith et al. (2019) found that accuracy is highest when the stimulus quality is not mismatched. Moreover, no attempts were made in the present test to emulate the stress that the witness may experience during a crime.

Another potential difference from an actual earwitness scenario is that all stimuli contain the speakers' modal voice with no influence of factors often found in criminal events, such as marked emotion. In general, the impact of emotionality on voice recognition has not been well investigated. In the third of three experiments conducted by Read and Craik (1995), participants who were unaware of the experiment's purpose were exposed to either a highly emotional target stimulus (both in terms of 'tone' and content) or a non-emotional target stimulus during a study phase. In the test phase, participants were then presented with a line-up containing an utterance by the target speaker as well as five foil voices. The target utterance was either identical to the stimulus heard in the study condition, or a different utterance by the target speaker matching the 'tone' of the original stimulus. As a third option, an utterance by the target speaker was presented that differed in 'tone' from the original stimulus. Emotionality of the original stimulus was not found to have an impact on memorability. It is, however, not clear what the specific phonetic correlates of an emotional 'tone' were, or how closely the emotions conveyed in the stimuli matched those in real-life scenarios.

In line with the described implicational hierarchy, the complexity of the participants' task increased with each test. The first test conducted is therefore assumed to be the "easiest" one for the average participant, while the last test is assumed to be the most difficult. The three tests can be categorised as follows:

1. A voice discrimination test in which the role of memory was reduced to a minimum and participants were aware of the task (Chapter 6).
2. A voice recognition test in which participants were aware of the task (Chapter 7).
3. A voice recognition test in which participants were not aware of the task (Chapter 8).

Since the PPV model (Lavan & McGettigan, 2023) suggests that voice discrimination and voice recognition are based on a shared cognitive mechanism, both tasks are relevant for the assessment of earwitnesses. If this is the case, performances in these tasks are likely to be positively correlated. To account for this hypothesis, a subgroup of the listeners who participated in the first test (discrimination test) were reinvited second test (recognition test). Despite the assumed relatedness of both tasks, discrimination is assumed to be the easier task, as it is a more immediate judgement, whereas recognition necessarily involves memorisation (cf. Section 3.1). This hierarchy is also plausible from a logical point of view, as an earwitness cannot necessarily be expected to remember a voice from a crime scene, if they have failed to tell that voice apart from other voices heard during or after the criminal event. A further complication was added in the third test, which differs from the second test in that participants were not aware that a recognition test would be conducted after first exposure to voices; instead, their attention was directed away from the speaker's identity. In all other aspects, tests two and three were identical.

Based on the increasing complexity between tests, it was assumed that with each test the average listener performance would decrease, while the variability of performances across listeners would increase.

To facilitate a comparison of the individual tests' results, only the type of voice processing required from the participants differed between tests, while other independent variables were kept stable. For instance, the stimuli for all tests were sourced from the same corpus (DyViS, Nolan et al., 2009) and stimulus durations were not altered between tests. Moreover, the same online platforms and tools were used for creating all three tests and the participant samples had a similar demographic composition. These measures counteract the interpretability issues of McGehee's (1937) seminal study, in which several independent variables were manipulated in different versions of the experiment, making it difficult to attribute performance differences between tests to individual manipulable variables (cf. Section 3.4.1).

Note that none of the three tests addressed speaker individualisation as only unfamiliar voice recognition was tested (cf. Section 3.1). Chapter 4 showed that unfamiliar and familiar voice recognition are likely to be neurologically distinct, but functionally similar. Therefore, familiar voice recognition tests would have been a relevant addition to this study. However, the Covid-19 pandemic impeded the work with pre-existing social networks and familiar-famous voice recognition tests were avoided due to the controllability issues described in Section 4.1.2.1.

The current baseline tests also explore the possibility that direct testing may not be a necessary means for assessing witness variables, and that future research may thus take more categorical approaches. To this end, the interpretation of each test's results considers the possibility that there might be observable predictors for participant accuracy, such as participant age, sex, or confidence. If robust predictors are identified, a screening test would be superfluous. Interactions between confidence and accuracy are of particular interest, as an earwitness's confidence is currently one of the few parameters that may help a trier of fact determine the weight of VP evidence (cf. Section 3.5). In this connection, Deffenbacher (1983; 1980) proposed the so-called "optimality hypothesis", according to which the likelihood of finding statistically reliable positive correlations between witness confidence and accuracy depends on how optimal the information-processing conditions are for the witness. The more ideal these conditions are, the better witnesses should be at assessing the accuracy of their memory through expressed confidence ratings. While the overall body of evidence provides fairly strong support for the basic tenets of this hypothesis, some studies did not find a decrease in the correlation between confidence and accuracy under suboptimal conditions (e.g. Semmler et al., 2018; Wixted & Wells, 2017). Nonetheless, no study has – to my knowledge – indicated that optimal conditions may result in lower correlations between these two metrics. The simplified event variable combinations created for these baseline tests, such as high audio quality and non-mismatched recording conditions between study and test, are therefore ideal for investigating the relationship of participant confidence and accuracy, as real-world crime scenes are likely to exhibit less optimal conditions for a witness.

5.2 "Super recognisers"

The term "super-recognizer" (henceforth "super recogniser" / SR) was coined by Russell et al. (2009) to refer to people who excel at face recognition tasks. The authors were contacted by four people who claimed to have better than average face recognition ability. To test these claims, the authors designed two face recognition tests. This included the "Before They Were Famous Test", which tested the participants' abilities to individualise celebrities from photographs taken before they were famous, as well as an extended version of the pre-existing "Cambridge Face Memory Test" (CFMT, Duchaine & Nakayama, 2006). The CFMT had originally been developed to identify individuals with prosopagnosia, i.e. face processing deficits (cf. Section 4.1.1). It exposes participants to face photographs of six unfamiliar people in a study phase and a total of 30 photographs in a first test phase, i.e. photographs of the six

faces introduced in study as well as distractor faces. For each photograph presented, participants judge whether it is an "old" face (introduced in study) or not. In a more difficult second test phase, which follows the same principle as the first one, digital noise is added to the stimuli as a confounding variable.

Russell et al. (2009) subjected the four self-diagnosed super recognisers and 25 control participants to both tests and observed that the SRs performed significantly above the controls, suggesting their self-reports of extraordinary abilities were valid. In a second study within the same publication, the four SRs were subjected to the Cambridge Face Perception Test (CFPT, Duchaine et al., 2007) along with 26 prosopagnosic participants and 26 controls. The CFPT is an unfamiliar face discrimination/sorting test and had also been originally developed detect prosopagnosia. In each trial, participants were presented with a target face and six test faces that had been morphed to contain 88, 76, 64, 52, 40, and 28% of the target face, respectively. Participants were then given one minute to arrange the faces in descending order of their resemblance to the target face. Each trial was evaluated by calculating the sum of deviations from the correct position for each test face. The test includes eight upright and eight inverted stimuli trials in randomised order. Russell et al. (2009) found that SRs outperformed controls and prosopagnosic subjects in this test as well. The SRs displayed larger inversion effects (poorer performance with inverted vs upright faces) than controls, who in turn showed larger inversion effects than developmental prosopagnosics; low inversion effects for prosopagnosic participants were in line with existing findings (Duchaine et al., 2006). Moreover, the differences between SRs and controls were comparable in magnitude to the differences between controls and developmental prosopagnosics. This indicates that face recognition ability represents a broad continuum, with prosopagnosia at the lower end and super recognition at the higher end. The discovery of super recognition indicated that the range of face recognition ability in the population was wider than previously thought.

Since Russell et al.'s (2009) original findings, numerous studies have investigated super recognisers across fields like cognitive neuropsychology and forensic psychology (e.g. Bate et al., 2018; Bobak et al., 2016; Dunn et al., 2020; Mayer & Ramon, 2023; Nador et al., 2021; Ramon, 2019; Ramon et al., 2016, 2019; Tardif et al., 2019). Depending on the exact type of face processing investigated, more specific terms have been coined for superior performances, such as "super matchers" (Bate et al., 2018) for individuals excelling at face discrimination, or "super memorisers" (Ramon et al., 2016) for individuals excelling at face recognition after especially long delays. For reasons of simplification, these phenomena will be referred to as

SRs for the purpose of this study, given that discrimination and recognition are co-dependent tasks for a witness and that all recognitions involve some sort of memorisation.

The criteria for defining super recognition have varied between different studies, as discussed in a synopsis by Bate et al. (2021). The authors suggest that from a behavioural point of view, SRs be defined as people who find it extraordinarily easy to recognise unfamiliar faces briefly seen before (Bate et al., 2021: 2161). The vast majority of studies defines SRs on an empirical basis, as individuals whose performance in a given face processing test surpasses the mean performance by at least two standard deviations (SDs) (Bate et al., 2021: 2156). This 2-SD-cutoff point has been used since Russell et al.'s (2009) seminal paper and reflects consistency with criteria used in the prosopagnosia literature to identify the opposite end of the face recognition spectrum (Bate et al., 2021: 2155).

Like many other cases in which voice processing research has been inspired by research on faces, the voice processing literature has suggested a parallel performance spectrum, ranging from phonagnosia at the lower end to super voice recognition at the high end. In this connection, the same threshold of two SDs is applied to identify the respective extreme performers (e.g. Aglieri et al., 2017; Humble et al., 2022; Mühl et al., 2018). While it is not the primary purpose of the three voice processing tests conducted as a part of this study to identify SRs, the described performance spectrum provides important reference points for the description of an individual earwitness's voice processing capacity. The results of the conducted tests will therefore be discussed before the background of this continuum.

5.3 Insight from psychometric tests

The present tests take inspiration from existing psychometric voice processing tests that were developed for the purpose of placing individuals on the performance spectrum outlined above. In turn, these tests had taken inspiration from pertinent face processing tests, such as the CFMT (Duchaine & Nakayama, 2006), CFPT (Duchaine et al., 2007), or the "Before They Were Famous Test" (Russell et al., 2009). Section 5.3.2 will discuss the most widely used standardised voice processing tests, i.e. the "Glasgow Voice Memory Test" (GVMT; Aglieri et al., 2017), the "Bangor Voice Matching Test" (BVMT; Mühl et al., 2018), and the "Jena Voice Learning and Memory Test" (JVLMT; Humble et al., 2022). In addition, some noteworthy other approaches for assessing voice processing skills will be discussed in Section 5.3.3. Both sections will explain why these existing tests do not provide strong implications for earwitness testimony in their current form.

5.3.1 Signal Detection Theory

Before psychometric tests can be discussed, the bases of "Signal Detection Theory" (SDT, Macmillan, 2002), which is commonly used to interpret the results of such tests, must be introduced. Signal Detection Theory considers all four possible outcomes of a voice identity judgement, i.e. hits, misses, correct rejections and false alarms (cf. Section 3.5.1). These measurements can in turn be transformed into ratios by dividing the raw count of the respective measurement by the greatest possible number of their occurrence, resulting in a hit rate (HR), false alarm rate (FAR), miss rate (MR) and correct rejection rate (CRR), respectively. Based on these rates, several indices can be calculated. The simplest index is the percent correct (PC), which reflects participant accuracy, i.e. the proportion of correct judgements made by a listener (PC = ((HR + CRR)/2)*100). However, the PC does not distinguish between the two possible types of mistakes, i.e. misses and false alarms. It is therefore advisable to complement this information with the index D prime (d′), which is defined as the difference between the participant's z-transformed hit rate and false alarm rate (d′ = z(HR) - z(FAR)). Since hits and false alarms cover all cases in which the participant gave a 'same' response, the index reflects the participant's sensitivity to a correct match. A d′ score of 0 is equivalent to chance performance, i.e. a 1:1 ratio of hits and false alarms, while a score of 3 indicates near-perfect discriminability.

5.3.2 Most notable tests

5.3.2.1 Glasgow Voice Memory Test

The GVMT was the first publicly available psychometric test intended to function as a "valid screening tool … for a preliminary detection of potential cases of phonagnosia and of "super recognizers" for voices" (Aglieri et al., 2017: 97). The test structure loosely follows that of the CFMT (Duchaine & Nakayama, 2006), in that study phases, in which participants are introduced to a set of stimuli, are followed by test phases, in which participants must decide whether a stimulus was part of the study phase or not. In total, the test comprises four such phases, i.e. a study and a test phase for voices as well as a study and a test phase for bell sounds, in this order. Bell sounds were included to allow for the detection of dissociations between vocal and non-vocal auditory processing. In each study phase, participants hear eight stimuli (voices or bells) in a fixed order (Aglieri et al., 2017: supplemental material), each presented three times in a row. In the recognition phases, they hear the same eight stimuli as well as eight foil stimuli in randomised order and make an old/new judgement for each stimulus.

This particular setup limits the GVMT's informative value for earwitness research, as the "old" stimuli presented in the voice testing phase are the exact same stimuli used in the voice study phase, rather than different stimuli produced by the same speakers. The participants' task is thus one of stimulus recognition rather than voice recognition. It is particularly difficult to draw conclusions about the reliability of earwitnesses from such a task, as any form of (forensic) voice comparison is based on the premise that even the same speaker cannot repeat an utterance in exactly the same way (cf. Section 2.1.1). An earwitness will therefore inevitably be confronted with differences between a memorised QS and a KS presented in a VP.

The stimulus design of the voice recognition task further reduces the test's potential for forensic applications. Sixteen native speakers of Canadian French (8 male, age unknown) were recorded producing the French vowel phoneme /a/ in isolation (Aglieri et al., 2017: 99). The resulting stimuli are very different from naturalistic utterances, having an average duration of only 487 ms (Aglieri et al., 2017: 99). The dispersion of individual durations around that mean was not characterised by the authors, meaning that individual stimuli may have been markedly shorter. As demonstrated with help of the summative framework of vocal identity presented in Section 2.1.2, an individual utterance can only convey a proportion of the speaker's vocal identity. In the case of isolated vowel tokens, this proportion is alarmingly small. It is therefore common in forensic settings to work with duration thresholds for analysed material to ensure that a sufficient proportion of the speaker's vocal identity is conveyed (cf. Section 2.1.4). In terms of vocal identifiers (cf. Section 2.2), these stimuli almost exclusively provide organic cues to speaker identity, as cultural and habitual identifiers depend on the presence of speech.

Some points can be made in favour of this stimulus design, especially as the authors did not have a forensic application of the test in mind. Presenting vowels rather than more complex utterances thus mitigates the risk of speech content processing being a confounding variable in what was conceived to be a voice rather than a speech recognition test. Nevertheless, it is important to acknowledge that in real-world situations, speech content and voice are intrinsically linked.

The authors report on a validation trial, for which 1120 adults (337 male, aged 18 – 86, mean = 26.7) completed an online version of the GVMT; the exact online platform used is not known. Additionally, one phonagnosic subject (female, aged 62) took an offline version of the test under laboratory conditions, alongside 63 age-matched control subjects (29 male, aged 18 – 74, mean = 26.7), 38 of which took the test in Glasgow and 25 in Montreal. The online group had a mean PC score of 78.15% (SD = 10.95; range = 37.5 – 100) and a mean d′ of 1.66

(SD = 0.69; range = -0.67 – 3.07). No significant differences were found between the online group and the offline control group (Aglieri et al., 2017: 104), so that a discussion can focus on the online groups results alone.

The results indicate calibration issues, as a ceiling effect was observed for participant accuracy (PC), suggesting the test may be unable to differentiate between high performing individuals. This is evidenced by the fact that no SRs could be identified, i.e. participants with a PC of at least 2 SDs above the mean, as this would be equivalent to an impossible PC threshold of 100.05% for the online group. However, 22 individuals could be identified at the opposite end of the spectrum. Compared to the online group, the phonagnosic subject's PC for voice recognition was 2.57 SDs below the mean, indicating the test can detect phonagnosia despite the reduced discrimination among top performers. Bell sounds were recognised with significantly higher accuracy by the online group but exhibited the same ceiling effect (mean = 83.39%, SD = 9.97, range = 43.75 – 100). The calibration issues likely stem from the lack of a measure for stimulus difficulty. Consequently, no attempt was made to predict the likelihood of correct responses for individual test items. Information of this kind could have helped compose the test in a way that ensures adequate coverage of the latent ability range.

5.3.2.2 Bangor Voice Matching Test

The BVMT (Mühl et al., 2018) presents participants with a simpler task than the GVMT. Whereas the GVMT is a recognition task that requires participants to memorise stimuli between a study and a test phase, the BVMT is an AX discrimination task. This means that participants hear different voice pairs, with each pair containing a study stimulus "A" and a test stimulus "X." After listening to each pair, participants make a direct judgement about whether stimulus X was produced by the same speaker as stimulus A. Memory demands are thus minimised.

The BVMT stimuli consist of sustained vowels, consonant-vowel-consonant syllables, and vowel-consonant-vowel syllables, recorded from several hundred native speakers of British English (aged 18 – 28). Although the authors mention that an additional two short paragraphs of text were recorded per speaker (Mühl et al., 2018: 2186), only the isolated syllables and vowels were used for test construction according to the reported methods. Some of the speakers were excluded due to either a "pronounced regional accent or vocal health issues" (Mühl et al., 2018: 2186), resulting in a final speaker pool of 182 female and 149 male speakers. The impact of this measure is hard to assess, as the simple structure and short duration of the stimuli (mean = 510 ms, SD = 110 ms) naturally limits the potential for regional allophonic variation.

Consequently, the stimuli exhibit similar problems as the GVMT's stimuli in terms of ecological validity, despite some added complexity from syllables.

In contrast to the GVMT, measures were taken to determine item difficulty. The authors first measured f0 and F1 "in the stable portion of the sustained vowel /e/" for each speaker (Mühl et al. 2018: 2186). Subsequently, two levels of abstraction were added to these raw measurements. First, the speakers were plotted on a 2-dimensional space according to the root-mean-square-normalised f0 and F1 measurements. An initial pool of 288 voice pairs was then created based on distances between speakers in this abstract space; the exact procedure of this step was not explained by the authors. Second, 457 listeners (135 male) completed a "long version" of the test, which contained all of the initial pairs. The authors then used *item response theory* (IRT) models to determine the difficulty of each item based on the obtained participant ratings. Finally, a shorter "validation" version of the test was created which only featured the 80 most efficient items (40 male pairs and 40 female pairs). While these steps effectively address latent ability coverage, they obscure links between participant accuracy and the original phonetic selection criteria of the items, especially for the validation version of the test.

The validation version was administered to 149 new participants (36 male, age range unknown, mean age = 22.47) in an offline laboratory setup, along with several other tests, including the GVMT, the "Glasgow Face Matching Test" (GFMT; Burton et al., 2010), a music perception test (Law & Zentner, 2012), and a digit span memory test. These tests were included to assess whether voice discrimination ability correlated with voice memory ability, face discrimination ability, general auditory processing, and general memory capability, respectively. However, the results demonstrated only weak to moderate correlations between the BVMT and the other tests. The weakest positive correlation was with the GVMT's voice recognition task ($r = .23$), indicating that the tests measure related but dissociable latent abilities. Note, however, that the p-values of these correlations were not reported (Mühl et al., 2018: 2189).

Performance on the test (mean PC = 84.57%) was normally distributed and demonstrated substantial individual differences (PC range = 61.25 – 97.5, SD = 7.2) in voice discrimination ability. Note that 50% is equivalent to chance level in this AX-discrimination design. This indicates that the test was better calibrated than the GVMT, although it must be considered that the GVMT was validated on a significantly larger participant sample. While the authors define 'phonagnosia' and 'super recognition' as the extreme poles of the voice processing spectrum (Mühl et al., 2018: 2190), they do not discuss the results of their validation test in relation to

these concepts. It can, however, be deduced from the reported PC average and standard deviation that no SRs were found, while some participants fell into the phonagnosic range. This highlights limitations in detecting top performance. A detailed analysis of the results on the basis of SDT was not provided, which limits inferences about participant behaviour.

5.3.2.3 Jena Voice Learning and Memory Test

Out of the discussed tests, the JVLMT (Humble et al., 2022) is the most cognitively demanding task. The test involves three main phases – a study phase, a repetition phase, and a test phase. In the study phase, participants are familiarised with eight target voices by listening to different pseudo-sentences spoken by each voice (mean duration = 3727 ms, range = 2487–5603 ms, SD = 505 ms) (Humble et al. 2022: 1361). Immediately after familiarisation with one voice, they complete a three-alternative forced choice (3AFC) trial, testing their immediate ability to discriminate that voice from two foils. The foils used for this purpose are not re-used at later stages of the test.

The subsequent repetition phase passively exposes participants to the voices again; no participant judgement is required. The previously unfamiliar voices can be characterised as trained-to-familiar voices after this step (cf. Section 4.2). During the final test phase participants performed a series of 3AFC trials and decided in each trial, which one of the presented three voices was featured in the study phase. A new set of non-repeating foils was used for this purpose. Each target voice was tested in multiple trials. In sum, this test setup introduces several novelties that had not been present in the GVMT and BVMT, such as an active training phase and the use of pseudo-language. Moreover, the 3AFC format used in the test phase might mitigate response bias, in cases where participants have a general bias towards responding "old" or "new" when presented with a binary choice.

The stimuli of the JVLMT were a huge step forward in terms of complexity. Nontheless, they differ from naturalistic speech in two important aspects. First, the JVLMT stimuli were created from read speech, which is less likely to elicit a similar degree of phonetic variability as spontaneous speech. Secondly, they exclusively contain sentences made up of semantically void pseudo-words, e.g. “ble sulpty debepts thek henbly stopapt” (Humble et al. 2022: 1355). The use of speech-invariant stimuli is in line with the JVLMT’s intended applicability to native speakers of different languages, while avoiding a language-familiarity effect (LFE). However, research with infant listeners (Johnson et al., 2011) and time-reversed speech stimuli (Fleming et al., 2014) suggests that LFEs are not necessarily dependent on language ability but rather on

sound structure. This might affect the JVLMT stimuli as they exclusively contain English phonemes and adhere to the phonotactic constraints of English. The authors of the test did not, however, find indications of LFE in their assessment of the data (Humble et al., 2022: 1366). A disadvantage of pseudo-speech stimuli is that a situation in which speaker and listener are known to share the same language cannot be replicated. This is especially problematic where the ability to recreate a specific composition of variables is desired, like in an earwitness cases.

Similar to the BVMT, the JVLMT's initial item pool was defined on the basis of phonetic criteria. The authors plotted the recorded stimuli on a 3-dimensional space, the dimensions being the z-transformed f0, formant dispersion, and harmonics-to-noise ratio of the items. Subsequently, an initial pool of 72 test triplets was created by defining equilateral triangles in this space, with small, medium, and large triangles representing high, medium, and low acoustic similarity, respectively. A long version of the test was created featuring all initial test items was administered to 232 online participants (101 female, 130 male, 1 other; aged 18 – 72, mean = 37.4, SD = 11.5) from different countries. Based on these participants' responses, a final validation version of the test was created that contained the 26 most selective items.

The final validation test was completed by 454 international participants (156 male, 296 female, 2 other), aged 18 to 74 (mean = 37.4, SD = 12.21), who took the test online, recruited through the online platform Mechanical Turk. The test battery further comprised the BVMT (validation version), GVMT and a digit span test. The validation test demonstrated good convergent validity through strong correlations with the BVMT and moderate correlations with the GVMT. It also showed good discriminant validity through weak correlation with a digit span test. Test scores were approximately normally distributed and ranged widely from 5 to 95% accuracy (mean PC = 51, SD = 18). This enabled the identification of individuals at the extreme ends of the ability spectrum, including four SRs and seven participants at the lower end of the spectrum, using the common 2 SD-cutoff criterion.

In sum, the test is exceptionally well calibrated and accounts for a wide range of latent ability. Nonetheless, the stimuli were constructed from pseudo-speech and are thus devoid of a propositional content. This eliminates an important source of complexity that characterises real-world speech processing. Both these aspects limit the JVLMT's informative value for earwitness research.

5.3.2.4 Comparison of existing voice processing tests

The designs and main findings of the discussed tests are summarised in the table below. Note that the tests are arranged in order of increasing task complexity (cf. discussions above) rather than by publication date.

Table 5: Comparison of the most prominent voice processing tests

Test	**BVMT (validation version)**	**GVMT (voices version)**	**JVLMT (validation version)**
Stimulus type	Sustained vowels; CVC-syllables, VCV-syllables (M = 510 ms)	Canadian French vowel /a/ (M = 487ms)	Pseudo-sentences, created with online tool "Wuggy", (M = 3727ms)
Stimulus presentation	80 AX discrimination trials (40 same pairs, 40 different pairs)	2 phases: - Study: 8 voices - Test: 16 voices (8 old, 8 new)	3 phases: - Study (8 voices) - Repetition - Test: 26 3AFC items
Tested processing task	Voice discrimination	Voice recognition, tapping into short-term memory	Voice recognition, tapping into long-term memory
Participants (in the original publication)	N = 457 Mean = 22.47 years SD = 7.27 Range not specified	N = 1120 Mean = 26.7 years SD = 11.1 Range 18 – 86	N = 454 Mean = 37.4 years SD = 12.21 Range = 18 - 74
Data collection	Offline Psychtoolbox-3	Online (platform not specified) + Offline controls	Online Psytoolkit
Collected participant responses	Same/Different judgements	Old/New judgements	3-alternative forced choices
PC score summary	Mean PC = 84.57 Range = 61.25 – 97.5 SD = 7.2 Span = 36.25	Mean PC = Range = 37.5 – 100 SD = 10.95 Span = 62.5	Mean PC = 51 Range = 5 – 95 SD = 18 Span = 90
Extreme performances	0 SRs Some participants in phonagnosic range (number not specified)	0 SRs 22 in phonagnosic range	4 SRs 7 in phonagnosic range
Duration	Approx. 10 min	Approx. 20-25 min	Approx. 22 min

5.3.3 Noteworthy other approaches

A study by Jenkins et al. (2021) examined whether people with exceptional super face recognition abilities are also more likely to possess superior voice recognition abilities, i.e. "whether super-recognition status generalises across face and voice modalities" (Jenkins et al., 2021: 592). Note that at the time, the JVLMT (Humble et al., 2022), which found voice SRs, had not yet been published. On the other hand, earlier voice processing tests, such as the GMVT and BVMT, had been unable to identify voice SRs in the validation versions that were discussed in the respective original publications (Aglieri et al., 2017; Mühl et al., 2018). Consequently, no SRs had been identified by means of testing when Jenkins et al. (2021) conducted their study.

Based on the assumption that face and voice processing are "underpinned by a common cross-modality mechanism", the authors hypothesised that diagnosed super face recognisers would outperform controls in voice processing tests (Jenkins et al., 2021: 592). However, this hypothesis is not necessarily compatible with the full range of neurological and cognitive evidence available. While some crossmodal integration likely occurs, especially at higher cognitive levels, there is strong evidence for the existence of distinct neural pathways for face and voice perception (Young et al., 2020: 406, cf. also Section 4.2). Furthermore, the failure of existing voice tests to identify SRs probably stemmed from (near-)ceiling effects in those tests (BVMT, GVMT) and the underlying calibration issues, rather than an actual absence of top-performers in the population.

To test their hypothesis, the authors re-recruited participants from a prior study who had completed the Glasgow Face Matching Test (GFMT) as well as extended version of the Cambridge Face Memory Test (CFMT). Based on those previous face recognition assessments, participants were categorized as 'super face matchers' (GFMT experts), 'super face recognisers' (CFMT experts), 'super face identifiers' (expert at both tests), or controls (mid-range scorers at both tests) (Jenkins et al., 2021: 601). All four groups completed the BVMT, the GVMT, and a bespoke Famous Voice Recognition Test consisting of 38 trials in which famous (N = 28) and foil (N = 10) voices were presented. The order of tests was identical for all participants. Seventy-six of the original 605 participants were excluded from the final sample, including 61 participants who scored below chance level at the BVMT and GVMT. The resulting final sample consisted of 529 participants with a mean age of 36.9 years (SD = 11.8, Range = 18–76, 64% female). It included 165 super face identifiers, 89 super face recognizers, 41 super face matchers, and 234 controls. In line with the hypothesis, it was predicted that super face matchers and identifiers would outperform the other groups on the

BVMT (voice discrimination), while super face recognizers and identifiers would excel on the GVMT and Famous Voice Test (voice recognition) (Jenkins et al., 2021: 592).

Note that for several reasons this methodology might not be entirely suitable for testing the authors' hypothesis:

1. Super recognition is defined by exceptional performance relative to a broader population. The authors estimate that only 2% of the population are 'super face identifiers' (Jenkins et al., 2021: 593). Thus, the identification of SRs (2 standard deviations above the mean) depends on the performance distribution in a given test sample. If the hypothesis that equivalent face and voice processing skills are positively correlated holds true, the authors are administering the BVMT to 206 participants who are likely to be exceptional voice matchers (super face identifiers and super face matchers) and 323 other participants (controls and super face recognisers). Likewise, they are administering the GVMT to 254 participants who are likely to be exceptional voice super recognisers (super face identifiers and super face recognisers) and 275 other participants (controls and super face matchers). This means that people with skills assumed to reflect about 2% of the population make up about 39% and 48% of the respective samples. Both tests should in this case generate strongly left-skewed distributions, with elevated means and standard deviations, making it impossible to find superior performances based on a group norm of two standard deviations above the mean. This problem is exacerbated by the known ceiling effects in the original BVMT and GVMT validation studies, which failed to identify top performers in representative samples (Aglieri et al., 2017; Mühl et al., 2018).
2. If the authors have identified people with superior face processing skills in previous studies and their hypothesis holds true, reinviting a sample of participants that matches the distribution of the obtained face test scores should be a sufficient criterion for finding superior voice recognition skills.
3. The authors excluded 69 participants who scored below chance on the BVMT and GVMT without explaining this decision. However, removing low scorers on the dependent variables is questionable, since poor performance could simply reflect low latent ability rather than inattentiveness. This is especially concerning if the excluded participants were those categorised as super face matchers/recognisers/identifiers (the independent variable). Omitting their low voice scores could artificially inflate the apparent correlation between face and voice recognition skills. Unless the authors have

clear evidence that the participants were not properly engaged, excluding those with low scores risks further skewing the sample and results.

4. The fixed order of test administration (BVMT, GVMT, FVRT) for all participants may impact results. This is especially true for the GVMT, which is the only unfamiliar voice memory test in the battery. By the time the GVMT's test phase took place, the participants had heard the eight stimuli from the GVMT's study phase as well as all of the BVMT's 80 stimuli. If GVMT foils seem familiar due to resemblance to BVMT voices, performance could be confounded. Additionally, with a total battery duration of 40-50 minutes (Jenkins et al., 2021: 594), differing fatigue levels across tests are a concern. Ideal methodology would counterbalance test sequence across participants to allow clearer conclusions about each measure.

Similar to their validation versions, the GVMT and BVMT showed ceiling and near-ceiling effects in this study, respectively. Average participant sensitivity (d′) decreased between tests, with group averages ranging from 2.39 to 2.74 for the BVMT (no overall mean provided), 1.26 to 1.40 for the GVMT (no overall mean provided) and .58 to .80 for the Famous Voice Recognition Test (no overall mean provided). It is difficult to assess whether this decrease stems from an increase in task difficulty or a fatigue effect as described above (Pt. 4). D prime is known to decrease as participant fatigue increases (cf. Matthews & Desmond, 2002).

Comparing performance on face and voice recognition tests is not a wholly novel approach. For instance, Mühl et al. (2018: 2189) included the GFMT in the test battery for the BVMT's validation test, only finding a low-to-moderate positive correlation between the tests ($r = .24$, $p = .004$). Similarly, the highest correlation found by Jenkins et al. (2021: 599) was between GFMT and BVMT ($r = .30$, $p < .001$). Progressively weaker positive correlations emerged between BVMT and CFMT ($r = .25$, $p < .001$), Famous Voice Recognition Test and CFMT ($r = .15$, $p < .001$), GVMT and CFMT ($r = .12$, $p < .001$), GVMT and GFMT ($r = .11$, $p < .05$), and Famous Voice Recognition Test and GFMT ($r = .11$, $p < .05$). As expected, the two face tests (CFMT and GFMT) were more strongly intercorrelated ($r = .59$, $p < .001$).

The relatively robust correlations between the BVMT and both face processing tests could reflect shared mechanisms, especially as the BVMT accuracy scores showed the left-skewed distribution (cf. Jenkins et al., 2021: 595, Fig. 1) expected in a sample enriched with top-performers (see Pt. 1). In contrast, GVMT accuracy was more normally distributed and Famous Voice scores were right-skewed, implying greater divergence from face recognition abilities.

Unfortunately, the authors did not consider PC distributions in their discussion, nor are summary statistics for accuracy provided.

The authors identified several participants with exceptional voice processing capabilities. Note, however, that a criterion of 1.5 standard deviations above the mean was chosen instead of the commonly threshold of two standard deviations. This lenient approach may have been necessary in light of the enriched participant samples and the GVMT's and BVMT's problems at discriminating between top-performers. While this more lenient criterion allowed the authors to identify some individuals with superior voice recognition, it is unclear whether any participants would have qualified under the established 2 SD-criterion. This makes it difficult to compare the results to prior work on super recognition. Overall, the superior performers were found in all of the conducted voice processing tests. The experiment group with the highest proportion of super voice recognisers were the 'super face identifiers'. However, all four experiment groups produced 'super voice recognisers', including the controls.

In sum, the study does not provide strong evidence for a generalisation of super recognition status across modalities. This stresses the danger of using combined assessment methods for eye- and earwitnesses in forensic contexts.

5.4 Measures taken to ensure ecological validity

The discussion of existing psychometric voice processing tests has revealed that some design choices limit the informative value these tests offer for earwitness research. The following list therefore summarises key measures taken in the present study to ensure the ecological validity of the conducted voice processing tests for forensic applications:

1. The tests should use naturalistic stimuli, ideally sourced from spontaneous speech. This measure ensures that organic, cultural, and habitual identifiers are present in the samples.
2. The stimulus duration should be forensically realistic. Schweinberger and Zäske (2018: 542) suggest a minimal duration of 1.5 seconds for general voice processing tests. Sample durations in forensic tests should be markedly longer. For instance, to my knowledge, the shortest duration for VP samples discussed in the literature is 15 seconds (cf. Pautz et al., 2023). Very short samples may not be sufficiently representative of a speaker's overall vocal identity (cf. Chapter 2).

3. Stimulus difficulty should be controlled to allow for a prediction of item difficulty, and ultimately for calibrating the test. While IRT was shown to be an effective means of controlling difficulty, a test with forensic implications may benefit from a method that
    a. is less complex and therefore easier to reproduce by a wide range of professionals
    b. allows for drawing a direct connection between difficulty and raw measurements of phonetic criteria
4. Different stimuli should be used for study and test, even for same-speaker items. This mitigates the risk of participants recognising the sample (e.g. recording conditions) rather than the voice.
5. Similarly, recording conditions (imprint fidelity) should be kept constant across all test items to avoid test items standing out for reasons other than the featured voice.
6. Large test batteries should be avoided. While exploring the correlation between voice processing and other abilities may be insightful, long batteries may induce fatigue or memory effects. This is especially problematic for the intended baseline character of the present tests.

6. Test 1 – Individual differences in voice discrimination

6.1 Purpose

In line with the implicational hierarchy outlined in Section 5.1, this first voice processing test is assumed to be the least complex out of the three tests conducted. It is an unfamiliar speaker discrimination task. As demonstrated in Section 3.1, speaker discrimination is a more basic form of voice processing than speaker recognition or speaker individualisation as it describes a direct comparison of different samples/voices. Memorisation, which is a central characteristic of recognition, is thus minimised. The test follows an AX design, which is similar to the BVMT's setup, with only a short delay of 10 seconds between compared voice stimuli. While earwitnesses necessarily perform a kind of voice recognition (cf. Section 3.1), there are several reasons for beginning this series of baseline tests with a discrimination task.

Firstly, discrimination is a prerequisite for unfamiliar speaker recognition. For instance, in order to memorise the voice of an unfamiliar perpetrator, an earwitness must first be able to discriminate that voice from others. The likely functional integration of these two aspects is predicted by the PPV model (Lavan & McGettigan, 2023, cf. Section 4.2.3.2). Secondly, the success of a VP depends on an earwitness's voice discrimination skills, as the witness must be capable of discriminating between voices presented in a VP.

What sets the present test apart from most existing voice recognition and memory tests is the use of naturalistic stimuli, which is intended to make the results more applicable to earwitness investigations. Moreover, the test also makes use of reaction time (RT) measurements for all listener judgements. The role of RTs in voice recognition tests is not well-investigated and may prove to be a more reliable predictor for accuracy than participant confidence, since it potentially captures a more instinctive reaction to the stimuli.

6.2 Methodology

The experimental procedure and recruiting procedure were approved by the ethics committee of the University of York's Department of Language and Linguistic Science.

6.2.1 Stimulus design

Stimuli were extracted from task 1 of the DyViS corpus which consists of mock police interviews with 100 speakers of *Standard Southern British English* (SSBE), aged 18 to 25 years (Nolan et al., 2009). The age range is comparable to the age range of 18 to 28 years of the BVMT speaker pool (Mühl et al., 2018: 2186). The dataset was chosen because the foil recordings used for the construction of VPs in the UK are usually taken from police interviews of unconnected criminal cases (McDougall, 2021; Nolan, 2003). Moreover, SSBE speakers with no regional dialect features were chosen to counteract the accent-familiarity effect (cf. Njie et al., 2022). This decision is a compromise and based on UK listeners' general familiarity with SSBE, despite the accent's geographical ties to the southeast of England. It is therefore justifiable to recruit listeners from all over the UK, as the use of SSBE in the stimuli mitigates the risk of the speakers' accent being a confounding variable.

Twelve speakers were excluded because of different recording conditions (N = 3), an accent other than SSBE (N = 1), or noteworthy features which listeners might perceive as disordered speech (above all various types of functional speech disorders affecting the production of alveolar fricatives, such as an interdental lisp; N = 8). Different recording qualities bear the danger that participants recognise the recording conditions rather than the featured voice. Similarly, recordings featuring different accents or noteworthy features like speech disorders could stand out because they display a greater degree of linguistic or idiosyncratic variability than the other stimuli. Forty-eight of the 88 remaining speakers were used for the present test, while the remaining 40 speakers were used for the other two tests conducted in this study.

Two 10-second-long extracts were taken from each of the 48 speakers: one for the first exposure to the voice (i.e. the study condition), and one for subsequent recognition (i.e. the test condition). Note that 10 seconds is the overall duration of the sound files used as stimuli, including pauses. All stimuli contained at least seven seconds of net speech, which is considerably longer than the stimuli used in existing tests (detailed discussion in Section 5.4). Having stimulus files of equal length facilitated later use in the software *PsychoPy*, which was

used to create the experiment, and helped predict the likely duration of the experiment, irrespective of the exact files featured in the final trials.

Every stimulus consisted of one or more complete intonation phrases. Although the stimuli featured spontaneous speech, there was relatively little diversity in the propositional content. The speakers were asked similar questions by the interviewer and adopted the same fictional persona for the purpose of the task. Topics that came up in the interview include neutral descriptions of e.g. the interviewee's daily routine, workplace, colleagues, or friends. Two non-overlapping lists of possible conversational topics were created and it was ensured that the two stimuli extracted per interviewee were not taken from the same list. The stimuli created from the first list were only used for the exposure to the voice, while the second list was only used for the subsequent test phase. An identical propositional content in both the exposure and the test phase could be a clear indicator that the stimuli were produced by different speakers. At the same time, using the same stimuli in both exposures might lead to a recognition of the stimulus rather than the speaker (which might, for example, have impacted the results of the GVMT (Aglieri et al., 2017)).

All 96 sound files were converted from stereo to mono, the peak amplitude normalised (relative of a peak level of -1 dBFS), and any DC offset removed. These tasks were conducted via the programme *Audacity* (version 2.3.3). In a few cases, mild audible background noise was removed using Audacity's "noise removal" function. All these measures served the purpose of homogenising the sound quality of the files, so that differences in recording quality would not impact the listeners' judgements. All stimuli files were stored in .wav format with a sampling rate of 44.1 kHz at 16 bits (output 705 kBit/s).

6.2.2 Stimulus difficulty and hypotheses

In line with the design principles for baseline testing outlined in Chapter 5, it was predicted that the fundamental frequency (f0), which is the acoustic correlate of pitch, would account for a large proportion of perceived between-speaker difference. Pitch is known to play an important role in judgments of voice similarity (Baumann & Belin, 2010; Eriksson & Wretling, 1997; Nolan et al., 2011; Remez et al., 1997; Sørensen, 2012). Thus, a low f0-difference between speakers would make it harder to tell recordings of different speakers apart, while a high f0-difference would make it harder to tell recordings of the same speaker together. The homogeneity of the accents featured in the DyViS corpus ensures a relative stability of idiolectal

features of voice, ensuring that the participants' judgements of voice similarity will be predominantly informed by organic and habitual features.

A *Praat* (Boersma & Weenink, 2022) script was used to extract the f0-contour and to determine the average f0 of each sound file. The speaker pool had a mean f0 of 106.7 Hz (Min = 78.5 Hz, Max = 142 Hz, SD = 14.4 Hz). The average f0-difference between the two sound files obtained from a given speaker was 4.8 Hz (Min = 0 Hz, Max = 19 Hz, SD = 4.3 Hz). Subsequently, 96 voice pairs of various difficulty were created from the 96 DyViS recordings: 48 same-speaker pairs and 48 different-speaker pairs. Difficulty was determined by the average f0-difference between the extracts in a pair. Three stimulus lists were created that were deemed equally difficult, each list containing a different 32 pairs (16 same-speaker pairs and 16 different speaker pairs). Each of the 48 DyViS speakers was featured once per list, either with both recordings, in the case of a same-speaker pair, or one recording, when featured in a different-speaker pair. The reason for having three different stimulus lists with unique voice pairs was to test whether the chosen principle of determining item difficulty by means of f0 difference works irrespective of the specific stimuli used. Participants could thus be divided into three distinct groups for the purpose of the experiment, where each group is presented with a different stimulus list. Similar controls do not exist for the BVMT and GVMT, which expose all participants to the same test items.

It was hypothesised that participant performances would vary significantly within each participant group as well as across all participants. It was further hypothesised that the performance spectra would be comparable between the three participant groups.

6.2.3 Participants

One hundred British participants (50 male, 50 female) were recruited via *Prolific* with the aim of obtaining a representative sample of the population. The age range was 18 – 68 years, with a mean and median of 36, and SD of 13.8. The goal was to have an equal number of male and female participants and a median age close to the British population median of 40.7 years.[19] The experiment was only offered to *Prolific* users who stated that they were UK nationals, native speakers of English, and currently residing in the UK. To increase the likelihood of participants taking the task seriously, users with a "Prolific score" lower than 90/100 could not participate. The score is a ratio of the participant's successfully completed and rejected studies.

[19] https://www.statista.com/statistics/281288/median-age-of-the-population-of-the-uk/; last accessed 25/01/2023.

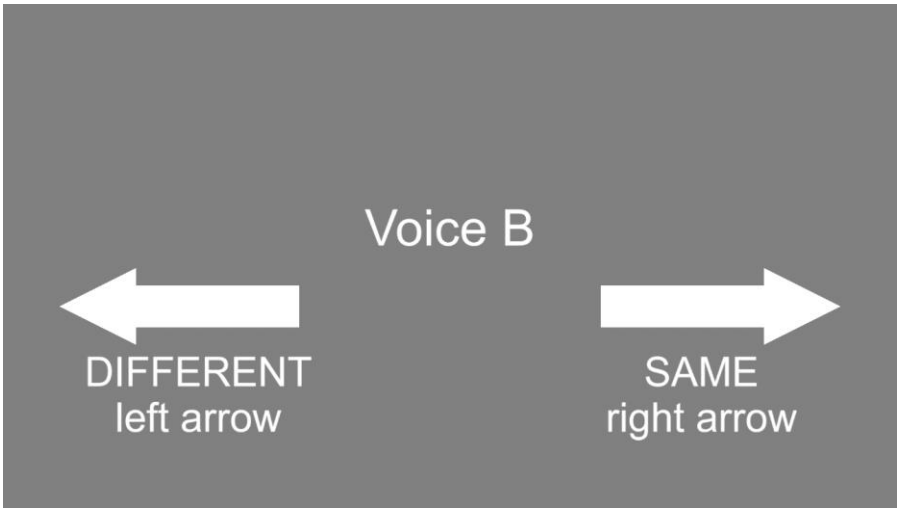
All users who signed up had a score greater than or equal to 95 (mean 99.3). To best control for participant age and sex, participants were recruited in four batches over the course of three days. After each recruitment call, the demographic composition (age and sex) of the existing sample was assessed to determine the group of *Prolific* users to be contacted in the next call. Otherwise, the recruitment calls were identical in content. All participants reported normal or corrected-to-normal hearing. Participants spent on average 20 minutes on the task and were paid £2.50 (equivalent to an hourly rate of £7.50).

6.2.4 Study procedure

To qualify for participation in the experiment, participants had to have access to a quiet room, be using a device with a physical keyboard (PC or laptop) and earphones/headphones, and be able to complete the task in one sitting. Participants knew that the test was unlikely to take more than 20 minutes due to the predetermined length of the stimuli, which had also been confirmed in a pilot study. First, a *Qualtrics* survey was used to elicit informed consent, as well as to collect meta-data, including age, sex, nationality, place of birth, mother tongue(s), and foreign language skills of the participants. Subsequently, the test takers were redirected to the *Pavlovia* server, where the actual test was hosted. The participants were given detailed instructions about the procedure before the test began.

Each participant was assigned one of the three stimulus lists. There were 32 trials based on the voice pairs defined in each list. The pairs were drawn from the list in random order for each listener. During an obligatory test trial participants were familiarised with the procedure and given an opportunity to adjust the volume of the input to their headphones. The test trial was identical for all participants and consisted of a different-speaker pair with a large f0 difference.

Figure 10: View of participants' screen during the test phase of each trial



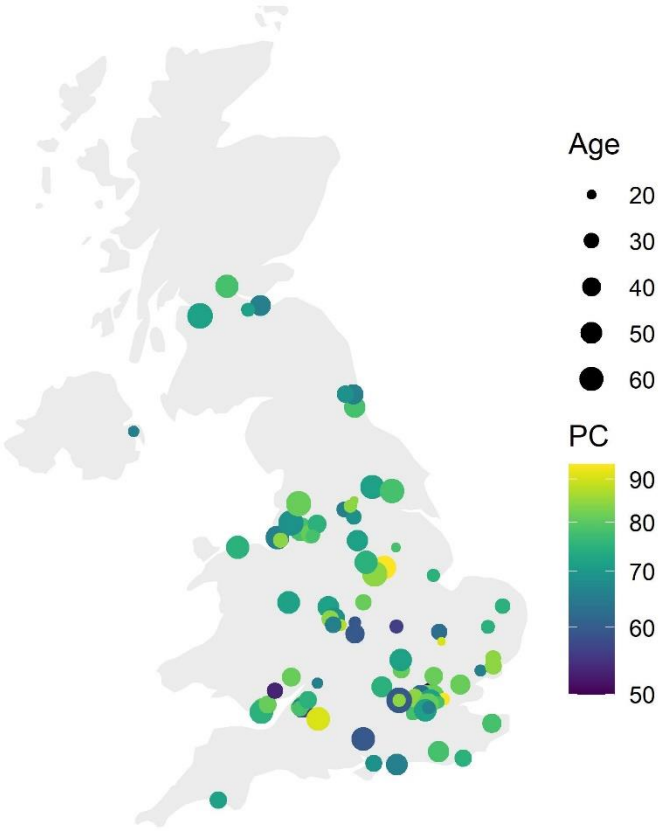
Each experimental trial of the AX discrimination task consisted of an exposure phase and a test phase. During the exposure phase, the participant saw the words "Voice A" on their screen, while the first 10-second-long stimulus of each voice pair was played. A 10-second-long countdown followed, during which the participants listened to a recording of unintelligible crosstalk.[20] Carterette & Barnebey (1975: 256) observed that voice recognition accuracy does not differ significantly for retention intervals of 0, 15, and 45 seconds. A relatively short interval was therefore chosen for the present test to minimise the overall duration. During the test phase, participants saw "Voice B" in the centre of their screen while the second stimulus played. A left-facing and a right-facing arrow icon were displayed at the respective sides of the screen. Additionally, the corresponding labels "DIFFERENT – left arrow" and "SAME – right arrow" were presented underneath the icons (Figure 10). Participants then pressed the left or right arrow button on their keyboard depending on whether they thought the Voice B stimulus was identical to the previously heard Voice A stimulus or not. The reaction time (RT) between the onset of the Voice B stimulus and the keyboard entry was also measured, unbeknownst to the participants. It was not possible to play either Voice A or Voice B more than once. The task was predominantly self-paced, in that participants had to press the space bar on their keyboard after each trial to start the next one, allowing them to take pauses between trials. At the end of the test, participants provided an overall rating of their confidence in making same-/different-speaker judgements using a 6-point Likert scale.

[20] https://www.youtube.com/watch?v=h2zkV-l_TbY&t=9282s; last accessed 25/01/2023.

6.3 Results

A map of the participants' locations within the UK is provided in Figure 11. Each dot represents one participant; size and colour of the dots reflect age of the participant and the PC score they achieved in the test, respectively.

Figure 11: Geographical map of the participants' locations within the UK



6.3.1 Participant-based analysis

Table 6 provides summary statistics for our test in comparison with the BVMT and GVMT. A comparison with the BVMT is particularly interesting as it was also an AX discrimination task. All calculations were conducted in *R* (version 2022.07.1).

Results support the first hypothesis, that performances would vary markedly across participants. In terms of PC, participant accuracy ranged from 50% (chance performance), to 93.75%, with a mean and median value of 75% (SD = 9.06), i.e. the exact midpoint between chance and perfect performance. The Jarque-Bera Test (Jarque & Bera, 1987), which tests the

null hypothesis that the data is normally distributed, confirmed a normal distribution of PC scores for the total sample of 100 participants ($p = .35$).[21]

The second hypothesis, that comparable participant performance spectra would be achieved across stimulus lists, was also supported. This outcome therefore validates the choice of f0-difference between stimuli as measure of item difficulty. Stimulus lists A, B, and C elicited average PC scores of 75.4% ($N = 31$, $SD = 10.4$), 75.5% ($N = 34$, $SD = 7.9$), and 74.1% ($N = 35$, $SD = 9.0$), respectively. A type 1 ANOVA revealed no significant differences between these means ($F(2, 97) = .26$, $p = .77$). The PC distributions of the individual stimulus lists were assessed with the help of the Shapiro-Wilk Test (Shapiro & Wilk, 1965), which is preferred for samples with fewer than 50 observations (Mishra et al., 2019: 70). The results suggest that the three lists have the skewness and kurtosis of a normal distribution (List A: $W(31) = .95$, $p = .12$; List B: $W(34) = .97$, $p = .48$; List C: $W(35) = .96$, $p = .32$).[22] Additionally, a Rasch model was fitted using the *ltm* package for R (Rizopoulos, 2006), to assess the relationship between PC scores and participant ability. It revealed that the lists cover the participants' latent ability to a similar extent (List A: 89.3%, List B: 88.5%, List C: 90.76%). The high level of similarity across the stimulus lists justifies that further analysis of the data be conducted on basis of the total sample of 100 participants, i.e. with data from all three lists combined.

As in the BVMT, participant accuracy did not correlate with overall test duration ($r = -.1$, $p = .31$). (Since reaction times were measured in the experiment, the cumulative RT across all of a participant's trials was used for this measurement, rather than the total time that the participant spent on the experiment. The advantage is that delays beyond the core task of the experiment are not reflected in the cumulative RT). Furthermore, no correlation could be observed between participant accuracy and confidence ($r = .09$, $p = .39$).[23] The accuracy span (Max_{PC} - Min_{PC}) of 43.75 is wider than the accuracy span of 36.25 elicited by the validation version of the BVMT, which suggests that the present test might be better at differentiating between individual listener capabilities. Given the smaller sample size of the test (100 compared to 149), this is especially surprising, as the validation version of the BVMT is based on a subset of the long version's items, which were chosen by means of *item response theory* (IRT) "to span a wide range of ability levels" (Mühl et al., 2018: 2187). Unfortunately, neither the

[21] Note that the test's true alpha level for a sample size of 100 is 0.1.
[22] Alpha level = 0.05. Null-hypothesis = normal distribution.
[23] NB 13 participants were excluded from the confidence analysis, as compatibility issues with a particular web browser prevented them from submitting a confidence rating.

accuracy span nor the standard deviation of the PC scores elicited by the long version of the BVMT are known, which makes it hard to assess the effectiveness of this measure.

Table 6: Comparison of the present test's results with results of the BVMT & GVMT

Present Voice Discrimination Test, N = 100 (50 male)					
	Min	*Max*	*Mean*	*Median*	*SD*
Age	18	68	36	38.85	13.8
PC	50	93.75	75.00	75.00	9.06
d´	0.00	2.94	1.38	1.35	0.57
Bangor Voice Matching Test, N = 149 (36 male) – validation version					
Age	*NA*	*NA*	20.49	*NA*	4.6
PC	61.25	97.5	84.57	*NA*	7.2
d′	*NA*	*NA*	*NA*	*NA*	*NA*
Bangor Voice Matching Test, N = 457 (135 male) – long version					
Age	*NA*	*NA*	22.47	*NA*	7.27
PC	*NA*	*NA*	75.99	*NA*	5.55
d′	*NA*	*NA*	*NA*	*NA*	*NA*
Glasgow Voice Memory Test (voices version), N = 1120 (337 male)					
Age	18	86	26.7	*NA*	11.1
PC	37.5	100	78.15	*NA*	10.95
d′	-0.67	3.07	1.66	*NA*	0.69

The d′ scores indicate that the present test is well-calibrated. The lowest measured score was 0, which is equivalent to chance performance. Negative d′ scores were never produced, meaning that none of the participants produced more false alarms than hits. At the same time, a ceiling effect was avoided, as the test would be capable of identifying performances between the current best score of 2.94 and near-perfect discriminability at a d′ of 3. The GVMT, for instance, reached a ceiling with a maximal d′ of 3.07.

With regard to extreme performances, two potential 'super recognisers' could be identified, which are commonly defined as participants with a PC of at least 2 SDs above the mean performance (Bate et al., 2021). In the case of the present test, this translates to a PC cut-off point of 93.12%. In comparison, the BVMT and GVMT (Aglieri et al., 2017: 103) did not find potential super recognisers.[24] On the other hand, four of the 454 participants who took part in the JVLMT (Humble et al., 2022: 1362) were identified as potential super recognisers.

Four participants in the present test could be identified at the opposite end of the spectrum, i.e. at least 2 SDs below the mean, which is equal to a PC below 56.88% in our test. The same

[24] NB that the authors of the BVMT did not explicitly state whether potential super recognisers were found. The authors did, however, report the mean PC score and the SD of PC scores (cf. Table 3), from which it can be reconstructed that no participant reached a PC score of at least 2 SDs above the mean.

threshold is usually applied when identifying cases of developmental phonagnosia (Roswandowitz et al., 2014). The GVMT (Aglieri et al. 2017: 103), which had 1120 participants identified 22 individuals with potential phonagnosia, while the JVLMT which had 454 participants found seven listeners in this range (Humble et al., 2022: 1362). The authors of the BVMT (Mühl et al. 2018) did not report the number of participants with a PC score of at least 2 SDs below average, but the summary statistics (cf. Table 6) indicate that some participants fell into this performance range.

6.3.2 Trial-based analysis

A trial-based analysis was conducted in order to determine the effects that had a significant impact on the outcome of an individual trial, i.e. each voice pair. To this end, a generalised mixed-effects model (GLMM) was fitted, using the *glmer* function of *R*'s *lme4* package (Bates et al., 2015). A generalised model was chosen because the dependent variable was a binary "correct response" score, which took the value of 1 in the case of a hit or correct rejection made by the participant, and the value 0 in case of a false alarm or miss. The assessed independent variables comprised five fixed effects: the participant's age (z-transformed across participants), the participant's sex, the participant's RT for the judgement (z-transformed within individual participants' RTs), the trial number (i.e. a number in between 1 and 32 depending on the position of the pair in the test), and the f0-difference between the stimuli in the pair. Participant identity (in the form of the participant number) was chosen as a clustering variable. A highly significant effect was found for RT ($z = -11.24$, $p < .001$), trial number ($z = -3.73$, $p < .001$), and f0-difference ($z = -3.4$, $p < .001$). No significant effect was found for participant age ($z = -.29$, $p = .77$) or sex ($z = 1.17$, $p = .24$).[25]

[25] In this study, the measure of predicted item difficulty was defined as the f0 difference (measured in Hz) between both stimuli in a voice pair. This value was also used as a fixed variable in the GLMM. An anonymous reviewer who assessed this experiment for publication in the *International Journal of Speech, Language and the Law*, pointed out that the logarithmic semitone scale might be more salient to a listener. The reason for measuring pitch in Hz rather than semitones was to establish consistency with the literature. To address the reviewer's valid concerns, the GLMM was run again, exchanging the fixed effect "f0 difference between stimuli in the pair" with the semitone difference between stimuli in a pair. The rounded p-values remained unchanged. The z values (test statistic) only differed in decimal positions. The specific values for the effects are: RT ($z = -11.24$, $p < .001$), trial number ($z = -3.76$, $p < .001$), semitone difference ($z = -3.72$, $p < .001$), age ($z = -.29$, $p = .77$), sex ($z = 1.17$, $p = .24$).

Figure 12: Predicted probability of a correct response in relation to z-transformed RT

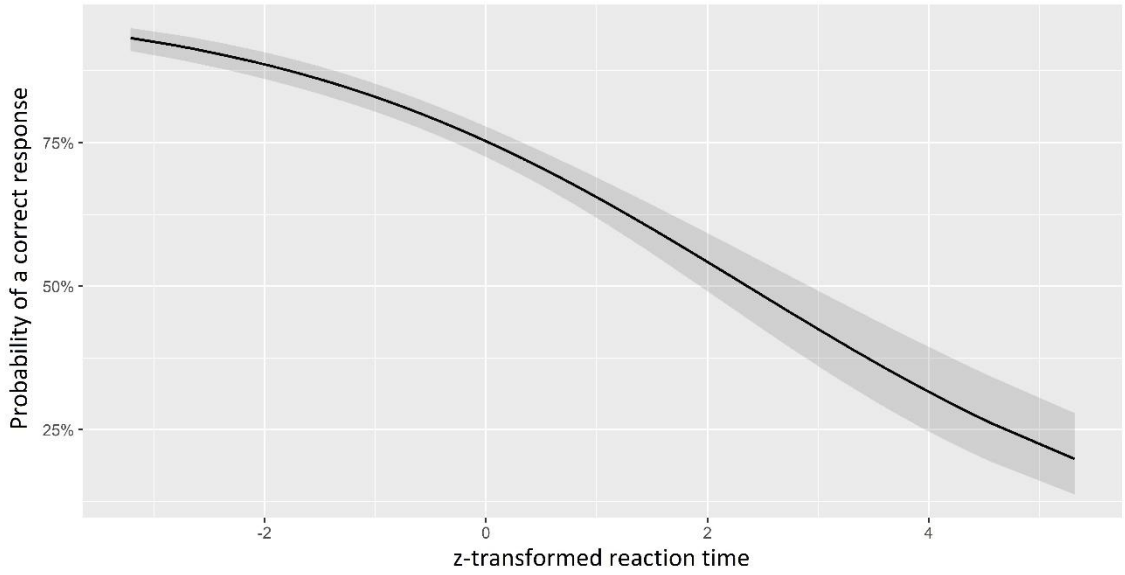
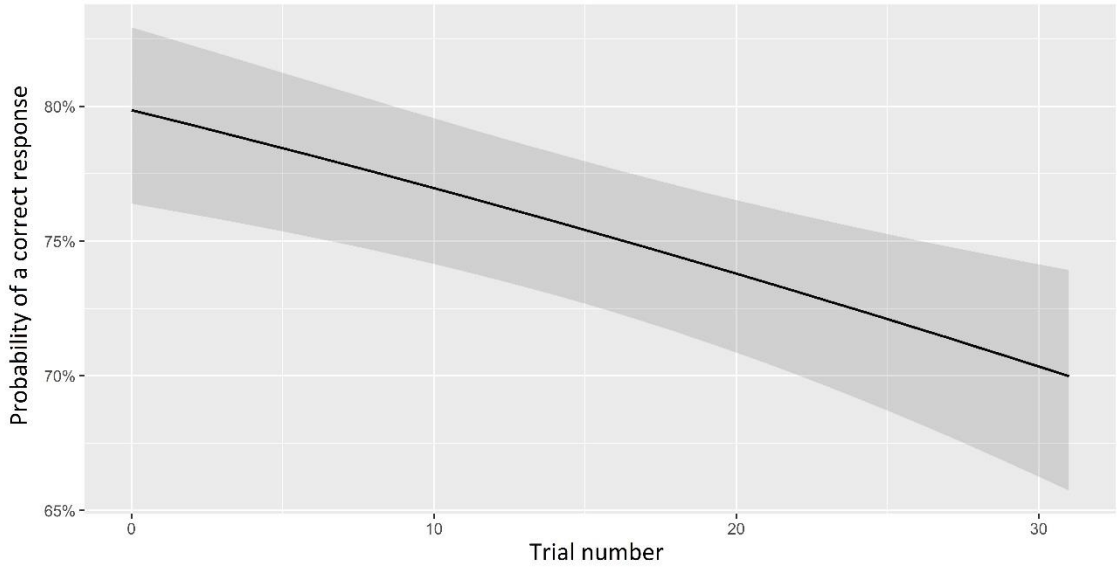


Figure 13: Predicted probability of a correct response in relation to trial number



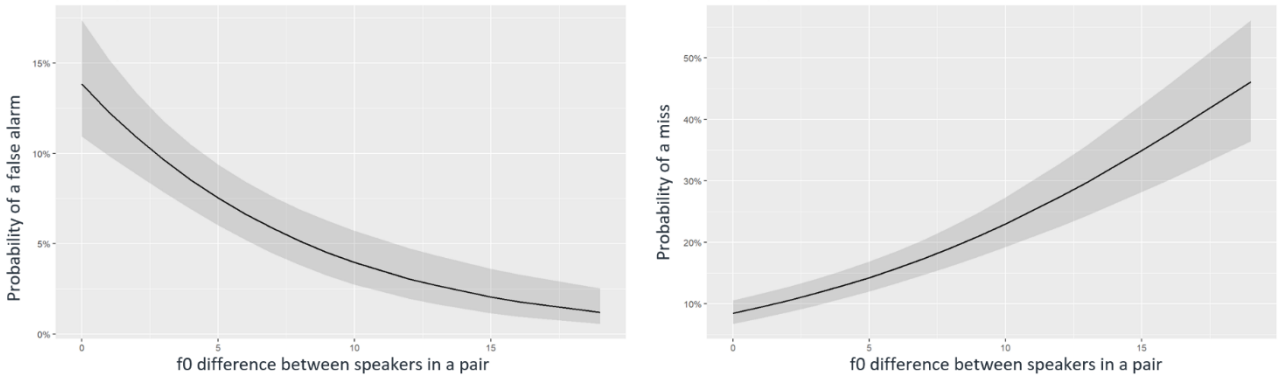
Predicted probability plots were created for the significant effects with the help of the package *sjPlot* (Lüdecke, 2022) for R. As shown in Figure 12, the z-transformed RT was a strong predictor for the correctness of a participant judgement, with a relatively narrow error margin. The faster a judgement was compared with the participant's other judgements, the more likely it was to be correct. The most extreme RTs deserve particular attention, as the fastest

judgements lead to a predicted probability of a correct response just shy of 80%, while the predicted probability of the slowest responses is below 25%.

A similar but weaker effect on accuracy was revealed for trial number (Figure 13). Items that appeared later in the test had a lower probability of eliciting a correct judgement than items featured early in the test. This is especially important as the order of test items was randomised for each participant. However, all predicted probabilities based on trial number fall into the 70 to 80% band.

The effect of f0 difference is more difficult to assess since a low f0 difference was predicted to make it harder to tell different voices apart and easier to tell samples of the same voice together. Consequently, the magnitude of f0-difference is assumed to have the opposite effect on misses and false alarms. The GLMM was therefore run two more times, with the dependent variable changed to false alarms and misses, respectively. In both cases, f0 difference remained a highly significant effect (false alarms: $z = -5.82$, $p < .001$; misses: $z = 8.84$, $p < .001$). As shown by Figure 14, the effect had a different direction in the two scenarios, and was stronger for misses than for false alarms, which further supports our choice of f0 as measure of item difficulty.

Figure 14: Predicted probabilities of false alarms and misses based on f0-difference



6.4 Discussion

6.4.1 Implications for voice processing tests

6.4.1.1 Stimulus quality

In contrast to previous voice processing tests, the present test makes use of naturalistic stimuli, with a similar propositional content across stimuli. Hence, the test exposed the participants to more phonetic detail than the vowels and syllables that were used for the stimuli of the GVMT and BVMT respectively. While the scripted pseudo-speech stimuli of the JVLMT were a huge step forward in terms of naturalness, they were devoid of propositional content. While pseudo-speech allows for testing native speakers of different languages, while avoiding a language-familiarity effect (LFE), they lacked forensic realism.

The naturalistic stimuli used here elicited a wider range of participant accuracy than the BVMT, which followed a similar AX design. Since the phonetic complexity of the stimuli is likely to be the driving force behind this effect, this provides a new perspective to the general assumption in the literature that discrimination performance declines with an increased stimulus variability (Smith et al., 2019: 273), as the present test elicited more extreme performances on both ends of the spectrum. In summary, naturalistic stimuli provide a higher degree of ecological validity in situations where the testing focus is on adaptability rather than generalisability.

6.4.1.2 Reaction time effect

In most online-based voice processing tests, including the GVMT, BVMT, and JVLMT, RT measurements were either discarded or not taken. Such measurements were avoided because they might be differently impacted by the hardware, operating system, internet browser and connection speed (e.g. Aglieri et al. 2017: 108). It was decided to include RT measurements as they are not consciously influenced by the participant, and might therefore be a valuable complement to confidence ratings. Large-scale comparisons of different online and offline setups confirm that online studies do not provide the same precision for RT measurements as lab-based studies (Bridges et al., 2020). However, the particular combination of *PsychoPy* as experiment builder and *Pavlovia* as host platform achieved close to millisecond precision. Measurements were precise to 3.5 ms for all tested combinations of web browsers and operating systems (Bridges et al. 2020: 1). Moreover, RT measurements were stable within a given setup, i.e. across the trials of individual participants (Bridges et al., 2020: 21). RT measurements

should therefore lead to interpretable results if a within-participant design is chosen, whereby the RTs of a participant are analysed in relation to the remainder of that participant's RTs (Anwyl-Irvine et al., 2021). To this end, RTs were z-transformed for each participant and included as an explanatory variable in the model. A strong inverse correlation could be identified, meaning that a given participant's faster judgements were more likely to be accurate than the slower judgements. RT speeds can generally be used to distinguish between different types of recognition memory, i.e. familiarity and recollection memory. The enhanced speed of correct decisions could therefore be indicative of a familiarity effect. Whether this is the case would need to be addressed in dedicated studies. If a reliable connection between accuracy and speed was confirmed, however, RT measurements might be a valuable metric for VPs and general voice processing tests.

6.4.1.3 Trial number effect

The effect of test duration on accuracy is rarely discussed for voice processing tests. Yet, the present test found a highly significant effect of trial number, i.e. the position of the item in the test, on participant accuracy. Items were less likely to be judged correctly when they appeared later in the test. The effect could be caused by fatigue, due to the increased cognitive load as the test progresses (Seale-Carlisle & Mickes, 2016). It could also be a memory effect, resulting from the fact that at a later point in the test participants will have encoded more voices than after the initial trials. This might cause some interference when decoding voices during the later trials. Despite being small in our test, the effect size could build up when test batteries are used to establish correlations with other tests, as it was the case for the BVMT, GVMT, and JVLMT. The test battery of the JVLMT, for instance, took participants 65 minutes on average to complete (Humble et al. 2022: 1361).

6.4.2 Implications for earwitness testimony

6.4.2.1 Performance spectrum

Compared to the complex potential combinations of event variables in earwitness scenarios, this test presented participants with a relatively simple task. Yet, considerable between-listener variation was found, ranging from chance performance to near-perfect accuracy. Knowing about the position of an individual witness on this spectrum could help assess the weight of evidence elicited with the help of a VP.

Figure 15 and Figure 16 compare the usefulness of different metrics that can be used to communicate the results of a screening test for earwitnesses (or the results of a voice recognition test in general). Each dot represents a participant; colour and size of the dot indicate the participant's sex and age, respectively. In both figures, the horizontal axis indicates a participant's PC value (accuracy). Thus, the two potential 'super recognisers' (PC of at least 2 SDs above the mean) are the two points furthest to the right in both figures. Conversely, the four points furthest to the left represent the four participants at the opposite end of the spectrum (PC of at least 2 SDs below the mean) .

The 'super recogniser' concept is, however, to be applied with caution in the context of earwitness testimony, as it is dependent on the performance of other participants in the same screening test, i.e. the group norm. It might be more informative for a trier of fact to be presented with a criterial norm for what is (un)acceptable witness behaviour, e.g. by defining a sensitivity threshold based on a d′ score. Figure 15 illustrates the advantage of expressing participant performance by means of d′ rather than PC. While a clear correlation between the two values exists ($r = .97$; $p < .001$), several outliers can be observed who showed markedly higher sensitivity (d′ scores) than other participants with the same PC value (cf. the five female participants highlighted by the black circle).

For the purpose of communicating with a trier of fact, the *positive predictive value* (PPV) might be of even greater informative value (Semmler et al., 2018). The PPV is defined as the proportion of correct identifications in relation to the total number of cases in which the participant came to the conclusion that speakers were identical, i.e. the combined hit and false alarm rates (PPV = HR/(HR+FAR)). By excluding all cases where the listener concluded that speakers were different (misses and correct rejections), the PPV answers the exact question that is raised by VP identification evidence: "given that this witness made a suspect ID, what is the probability that the ID is correct?" (Semmler et al. 2018: 404-405). It thus represents the

precision with which a participant makes correct selections. The present test's results showed a greater correlation between PC and d′ scores ($r = .97$; $p < .001$) than between PC and PPV scores ($r = .79$; $p < .001$). This is unsurprising as both PC and d′ consider all listener judgements, while the PPV is restricted to same-speaker judgements. The advantage of the PPV is exemplified by Figure 16, which shows that a maximal PPV of 1 is not necessarily linked to a high PC score. Thus, the five female participants with a PPV of 1 (cf. Figure 16) had PC scores ranging from 65.6 to 93.8 (only one of them being a potential 'super recogniser'). Note, that these are the same five participants who are highlighted in Figure 15. These individuals had perfect accuracy whenever they concluded that two voices in a pair were produced by the same speaker. For the present test, which contains equal numbers of same-speaker pairs and different-speaker pairs, this means that they did not produce any false alarms. This high level of precision when making same-speaker judgements is not necessarily reflected by the PC score, as some listeners with a PPV of 1 produced a high number of misses when making different-speaker judgements, resulting in an overall lower accuracy.

In analogy to the PPV, a negative predictive value (NPV), could be calculated for cases in which no identification was made by the witness to indicate the proportion of correct rejections in relation to all cases in which no identification was made in the screening test (NPV = CRR/(CRR+MR)).

Figure 15: Relationship between percentage correct (PC) and D prime score

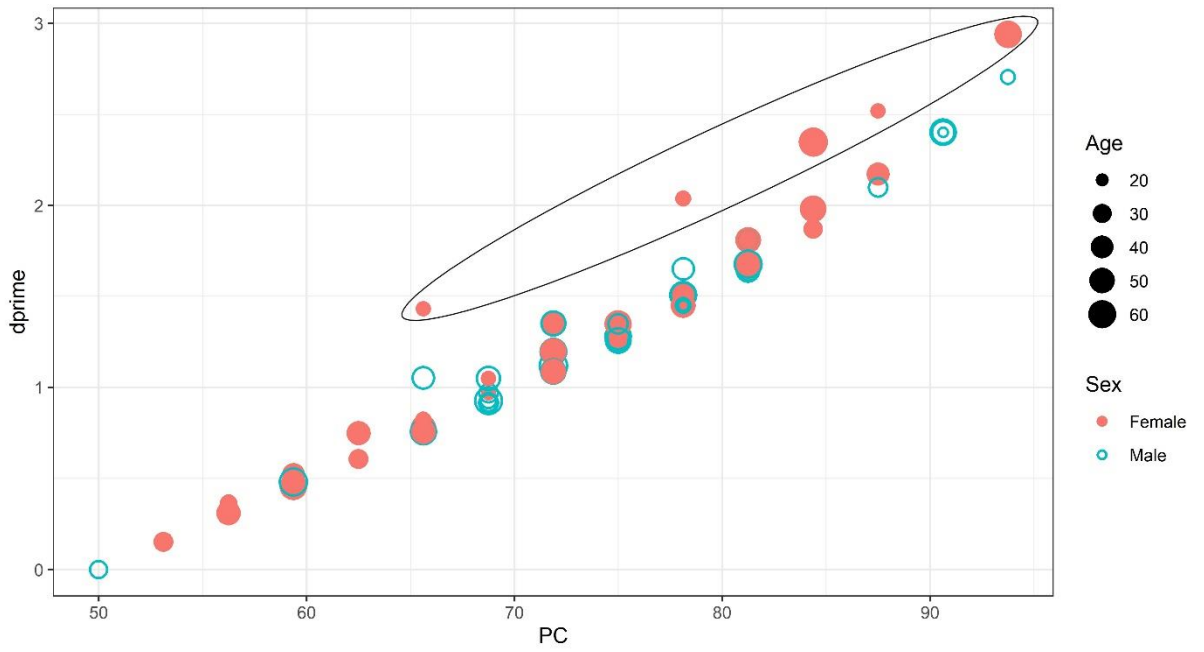
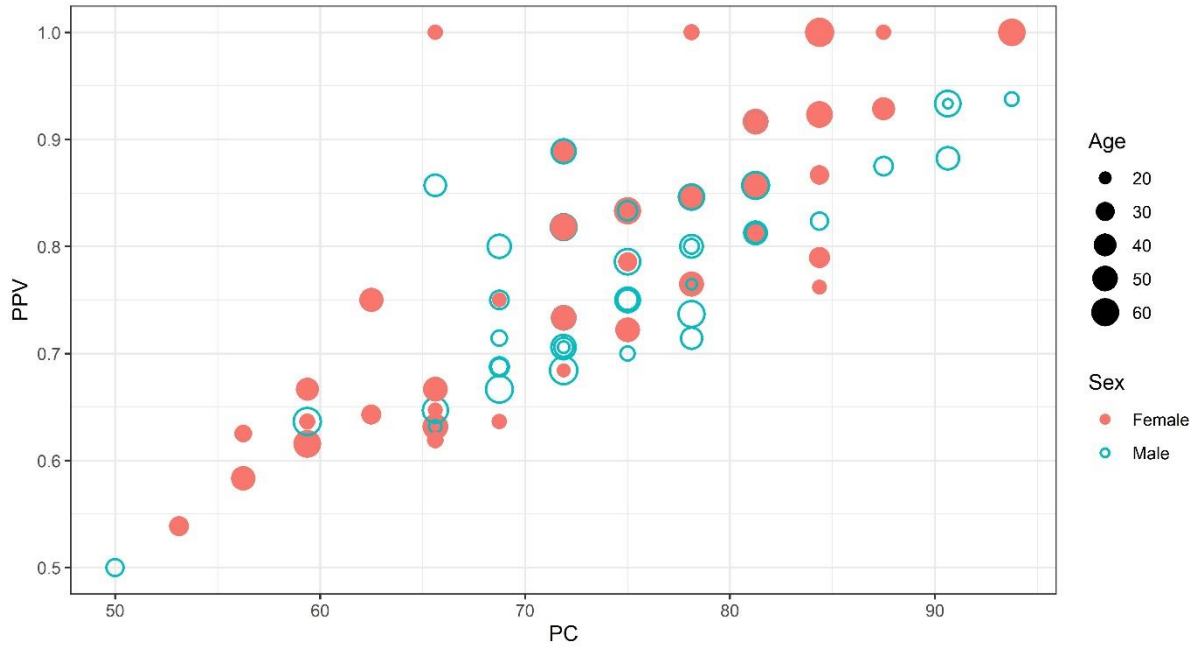


Figure 16: Relationship between PC and positive predictive value (PPV)



6.4.2.2 Test design

The present test confirmed the low correlation between confidence ratings and accuracy (Kerstholt et al., 2004; Yarmey, 1995), at the same time as revealing a strong effect of RT on accuracy. RT might therefore be a valuable metric for earwitness screening tests if such a procedure were to be introduced. RT measurements might also bear a similar potential for VPs with a sequential mode of stimulus presentation. Such parades might further benefit from additional research into the observed trial number effect. The effect corroborates observations made by Zetterholm et al. (2012) that voices occurring at a late point in a VP are more likely to be identified as the culprit.

If this trial number effect is a memory effect, this might also have implications for the execution of a complementary screening test for earwitnesses. That is, it might be advisable to conduct such a test after the VP, to prevent a scenario in which the memory of the screening test's stimuli interferes with the witness's performance in the VP.

6.5 Limitations

The test reported here exhibits some shortcomings compared to other voice recognition tests. For instance, participants were exclusively exposed to male voices. If the test claimed general applicability, this would conflict with existing findings regarding own-sex bias in voice recognition tasks (Wilding & Cook, 2000). If used for the intended earwitness screening purpose, the composition of the stimuli might, however, be adapted to the event variables of each case, presenting the listener with only male or female voices, depending on the sex of the suspect.

The present test was intended to function as a baseline test to motivate further research into witness variables. The test mainly assessed voice discrimination skills of lay listeners. It therefore does not address all relevant characteristics of earwitness situations, particularly the role of memory. The GVMT, which tapped into short-term memory, elicited a significantly wider PC range (62.5%) than the present test (43.75%). The JVLMT was designed to address long-term voice memory (Humble et al. 2022: 1364) and elicited an even wider accuracy range (90%). A screening test for earwitnesses should therefore address memorisation capabilities, given the inevitable delay between a criminal event and the opportunity to conduct a VP. While the type of memorisation might impact witness accuracy, e.g. incidental as opposed to intentional memorisation (for an overview see Clifford, 1980), it is unclear whether this difference can be addressed by a screening test, in which the subject is necessarily aware of the task.

Lastly, it is important to consider that the present study did not include a retest. Consequently, it is unclear whether the participants who demonstrated high accuracy in this test would achieve a comparably high score in a different test that follows the same design principles. This reiterates the importance of applying the term “super recogniser” with caution in relation to the findings of the present study. The BVMT (Mühl et al. 2018), which is similar in structure to the present test, had a high retest reliability (r = .86). However, further empirical studies are needed to assess whether a test that uses naturalistic stimuli would elicit the same degree of consistency.

6.6 Conclusion

An AX voice discrimination task with naturalistic stimuli showed that lay listeners differ markedly in their ability to recognise unfamiliar voices, with accuracy ranging from chance level to near-perfect performance. Participants differed markedly in recognition accuracy (mean = 75%, range = 50% - 93.8%). Two potential 'super recognisers' were identified as well as four participants at the opposite end of the spectrum. The elicited accuracy range exceeds that of the BVMT, which followed a similar design but employed less complex and less natural stimuli.

While f0 difference between stimuli correlated with participant accuracy in this baseline test, a fully-fledged screening test for earwitnesses would benefit from a more comprehensive approach to item difficulty. That is, such a test should also consider further speaker-specific vocal features, such as voice quality. A screening test would also need to match certain inevitabilities of the earwitness condition. First and foremost, memory needs to be addressed to a greater degree. Furthermore, research is needed to better understand the impact of speaker gender, age, geographic background, and emotionality in such a test. The present test did not reveal correlations between test performance and definable participant characteristics, such as age, sex, or confidence. This is illustrated by the two potential super recognisers identified by the present test, a 58-year-old woman from Stockport (PC: 93.75; d′: 2.94; PPV: 1; confidence: 5/6) and a 24-year-old man from London (PC: 93.75; d′: 2.70; PPV: .93; confidence: 4/6). The findings question an expert witness's ability to estimate the credibility of VP evidence based on witness characteristics. A screening test might therefore better fulfil this function as it can provide the trier of fact with a d′ or PPV value that reflects participant behaviour. The test items can be adapted to the event variables of the crime; thus, creating new system variables for the legal system.

Finally, analysis of participant responses revealed two general trends. First, RT was a strong predictor for identification accuracy, in that faster responses indicated more accurate judgements. Secondly, participant accuracy declined over the duration of the test. Both trends might help inform the construction of VPs with a sequential mode of presentation.

7. Test 2 - Individual differences in voice recognition

7.1 Purpose

Test 1 demonstrated that phonetically untrained listeners possess varying levels of voice discrimination ability. The current test (Test 2) additionally introduces memory processes into the testing paradigm. Therefore, the latent ability being measured is voice recognition rather than mere voice discrimination. In accordance with the implicational hierarchy outlined in Section 5.1, voice recognition is considered a more complex task than voice discrimination.

As the testing strategy of this study was to introduce one specific new source of complexity between tests (in this case memory), other variables that may have an impact on participant accuracy, were kept consistent. For instance, the stimuli used for the present test had the same length and quality as the stimuli used in the previous discrimination test. Moreover, the stimuli were sourced from the same corpus and a similar participant cohort was recruited.

Since listeners and speakers were not familiar with each other, latent ability did not comprise speaker individualisation, which would require some form of familiarity with the speaker (cf. Section 3.1). Consequently, the test's primary implications concern unfamiliar earwitness testimony.

As explained in Section 3.1, a voice recognition test should clearly define the memory processes involved. The present test mirrors the study setup of the *Glasgow Voice Memory Test* (GVMT), which tapped into short-term memory (Aglieri et al., 2017: 108) by requiring participants to memorise eight voices between a study and a test phase. The present test was thus assumed to primarily tap into short-term memory as well, while the more complex stimuli may have been more cognitively demanding than the GVMT's vowel stimuli.

The inclusion of a memory component in this test made the task more analogous to an earwitness's real-world experience. However, the construction of VPs is time-consuming, so that earwitnesses are typically tested after relatively long delays. Consequently, real-world earwitness testimony also comprises a long-term memory component, which was not emulated here.

Test 1 revealed reaction time (RT) to be a strong predictor of accuracy in voice discrimination tasks, suggesting implications for voice lineup design and earwitness screening (Section 6.4.1.2). RTs were therefore measured in the present test as well. Furthermore, this test required participants to rate the target voices' distinctiveness during the study phase. As argued in Section 2.3.1, voices that are less prototypical (and thus more distinctive) may be

recognised more easily than less distinctive voices. Thus, distinctiveness ratings were expected to correlate with participant accuracy. Distinctiveness ratings were impractical in the previous AX discrimination test, as they would have interrupted the individual AX trials, potentially confounding immediate discrimination accuracy and RT measurements. However, the study-then-test design used for the present test allowed for collecting distinctiveness ratings separately during study, avoiding a disruption of RT measurement during test.

7.2 Methodology

The experimental procedure and recruiting procedure were approved by the ethics committee of the University of York's Department of Language and Linguistic Science.

7.2.1 Participants

One hundred new participants (50 male, 50 female) were recruited through *Prolific* with the aim of obtaining a representative sample of the British population. The age range was 20 to 70 years, with a mean of 44, a median of 43.5, and standard deviation of 12 years. The aim was for an equal number of male and female participants and a median age close to the British population median of 40.7 years.[26] In order to participate, individuals had to be UK nationals, current UK residents and identify as native English speakers. Users with a Prolific score lower than 95/100 could not take part; all participants recruited had a score greater than or equal to 96 (mean 99.5). The comparable age, gender, and geographic makeup between the newly recruited group and the group recruited for Test 1 enables direct comparison of both tests' results. Recruitment occurred over two days in four identical calls, staggering based on age and gender to best sample these demographics. Participants who took part in Test 1 could not participate to avoid memory of the first test's stimuli being a confounding variable. All participants reported normal or corrected-to-normal hearing. Participants were paid £1.75 and completed the task in 14 minutes on average (equivalent to an hourly rate of £7.50).

[26] https://www.statista.com/statistics/281288/median-age-of-the-population-of-the-uk/; last accessed 25/01/2023.

7.2.2 Stimuli

The stimuli were again sourced from task 1 of the DyViS corpus (Nolan et al. 2009), whose ecological validity for earwitness research has already been discussed in Section 6.2.1. The corpus's main advantages are that the speakers are matched for accent and demographic criteria (male speakers, age range = 18 – 25 years). Moreover, the featured SSBE accent mitigates the risk of an 'accent-familiarity effect' (Njie et al., 2022) when recruiting listeners from different parts of the UK (cf. Section 6.2.1).

Using the same method described in Section 6.2.1, two 10-second-long stimuli were extracted from each of the 100 DyViS speakers. Two stimuli were collected per speaker, i.e. one stimulus that could be used for the first exposure to the voice and one for subsequent testing. Each stimulus contained one or more complete intonation phrases and at least seven seconds of net speech. The propositional contents of the designated 'study' and 'test' stimuli did not overlap. That is, certain topics were reserved exclusively for study stimuli and others only for test stimuli, regardless of the speaker.

Twelve speakers were excluded due to particularly distinctive vocal features that made them stand out from the rest of the speaker pool, such as non-SSBE regional accent features or disordered speech (cf. Section 6.2.1). The remaining 88 speakers were split in two groups, with 48 used for the discrimination test (Test 1) and 40 speakers used for the present test. Consequently, none of the DyViS speakers were featured in both tests.

The 40 speakers available for the present test were sorted by the average f0 difference between their prospective 'study' and 'test' stimuli. The eight speakers with the highest f0 difference were excluded to homogenise the speaker pool.

The final speaker pool of 32 speakers had a mean f0 of 104.7 Hz (range = 83.5 – 137 Hz, SD = 14.25 Hz), which is comparable to the speaker pool of Test 1 (Mean 106.7 Hz, range = 78.5 – 142 Hz, SD = 14.4 Hz). Two separate stimulus lists ('A' and 'B') of 16 speakers each were created from these 32 speakers. The lists were similarly composed in terms of the featured speakers' average fundamental frequencies (measured across study and test stimulus for each speaker): List A had a mean f0 of 104 Hz (range = 83.5 – 134 Hz, SD = 13 Hz) and List B a mean of 105 Hz (range = 85 – 137 Hz, SD = 14 Hz). Shapiro-Wilk tests (Shapiro & Wilk, 1965) indicated that these mean speaker f0 values were normally distributed in both lists (List A: $W(16) = .95$, $p = .42$; List B: $W(16) = .94$, $p = .30$). [27] It was assumed that a stimulus list with

[27] Alpha level = 0.05. Null-hypothesis = normal distribution.

a normal distribution of average f0 values (within a range typical for the relevant population) would allow for eliciting normally distributed accuracy scores. The reason for having two similar stimulus lists with different voice pairs was to assess whether this principle of test calibration by means of f0 composition would work irrespective of the specific stimuli used. In the GVMT (Aglieri et al., 2017), which followed a similar study-then-test design, no measures were taken to control stimulus difficulty (cf. Section 5.3.2.1).

The stimuli were modified in the same way as those used in Test 1. All files were converted from stereo to mono, the peak amplitude normalised (relative of a peak level of -1 dBFS), and any DC offset removed. These modifications were performed using the audio editing software *Audacity* (version 2.3.3). In a few cases, mild background noise was reduced using Audacity's "noise removal" feature. The goal of these measures was to eliminate differences in the original recording quality as a potential confounding variable. Modified stimuli were saved as 44.1 kHz, 16 bit .wav files (705 kBit/s output).

7.2.3 Study Procedure

Participants had to meet the same eligibility criteria as those in the voice discrimination test. This included access to a quiet room, to a device with a physical keyboard (PC or laptop) and to earphones/headphones. Furthermore, they had to be able to complete the task in one sitting. Participants were informed that the test would take approximately 15 minutes, as confirmed during piloting. First, a *Qualtrics* survey elicited informed consent and participant metadata including age, sex, nationality, place of birth, native language(s), and foreign language skills. Participants were then redirected to the voice discrimination test hosted on *Pavlovia*. An information text explained the testing procedure in detail.

Prior to beginning the test, participants were required to complete a sound check to verify their earphones or headphones were functioning and to adjust volume as needed. The sound check utilised a recording of unintelligible crosstalk[28], i.e. the same clip that had served as a cognitive distractor during the AX discrimination trials in Test 1. Playing a voice stimulus was avoided at this point in order not to confound the voice study phase that was to follow.

The participants were then assigned one of the two stimulus lists by means of randomisation without replacement. Participants were not aware of this step. While an equal distribution between the lists was intended, the randomiser made false assumptions about

[28] https://www.youtube.com/watch?v=h2zkV-l_TbY&t=9282s; last accessed 25/09/2023.

recruitment numbers for each list due to several participants starting but not completing the study. This resulted in a slight over-recruitment for List B (N = 52) compared to List A (N = 48).

The study phase exposed participants to eight target speakers from the allocated stimulus list. The target speakers were predetermined for each list and thus identical for all participants allocated to the same list. However, to control for order effects, the eight voices were presented in a randomised order.[29] Each target speaker's study stimulus was played three times consecutively, with a three-second silent interval between repetitions, accompanied by an on-screen countdown from three to one. During each stimulus playback, the screen displayed the number of the current voice (out of eight) and repetition (out of three). After hearing all three repetitions of a target speaker's study stimulus, participants rated the distinctiveness of the speaker's voice on a scale from one ('very inconspicuous') to six ('very distinctive'). The study phase proceeded largely self-paced, requiring participants to press the space bar of their keyboard after the distinctiveness rating before being able to advance to the next voice. However, participants were discouraged from taking long breaks.

After completion of the study phase, an info text explained the procedure of the test phase, which assessed the participants' ability to recognise the eight speakers presented in study. The test phase consisted of 16 randomised trials, one for each speaker on the stimulus list. The eight speakers who were not featured in the study phase functioned as foils. Consequently, the study phase comprised an equal number of target-present and target-absent trials; trial order was randomised.

In contrast to the study phase, the test stimuli were only played once. During each trial, participants listened to the test stimulus, while a left-facing and a right-facing arrow icon were displayed on the respective sides of the screen, with the corresponding labels "OLD" (left arrow) and "NEW" (right arrow) underneath. The participants then pressed the left or right arrow key depending on whether they thought the voice was "old" (featured in study) or "new" (a foil voice). Reaction times (RT) between the onset of the stimulus and the keyboard entry were measured, unbeknownst to the participants. Each test trial ended with the participant's keyboard entry. After each trial, a screen with a progress bar was presented indicating the proportion of trials completed. The space bar had to be pressed to start the next trial, so that the test phase

[29] In the GVMT, which followed a similar design, the order of the study stimuli was not randomised.

was predominantly self-paced. Participants were instructed only to take pauses between trials and not during trials.

At the end of the test, participants provided an overall rating of their confidence in having made correct old/new judgements, using a six-point Likert scale. The average participant completed the entire study (*Qualtrics* questionnaire + *Pavlovia* voice recognition test) in just under 14 minutes.

7.2.4 Hypotheses

Several hypotheses were made regarding the outcome of the test, based on the assumed testing hierarchy (1), the results of other voice recognition tests (2), earlier findings from Test 1 (3-4), and the research literature (5-6). It was hypothesised that:

(1) The test would present listeners with a more difficult task than the previous voice discrimination test (Test 1), due to the addition of memory processes.

(2) The test would present listeners with a more difficult task than the voices version of the GVMT (Aglieri et al., 2017), which followed a similar design but used less complex stimuli.

(3) The test would demonstrate that f0 differences between speakers influence item difficulty and thereby test calibration, as observed in Test 1.

(4) The test would reveal similar reaction time and trial number effects as Test 1.

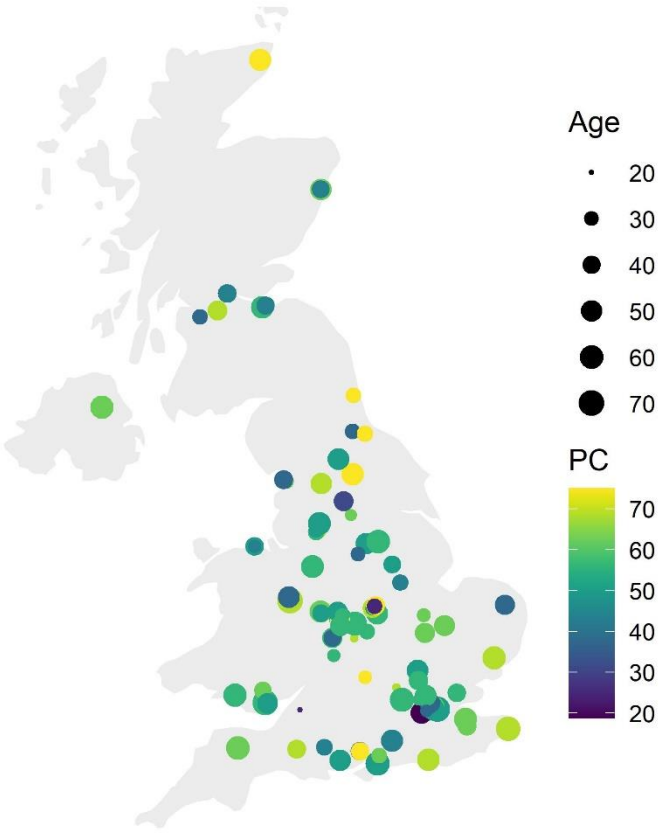
(5) The test would not reveal an impact of confidence on accuracy, as such a relationship is rare in the literature.

(6) The test would reveal that, in line with the literature, distinctive voices will be recognised with higher accuracy.

7.3 Results

Figure 17 shows the geographical location of the 100 participants within the UK. Each dot represents one participant. The size of the dots indicates the participant's age, while the colour reflects the participant's accuracy score in the present test.

Figure 17: Geographical map of the participants' locations within the UK



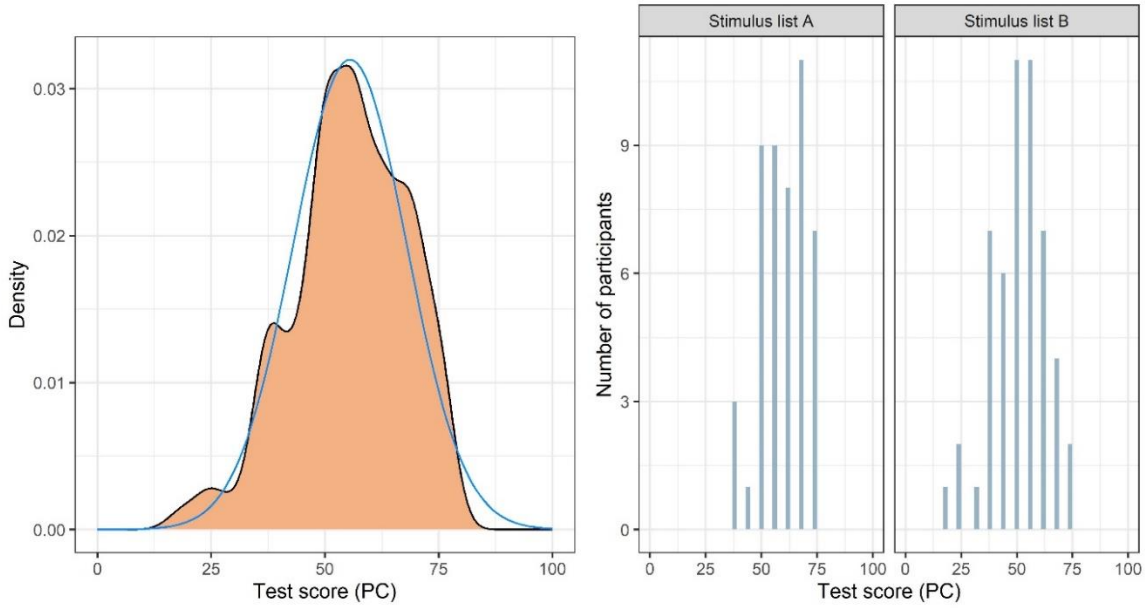
7.3.1 Participant-based analysis

The full participant sample (N = 100) achieved a mean accuracy (PC) score of 55.5% (SD = 12.47) on the present test, with individual scores ranging from 18.75% to 75%. As shown in Figure 18 (left panel), a density plot based on the of the full sample's PC score distribution closely overlaps with a normal curve. The Jarque-Bera Test (Jarque & Bera, 1987) confirmed normal distribution (p = .18).[30]

[30] Note that the Jarque-Bera Test's true alpha level for a sample size of 100 is 0.1. Null-hypothesis = normal distribution.

However, the histograms in the right panel of Figure 18 indicate that only List B elicited normally distributed PC scores, whereas the distribution generated by List A appears negatively skewed. Shapiro-Wilk tests verified that List B PC scores were normally distributed ($W(52) = .97$, $p = .15$), whereas List A scores violated normality assumptions ($W(48) = .95$, $p = .01$).

Figure 18: PC score distributions for the newly recruited sample (N = 100)



The implications of these distributional differences for further analysis must be carefully evaluated. As discussed in Section 5.3.2, some of the most notable voice processing tests elicited non-normal distributions in their validation versions. For instance, the Shapiro-Wilk test revealed strong normality violations for the GVMT's PC distribution ($p < .001$), which in turn led to ceiling effects (Aglieri et al., 2017: 100). Moreover, the GVMT's PC distribution was platykurtic, with a peak broader and lower than that of a normal curve (Aglieri et al., 2017: 100). Nonetheless, the creators of the GVMT considered the full sample for statistical analysis by restricting their methods to nonparametric tests that do not rely on normality (Aglieri et al., 2017: 100).

The GVMT is an important point of reference as it employed the same study-then-test design as the present test, down to the number of target and foil voices. A key difference, however, is that the authors of the GVMT did not try to calibrate their test, whereas an f0-based method was assumed to sufficiently calibrate the present test (hypothesis 3). The creation of two similarly composed stimulus lists was intended as a proof of concept for this method.

Consequently, the different distributions produced by List A and B allow for an initial assessment of hypothesis 3. On the one hand, calibration had some effect on both lists. For instance, an IRT-based 'Rasch model' – fitted with R's *ltm* package (Rizopoulos, 2006) – revealed that the lists covered participants' latent ability to a similar extent (List A: 93.38%, List B: 94.57%). Moreover, none of the lists produced a ceiling effect. On the other hand, List B produced a wider accuracy range (18.75 – 75%) which fully comprises that of List A (37.5 – 75%). The list's PC means also differ significantly (List A: 60.29%, List B: 51.08%), as confirmed by a Welch's t-test ($t(97.25) = 3.97$, $p < .001$). Overall, List B indicates that test calibration can be achieved by means of controlling the f0 composition of the stimulus pool, partially confirming hypothesis 3. This finding is, however, relativised by the less ideal PC distribution of List A, indicating that this method does not work irrespective of specific stimuli used.

Based on the above explanations, the subset of List B participants is particularly relevant for further analysis. Table 7 provides summary statistics for the full participant sample of present test, the List B subset, the previously conducted voice discrimination test (Test 1), and the voices version of the GVMT. A comparison with the BVMT (Mühl et al., 2018) is not deemed necessary, as it was in several ways outperformed by Test 1, which was also a discrimination task (cf. Chapter 6).

Table 7: Comparison of the present test's results with a selection of other tests

	Min	*Max*	*Mean*	*Median*	*SD*
Present Voice Recognition Test (full sample), N = 100 (50 male)					
Age	20	70	44	43.50	12.0
PC	18.75	75	55.5	56.25	12.47
d´	-1.59	1.25	0.26	0.28	0.61
Present Voice Recognition Test (stimulus list B), N = 52 (26 male)					
Age	20	67	42.5	42	11.2
PC	18.75	75	51.08	50	12.5
d´	-1.59	1.18	0	0	0.62
Previous Voice Discrimination Test (Test 1), N = 100 (50 male)					
Age	18	68	36	38.85	13.8
PC	50	93.75	75.00	75.00	9.06
d'	0.00	2.94	1.38	1.35	0.57
Glasgow Voice Memory Test (voices version), N = 1120 (337 male)					
Age	18	86	26.7	*NA*	11.1
PC	37.5	100	78.15	*NA*	10.95
d'	-0.67	3.07	1.66	*NA*	0.69

Mean performance in the present voice recognition test was lower than in the previously conducted voice discrimination test (Test 1). This applies to both participant accuracy (PC) and sensitivity (d′). Whereas the voice discrimination test (Test 1) produced an average PC of 75%, which is the exact mid-point between perfect accuracy (100%) and chance level (50%), the mean PC of the present test was practically at chance level for the List B subset (51.08%) and slightly higher for the full participant sample (55.5%), confirming the hypothesis (1) that the present test presented listeners with a more difficult task than Test 1. These means are directly comparable as chance level was identical in both tests. Moreover, both the full sample and List B subset of the present test elicited a PC span ($Max_{PC} – Min_{PC}$) of 56.25%, surpassing Test 1's span of 43.75 %. However, despite generating a wider spread of accuracy scores, the present test did not match the discrimination test's ability to differentiate between top performances. While Test 1 found two potential SRs, no participant reached a PC score of at least two standard deviations above the mean in the present test. On the other hand, three participants were identified at the opposite end of the spectrum with PC scores of 18.75 (N = 1) and 25 (N = 2). Note that all participants in this range are from the List B subset. They fall into the potential phonagnosic range irrespective of whether the mean and SD of the full sample are used or that of the List B subset.

A comparison of the present test with the GVMT (Aglieri et al., 2017) confirms the hypothesis (2) that the present test was more difficult. The GVMT elicited a mean PC of 78.15% which is significantly higher than that of the present test (55.5%). A comparison with the full sample of the present test is feasible in this case as calibration issues are of minimal concern for a comparison with the GVMT, which was not calibrated at all. These means are directly comparable, as in addition to a shared chance level, both tests consisted of 16 test trials, so that PC scores are bound to the same increments of 6.25%. While the GVMT's elicited accuracy span of 62.5% exceeded that of the present test (56.25%), comparison is hindered by the GVMT's ceiling effect. A more informative comparison of performance ranges can be made on the basis of D prime. The participants of the GVMT's validation version achieved a mean d′ of 1.66, with individual scores ranging from -.67 to 3.07. The top figure is close to an ideal d′ value of approximately 3, which indicates exceptional ability to distinguish between signal (target voices) and noise (foil voices). On the other hand, the present test elicited an average d′ value of about .26, which is essentially equivalent to guessing, and a total range from -1.59 to 1.25 (for the full sample). Since the present test closely mirrors the GVMT's design, the substantial decrease in ability to tell apart target and foil voices is most likely due to the key

difference between both tests, i.e. stimulus complexity. In turn, this means that the exceptionally high d′ scores elicited by the GVMT are not necessarily indicative of exceptional listener ability, but rather of low task difficulty, as the test reached a ceiling.

7.3.2 Trial-based analysis

A trial-based analysis was conducted in order to determine the effects that had a significant impact on the outcome of an individual test trial, i.e. each old/new judgement. To this end, a generalised mixed-effects model (GLMM) was fitted, using the *glmer* function of *R*'s *lme4* package (Bates et al. 2015). A generalised model was chosen because the dependent variable was a binary "correct response" score, which took the value of either 1, when a judgement was correct, or 0, when a judgement was incorrect. Seven independent variables were considered as fixed effects: Participant age, participant sex (reference level "male"), participant confidence, the trial number (i.e. a number between 1 and 16 depending on the position of the trial within the test), the participant's reaction time (z-transformed within individual participants' RTs), the f0 difference between a target speaker's study stimulus and test stimulus, as well as the participant's rating of a study voice's distinctiveness (on a scale from 1 – 6). Participant identity was used as a clustering variable.

Crucially, two of these variables only apply to target-present trials, i.e. the f0 difference between study and test stimulus and the voice distinctiveness ratings from the study phase. Consequently, the decision was made to run separate models for target-present and target-absent (foil) trials. The dependent variable in the first model was a correct response in a target-present trial, i.e. a hit being counted as 1 and a miss as 0. The dependent variable in the second model was a correct response in a target-absent trial, i.e. a correct rejection being counted as 1 and a false alarm as 0.

Both models were fitted with the maximal amount of available independent variables; all of the seven variables listed above were used in the target-present model and five variables in the target-absent model (seven minus distinctiveness and f0 difference). No interactions were fitted between individual fixed effects. Participant identity (in the form of participant number) was used as a clustering/random variable in both models. Table 8 summarises the findings of the target-present model and Table 9 the results of the target-absent model. Each model was run for the full sample as well as for the subset of participants who were assigned stimulus list B.

Table 8: GLMM results (Dependent variable: Correct response to target voices)

	Full participant sample (N = 100)			Stimulus list B (N = 52)		
Fixed effects	***z (test statistic)***	***p (significance)***		***z (test statistic)***	***p (significance)***	
(Intercept)	-0.29	>0.7		-0.28	>0.7	
Age	0.32	>0.7		0.31	>0.7	
Sex	-0.74	>0.4		-1.12	>0.2	
Trial number	0.36	>0.7		-0.02	>0.9	
RT (z transf.)	-0.18	>0.8		0.58	>0.5	
Confidence	2.63	<0.01	**	2.56	<0.05	*
Distinctiveness	-0.19	>0.8		1.02	>0.3	
F0 difference	-1.45	>0.1		-2.67	<0.01	**

Table 9: GLMM results (Dependent variable: Correct response to foil voices)

	Full participant sample (N = 100)			Stimulus list B (N = 52)		
Fixed effects	***z (test statistic)***	***p (significance)***		***z (test statistic)***	***p (significance)***	
(Intercept)	-0.32	>0.7		-1.49	>0.1	
Age	-0.16	>0.8		0.74	>0.4	
Sex	1.84	<0.1	.	2.27	<0.05	*
Trial number	-1.20	>0.2		-0.13	>0.8	
RT (z transf.)	-2.00	<0.05	*	-1.87	<0.1	.
Confidence	1.63	>0.1		1.04	>0.2	

In contrast to the results of Test 1 and hypothesis 4, no trial number effect was found by any of the models. In line with hypothesis 4, however, a significant ($z = -2.00$, $p < .05$) RT effect was found for the full participant sample, albeit in target-absent trials only. This effect became insignificant when analysis was restricted to the List B participant subset but maintained a relatively low p-value ($z = -1.87$, $p < .1$). At the same time, a near-significant ($z = 1.84$, $p < .1$) effect of participant sex in target-absent trials became significant ($z = 2.27$, $p < .05$) when analysis was reduced to the List B participant sample. No such effect was found in target-present trials.

Against expectations (hypothesis 5), a highly significant confidence effect ($z = 2.63$, $p < .01$) was found in target-present trials, which reduced to a significant effect ($z = 2.56$, $p < .05$) when just the List B participant group was analysed. The List B participant subset further exhibited a highly significant effect of f0 difference between a test stimulus and the corresponding study stimulus in target-present trials ($z = -2.67$, $p < .01$).

Distinctiveness ratings did not produce significant effects in target-present trials, thus rejecting hypothesis 6. In other words, more distinctive voices were not recognised more accurately. None of the models found a significant effect of listener age.

Considering the directionality of the found effects, the following simplified summary statements can be made:

1. **In target-present trials**, greater **confidence** significantly increased the likelihood of a hit (full sample & List B).
2. **In target-present trials**, greater **f0 differences** between the speaker's study and test stimulus significantly increased the likelihood of a miss (List B).
3. **In target-absent trials**, faster **reaction times** significantly increased the likelihood of a correct rejection (full sample)
4. **In target-absent trials**, **male** listeners were significantly more likely to correctly reject a foil speaker (List B)

7.3.3 Correlations with the voice discrimination test

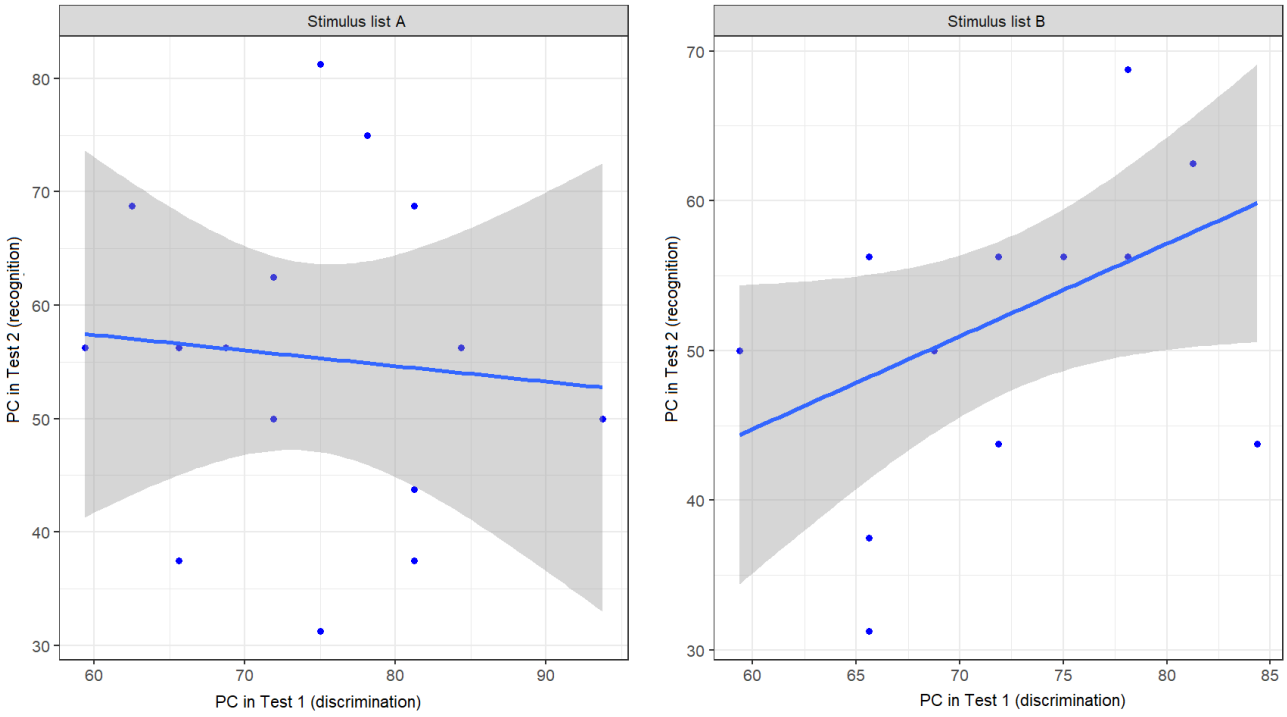
A group of 30 participants who took part in the voice discrimination test (Test 1) were reinvited for the present test to allow for an assessment of possible correlations between the two tasks. The group was comprised of 15 male and 15 female listeners and had an average age of 46.13 years (median = 47, range = 23 – 67). This re-recruited participant group had a mean PC of 73.44% in Test 1 (range = 59.38 – 93.75, SD = 8.63). It also included one of the two SRs identified in Test 1 (female, aged 58). When the re-recruited cohort completed the present test, in between 96 and 99 days had elapsed since they had taken Test 1 (depending on the participant). This relatively long delay was chosen to mitigate bias induced by the participants' memory of the Test 1 stimuli. Note that while both tests used similar stimuli, no stimulus was used in both of the tests, further mitigating memory-induced bias.

The re-recruited group achieved an average accuracy of 53.96% in the present test (SD = 12.33). A t-test (Welch) indicated that this mean did not differ significantly from the full sample (N = 100) of the cohort that was newly recruited for Test 2 ($t(48.3) = -.59$, $p = .55$). Despite these similarities, the results of this group were analysed separately from the newly recruited cohort and not included in the main statistical analysis. The individual PC scores ranged from 31.25 to 81.25%. The top score of 81.25% was achieved by only one participant who would qualify as a SR when judged against the main cohort's mean and SD. Like the main cohort, these re-recruited participants were allocated to either List A (N = 15) or List B (N = 15). In

contrast to the main cohort, both lists produced normally distributed accuracy scores for this re-recruited participant group (List A: W(15) = .97, p = .87; List B: W(15) = .95, p = .60).

Overall, no significant correlations were found between the PC scores achieved in Test 1 and Test 2 (r = .14, p = .43). When the stimulus lists were tested separately, the PC scores generated by List A did not correlate with the participants' performance in Test 1 (r = -.09, p = 0.75). On the other hand, the scores generated by List B showed near-significant moderate correlations with Test 1 scores (r = .49, p = .06). While not significant, this correlation is worth considering in light of the very small sample size. Scatterplots with fitted correlation lines illustrate the relationship between performance on Tests 1 and 2 for the sample of 30 re-recruited participants (Figure 19). The potential SR who achieved a 93.75% accuracy in Test 1 only answered 50% of the present test's trials correctly (the rightmost point in the left panel of the figure). Note, however, that she was assigned to List A in the present test, i.e. the list whose PC scores were generally not correlated with that of Test 1.

Figure 19: Correlations between Test 1 and Test 2 PC scores (by stimulus list)



Other studies have indicated low-to-moderate correlations between voice discrimination and voice recognition tasks. The authors of the BVMT (Mühl et al., 2018: 2189), a voice discrimination task, found a Pearson correlation of .24 (p = .004) between their test and the GVMT, a voice recognition test. Jenkins et al. (2021: 599) found a slightly higher correlation

of .30 ($p < .001$) between these two tests. However, both tests showed ceiling effects, making it difficult to assess the potentially different correlations that better calibrated versions would generate.

7.4 Discussion

7.4.1 Interpretation of significant effects

The PPV model (Lavan & McGettigan, 2023) hypothesises that the processing routes taken by unfamiliar and familiar voice signals in the brain are likely intertwined. Bi-directional interactions are assumed to exist between mainly feature-based unfamiliar and mainly Gestalt-based familiar voice processing. This means that while unfamiliar voice processing is likely to start off as a feature-based process during the formation of first impressions, communication with voice recognition units is possible early on. Moreover, VRUs are not necessarily person-specific, but can also be stereotype-specific, so-called 'personae' (cf. Section 4.2.3.2). This means that unfamiliar and familiar recognition are likely to be functionally integrated (cf. Section 4.2.3.2).

Given this potential importance of familiarity sensations even in unfamiliar voice processing, the RT measurements taken during the test phase are of particular interest, as recognition memory is typically subdivided into recollection and familiarity (cf. Yonelinas, 2002). Familiarity is the sensation that a stimulus has been encountered before. It tends to be a faster, more automatic process than recollection, which is a slower, more deliberate attempt to retrieve information (Yonelinas, 2002). Shorter RTs are therefore indicative of greater familiarity, which may explain the RT pattern found in Test 1. In the present recognition test, a more differentiated pattern emerges, as RTs did not have an impact on the recognition of voices presented in study. However, RT was found to impact participants' responses to the voices that were newly introduced in test, i.e. the foil voices. The effect was only significant for the full participant sample, but near-significant in the List B participant subset. Faster judgements in these target-absent trials significantly increased the likelihood of a foil being correctly rejected. Consequently, longer RTs increased the likelihood of a false alarm.

The reasons behind this pattern may be rooted in the experimental setup, since participants were not aware that RT measurements were taken. They were thus not under any pressure to make fast judgements. This may have led to a situation, in which listeners did not

make an immediate "old" judgement when hearing a target voice, even when they had an immediate familiarity sensation. Instead, the participants may have tried to nonetheless use recollection processes in these situations, until sufficient confidence was established that the judgement was correct. This would also explain why false alarms, which are also "old" judgements, were made after longer delays than correct rejections, reflecting the same attempt of active information retrieval. In contrast, listeners might not have felt this need for active recollection when they had the intuition that a speaker had not been present in study, as they might have considered their task to be above all the recognition of the target voices, rather than the correct rejection of the foils.

The present findings on confidence provide further tentative support for this hypothesis about participants' task perception. It was the strongest effect found in the data in terms of significance as well as in terms of effect size. The vast majority of studies on unfamiliar voice recognition have found participant accuracy and confidence to be generally unrelated (cf. Schweinberger & Zäske, 2018: 547). However, the separate analyses of target-present and target-absent trials conducted here, allow for a more nuanced discussion. In this connection, high participant confidence was found to be predictive of correct responses, but only in target-present trials. Remember, that this was an overall confidence rating, i.e. only one confidence rating was taken at the end of the experiment. A high overall confidence rating obtained after the test meant that the participant has likely produced a higher hit rate (and lower miss rate) than a participant with a lower overall confidence rating. No inferences can, however, be made about the correct rejection rate (or false alarm rate) from these confidence ratings. Based on this insight from the trials-based analysis, it should be possible to find significant correlations between confidence and hit/miss rate in a participant-based correlation test, but not so between confidence and correct rejection/false alarm rate. This could be confirmed; a highly significant moderate correlation was found between participants' confidence ratings and hit/miss rates ($r = .34$ / $r = -.34$, $p = .003$, $N = 100$). while no significant correlation existed between confidence and correct rejection/false alarm rate ($r = .17$ / $r = -.17$, $p = 0.13$, $N = 100$).

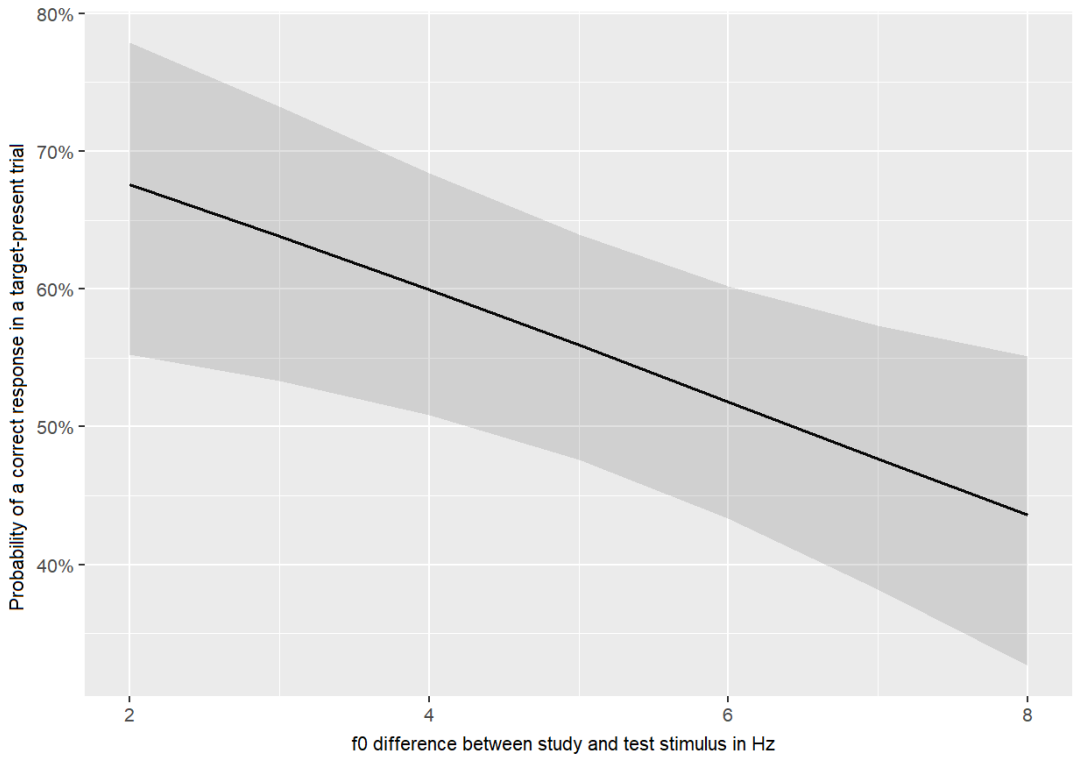
This is an important finding for earwitness research, where the prediction of false alarms is of particular interest. Moreover, it reiterates that in voice recognition tests – including VPs – the listeners might predominantly perceive their task to be the recognition of the targets rather than the correct rejection of the foils (which is an equal part of the actual task). Therefore, their confidence ratings may be more reflective of hits than of false alarms.

7.4.2 Calibration of voice processing tests

The calibration method impacted the two stimulus lists differently. Comparing the generalized linear mixed models (GLMM) run on the full sample to those run on just the List B participants reveals some similarities and differences. It can be observed that most significant effects are shared by both models, although changes in effect size and significance can be found.

The calibration discrepancy between the lists indicates that average f0 of the featured speakers was an insufficient criterion for stimulus list composition overall. The distribution of accuracy scores in List A was negatively skewed. This indicates that some speakers featured in List A may have stood out from the overall speaker sample for reasons other than f0, leading to a situation where those speakers were more easily identifiable as targets or foils, respectively. An analysis of the distribution curve indicated that the method worked better for List B. This is confirmed by the trials-based model which found that the f0 difference between a speaker's study and test stimulus had a highly significant effect on the outcome of target-present trials, but only for List B participants. These participants were less likely to correctly recognise target speakers with a high f0 difference between study and test stimulus. This effect, which was the strongest effect across all models, is illustrated in Figure 20 by means of a predicted probability plot.

Figure 20: Predicted probability of a correct response in a target-present trial (List B)



As shown, target-present trials in which the speaker's f0 closely matched that of the corresponding study stimulus had a 70% likelihood of eliciting a hit. This likelihood dropped below chance level for trials with a high f0 difference between study and test stimulus. The size of the effect indicates that f0 difference can be a key factor for recognition accuracy in a test calibrated by average speaker f0. On the other hand, the analysis of the whole sample, which also includes List A participants, indicates that this method does not work irrespective of the specific speakers used.

Better understanding of the many factors that impact the recognisability of certain speakers would have important implications for VP construction. It was hypothesised (6) that distinctive voices may be recognised with higher accuracy. However, analysis of the elicited data did not reveal a significant impact of distinctiveness ratings on recognition accuracy in target-present trials; neither for the full sample, nor for the List B subset. It must be considered that these distinctiveness ratings are not necessarily indicative of a speaker's distinctiveness in phonetic terms, but rather of the distinctiveness as perceived by a particular listener (cf. Section 2.3.1). Consequently, assessments may differ depending on the individual listener's prototype, or exemplars (cf. Section 2.3.1.2).

Finally, the better calibrated List B showed a strong and significant effect of participant sex in target-absent trials. Male listeners were significantly more likely to correctly reject a foil speaker and, in turn, less likely to produce false alarms. Note, however, that all speakers used in this test were male. It would therefore be an overgeneralisation to claim that male listeners are better at avoiding false alarms. The effect may rather be indicative of an own-sex advantage. Skuk and Schweinberger (2013) reported a similar effect in a familiar voice individualisation study with German high school students (discussed in Section 4.1.2.2). Male listeners correctly individualised male speakers more often than female speakers. In contrast, female listeners individualised speakers of both sexes at similar levels. However, their study did not include any target-absent trials as listeners were exclusively presented with familiar speakers. Consequently, the present findings provide a rare example for an own-sex advantage in target-absent trials. An own-sex advantage could not be determined for female participants in the present study due to the lack of female speakers. The authors of the GVMT discovered that female participants recognised female speakers better in their study (Aglieri et al., 2017: 102), while they did not find an own-sex advantage for male listeners.

7.5 Limitations

The current test shares some of Test 1's limitations, as the speech stimuli were sourced from the same corpus of male voices only. This homogeneous sample limits the generalisability of the findings and the ability to fully assess phenomena like the own-sex advantage found for List B participants. Future studies should therefore use a speaker pool balanced for sex.

While the present test introduced a memory component, it did not assess long-term memory over an ecologically valid timeframe. The brief retention interval likely underestimates witnesses' memory capabilities in real earwitness scenarios. Subsequent tests should incorporate delays of days or weeks between study and test phases to better approximate long-term memory demands. The JVLMT (Humble et al., 2022), which introduced only a small amount of long-term processing, already produced a significantly wider accuracy span of 90%.

The ecological validity is further limited by the overall rather short exposure to the target speakers' voices in the study phase (3x10s per speaker), as well as the introduction of eight target speakers within a brief period of time.

Additionally, participants here intentionally memorised the voices, unlike witnesses who incidentally encode voices during a crime. This intentional memorisation may not reflect real-world memory processes and performance. However, the deliberate memorisation was an important pre-cursor for assessing incidental voice encoding, which was separately investigated in a follow-up study (Chapter 8).Overall, the current test provides a controlled increase in complexity from Test 1, but future iterations should increase ecological validity by means of sample diversity, long-term memory assessments, and incidental encoding.

7.6 Conclusion

An unfamiliar voice recognition test was conducted that followed a 16-trial study-then-test design, with eight target and eight foil voices. The test revealed substantial individual differences in latent ability for voice recognition. Mean performance was lower compared to a prior voice discrimination task, confirming the assumed implicational hierarchy. The wider range in performance suggests higher individual variability in voice recognition ability compared to voice discrimination. These findings highlight the importance of including memory processes in potential earwitness screening.

Compared to existing voice recognition tests (e.g. GVMT), the present test provides a more ecologically valid assessment through use of naturalistic voice samples. The results offer useful insights for forensic research on voice recognition abilities. An interesting finding was that participant confidence correlated with individual SDT indices, especially hits and misses. This suggests confidence may be more predictive of performance than typically assumed. In contrast, perceived speaker distinctiveness did not predict accuracy.

Fundamental frequency was a less robust predictor of overall test difficulty in this voice recognition task than in the previous voice discrimination task. Nonetheless, it allowed for predicting item difficulty to some extent. Compared to the discrimination test, response times were a less reliable indicator of accuracy. The data also provided some evidence for an own-sex advantage, aligning with prior literature. However, limitations of the study design preclude conclusive assessment of this effect.

8. Test 3 - Individual differences in task awareness

8.1 Premise

Test 3 was designed to expand the testing baseline by incorporating task awareness into the experimental paradigm. In the most widely used voice processing tests, including the GVMT (Aglieri et al., 2017), BVMT (Mühl et al., 2018), and JVLMT (Humble et al., 2022), participants are aware of the task. They know that the test's purpose is to assess their voice processing skills and they are actively instructed to either compare voices for speaker identity (BVMT), or to memorise voices for later recognition of the speaker (GVMT & JVLMT). Many real-world situations lack this explicit task awareness. This applies to some criminal events. For instance, witnesses may not realise in the moment that they are witnessing a criminal event. This may e.g. apply to cases of telephone fraud. In other cases, witnesses may be aware of their involvement in a crime but focus on other information that they deem more relevant, such as what is being said rather than the voice of the speaker. If task awareness substantially impacts recognition accuracy, determining whether a witness consciously attempted to remember the perpetrator's voice could determine the weight of their testimony.

Task awareness is an estimator variable in the literal meaning of the word, in that it can only be considered in the categorical assessment of a witness's credibility (objective reliability), i.e. by guessing the likelihood of their judgement being correct in light of the known estimator variables. It cannot, unfortunately, be included in a potential screening test for earwitnesses (as suggested in Section 3.5.2), as such a test would necessarily be conducted in the context of a legal investigation, ideally after a VP has taken place. The witness would at this point know that their voice recognition skills are under scrutiny, and thus be aware of the screening test's purpose.

Certain types of tests do not lend themselves well to manipulating participant awareness of the test's purpose. For example, AX discrimination tasks require comparing two stimuli within each trial, immediately revealing the discrimination criterion. After just one trial, participants understand the test demands, preventing researchers from plausibly masking the intent. Thus, it would be impossible to create an unaware version of the BVMT, or Test 1 of the present study, as participants necessarily become aware of the task upon completing a trial. However, this issue can be avoided for tests employing a study-then-test design, such as Test 2 here and the GVMT. In these cases, information about the test's true purpose can be withheld when first exposing participants to voices in the study phase. The actual intent - speaker recognition - is only revealed before the test phase begins.

Consequently, the current test (Test 3) adapts Test 2, the key difference being that participants are not informed of the test's true purpose before the test phase. During the study phase, the test's intent is masked by instructing participants to memorise the content of the study stimuli. Consequently, they assume the subsequent test phase will evaluate their memory for specific details of the content. This task is forensically plausible, as an earwitness may attempt to remember information about what was said during a crime. By withholding the speaker recognition goal until after the study phase, Test 3 prevents participants from approaching the voices analytically during initial exposures.

8.2 Methodology

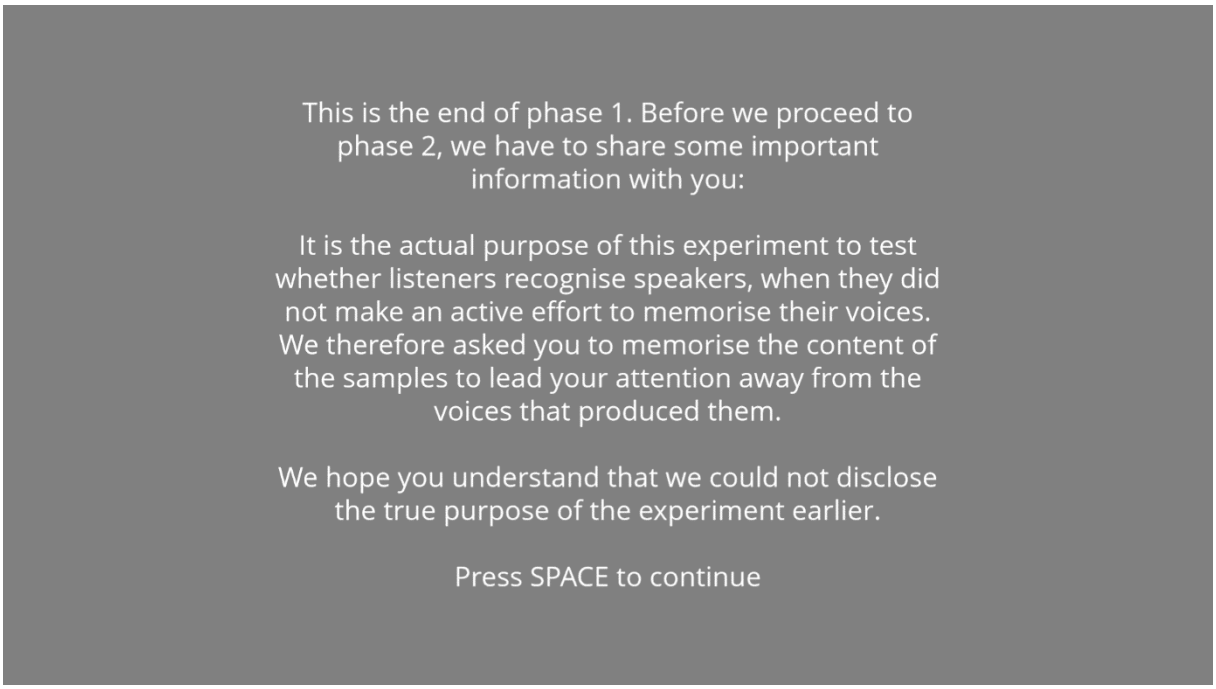
The experimental procedure and recruiting procedure were approved by the ethics committee of the University of York's Department of Language and Linguistic Science.

8.2.1 Differences from Test 2

The design was identical to that of Test 2. A detailed account of this design, including the used stimuli was provided in Section 7.2. In brief, a study-then-test design was employed with eight stimuli in the study phase and 16 stimuli in the test phase (8 targets and 8 foils). Given the more normal distribution of results elicited from stimulus list B in Test 2, all participants in the present test were assigned that list.

The only procedural difference between this test and Test 2 is the instructions given to participants at certain points. Participants were led to believe the test assessed speech (propositional content) memory rather than voice memory. Before the study phase, they were instructed to pay close attention to what was said and to memorise the content as best they could. To increase credibility, they were specifically told they could not take notes or use assisting devices. Study stimuli were repeated three times. For each study stimulus a distinctiveness rating was elicited after the third repetition. In contrast to Test 2, this rated the content's distinctiveness on a 1 ('very inconspicuous') to 6 ('very noteworthy') scale, rather than the voice's distinctiveness. After the study phase had ended, participants saw the following text revealing the actual purpose of the study (cf. Figure 21.)

Figure 21: Information text displayed after the study phase



8.2.2 Participants

Fifty new participants (25 male, 25 female) were recruited through *Prolific*. The age range was 23 – 81 years, with a mean of 45.02, median of 41.5, and SD of 14.47 years. Participants had to be UK nationals, current UK residents and self-identify as native speakers of English to be eligible for the study. Users with a "Prolific score" lower than 95/100 could not participate; all users in the final sample had a score greater than or equal to 96 (mean = 99). Recruitment occurred in two separate calls, over the course of a single day; one call for female and one call for male participants. Participants who took part in one of the previous two tests were not eligible to take part. All participants reported normal or corrected-to-normal hearing. Participants were paid £2.60 and completed the task in 15 minutes on average (equivalent to an hourly rate of £10.40).

8.2.3 Hypotheses

(1) It was hypothesised that the test would present participants with a more challenging task than Test 2.

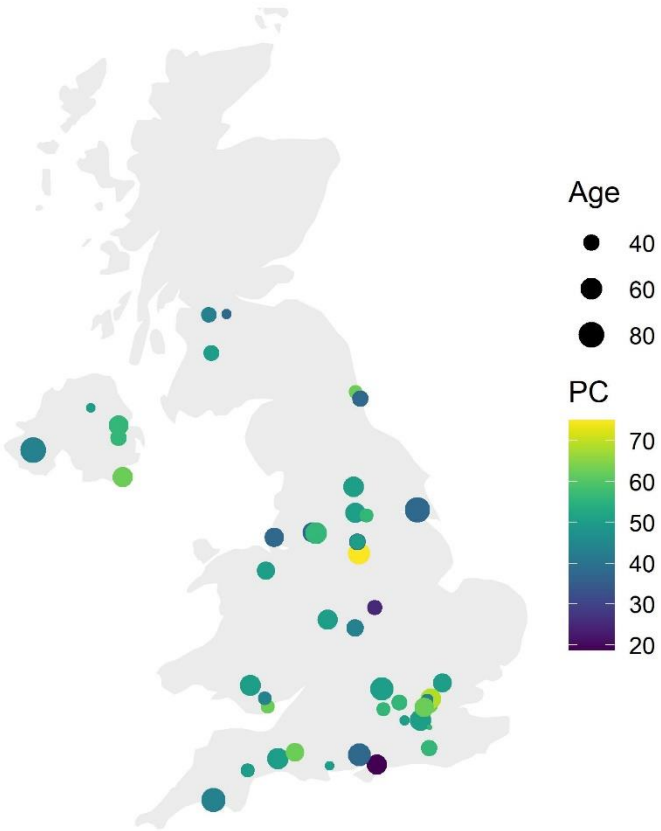
(2) It was hypothesised that higher confidence would be indicative of more accurate judgements in target-present trials, given the similar effect observed in Test 2.

(3) It was hypothesised that f0 difference between study and test stimuli would be correlated with accuracy in target-present trials, as observed for List B participants in Test 2.

8.3 Results

Figure 22 shows the geographical location of the 50 participants within the UK. Each dot represents one participant. The size of the dots indicates the participant's age, while the colour reflects the participant's accuracy score in the present test.

Figure 22: Geographical map of the participants' locations within the UK



8.3.1 Participant-based analysis

Table 10 provides summary statistics for the present test, juxtaposed with the outcomes of Tests 1 and 2. The most informative comparison is with the List B participant subset of Test 2, as the full participant sample of the present test was assigned List B.

Table 10: Summary statistics for the present test compared to Tests 1 and 2

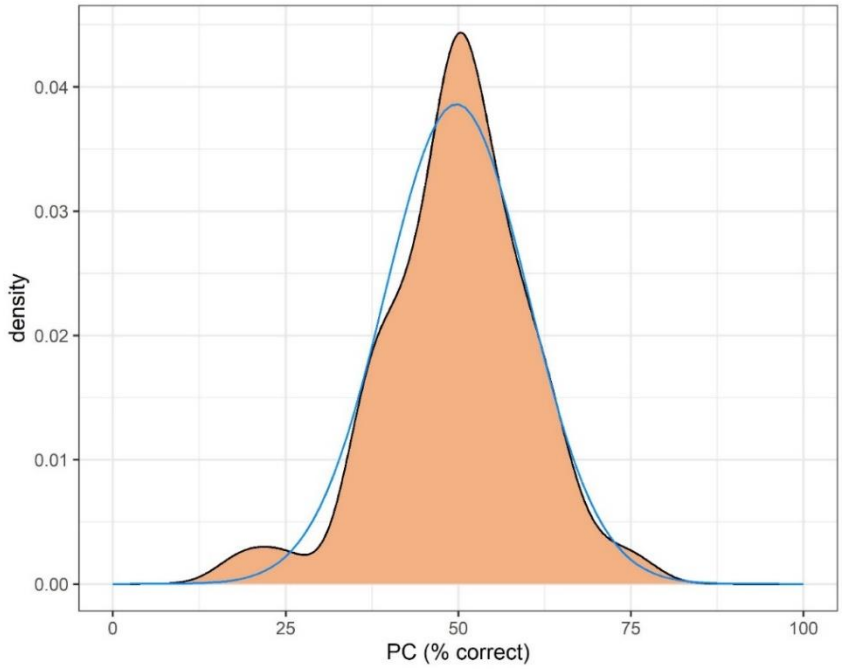
	Min	*Max*	*Mean*	*Median*	*SD*
Test 1 - Discrimination Test, N = 100 (50 male)					
Age	18	68	36	38.85	13.8
PC	50	93.75	75.00	75.00	9.06
d'	0.00	2.94	1.38	1.35	0.57
Test 2 - Recognition Test (full sample), N = 100 (50 male)					
Age	20	70	44	43.50	12.0
PC	18.75	75	55.5	56.25	12.5
d´	-1.59	1.25	0.26	0.28	0.61
Test 2 - Recognition Test (stimulus list B), N = 52 (26 male)					
Age	20	67	42.5	42	11.2
PC	18.75	75	51.08	50.00	12.5
d´	-1.59	1.18	0.00	0.00	0.62
Test 3 - Recognition test without task awareness, N = 50 (25 male)					
Age	23	81	45	41.50	14.5
PC	18.75	75	49.75	50.00	10.3
d´	-1.56	1.18	-0.01	0.00	

The present test elicited an average percent correct (PC) score of 49.75% (SD = 10.3), with individual scores ranging from 18.75 to 75, and a median of 50. In comparison, List B participants in Test 2 had elicited a mean PC of 51.08% (SD = 12.5), within the same range and with the same median. A Welch's t-test indicated no significant difference between the mean PC scores of both tests ($t(97.73) = .59$, $p = .56$). At first glance, this result seems to reject hypothesis (1), i.e. that the present test would be more difficult than Test 2. However, Test 2 performance was already at chance level, making it improbable to obtain a significantly lower average with the present setup, given the small sample size.

Crucially, the distribution of scores must also be considered when assessing test difficulty. In this connection, the Shapiro-Wilk test indicated normality violations ($W(50) = 0.95$, $p = 0.02$) for the present test, while Test 2 scores had been normally distributed (cf. Chapter 7). Figure 23 shows that the present test's PC distribution is leptokurtic, i.e. it has a higher and narrower peak than a normal distribution. This is unsurprising as the test elicited the same mean score as Test 2 and a markedly lower standard deviation. This means that most of the scores were concentrated near the mean, resulting in less variability overall. At the same

time, the distribution has heavier tails, i.e. more values can be found in the extreme ranges than in a normal distribution. This becomes apparent at the margins of the distribution where the density plot transgresses the blue normal curve.

Figure 23: Denisty plot of Test 3's PC score distribution with superimposed normal curve



Due to the lower SD of this leptokurtic distribution, the cut-off point for super recognition is also lower at 70.35%, compared to 76.08% for List B participants in Test 2. One participant achieved a score of 75% in the present test and therefore matches the criteria for super recognition here. In contrast, the List B top performer in Test 2 had a score of 75% as well but did not match SR criteria. The present test's cut-off for the potential phonagnosic range was 29.15%. Two participants scored below this range, with PC scores of 18.75 and 25%, respectively.

In sum, the greater accumulation of scores around the mean of 49.7% shows that more people in the present test were performing at chance level than in Test 2. This is an increase in participants who performed no better than someone who is randomly guessing. The test was thus more difficult in that it was less likely for participants to distinguish themselves from the mean. In turn, participants who likely based their decisions on some criterion could set themselves apart from the mean more easily, resulting in more extreme performances. The only difference between Tests 2 and 3 was task awareness, which is likely not given for an

earwitness. It might therefore – with all caution – be hypothesised that real-world unfamiliar voice recognition attempts likely elicit leptokurtic performance distributions for the phonetically untrained population. Depending on the prevalence of super recognition and phonagnosia, this might even lead to distribution with additional peaks at the extreme ends (cf. the lower performance margin in Figure 23).

8.3.2 Trial-based analysis

Table 11 and Table 12 show the outcome of the trial-based analysis. For maximal comparability, the same GLMM models were fitted as for Test 2. The tables therefore also present the outcomes of the models run on the List B participant subset of Test 2. In the present test, stimulus list B was used exclusively. In analogy to the analysis of Test 2, separate models were run for target-present (Table 11) and target-absent trials (Table 12).

Table 11: GLMM results (Dependent variable: Correct response to target voices)

	Test 3 - full sample (=Stimulus list B) (N = 50, 25 male)			**Test 2 - Stimulus list B (N = 52, 26 male)**		
Fixed effects	***z (test statistic)***	***p (significance)***		***z (test statistic)***	***p (significance)***	
(Intercept)	0.80	>0.4		-0.28	>0.7	
Age	-0.45	>0.6		0.31	>0.7	
Sex	0.07	>0.9		-1.12	>0.2	
Trial number	1.61	>0.1		-0.02	>0.9	
RT (z transf.)	1.99	<0.05	*	0.58	>0.5	
Confidence	-0.97	>0.3		2.56	<0.05	*
Distinctiveness	1.88	<0.1	.	1.02	>0.3	
F0 difference	-3.10	<0.01	**	-2.67	<0.01	**

Table 12: GLMM results (Dependent variable: Correct response to foil voices)

	Test 3 - full sample (=Stimulus list B) (N = 50, 25 male)			**Test 2 - Stimulus list B (N = 52, 26 male)**		
Fixed effects	***z (test statistic)***	***p (significance)***		***z (test statistic)***	***p (significance)***	
(Intercept)	1.30	>0.1		-1.49	>0.1	
Age	-0.82	>0.4		0.74	>0.4	
Sex	1.56	>0.1		2.27	<0.05	*
Trial number	-2.46	<0.05	*	-0.13	>0.8	
RT (z transf.)	-2.77	<0.01	**	-1.87	<0.1	.
Confidence	-0.36	>0.7		1.04	>0.2	

Like in Test 2 (List B subset), f0 difference between a study and test stimulus had a highly significant impact on correct responses in target-present trials ($z = -3.10$, $p < .01$), which confirms hypothesis 3. In terms of size, the effect was even larger than in Test 2. In contrast to Test 2 (List B subset), RTs had a significant effect on accuracy in target-present trials ($z = 1.99$, $p < .05$). The confidence effect found in Test 2 could not be confirmed in the present test, thus rejecting hypothesis 2.

In target-absent trials RT had a highly significant effect on accuracy ($z = -2.77$, $p < .01$). This means that in contrast to Test 2 (full sample), RT was predictive of accuracy in both target-present and target-absent trials. The significant effect of participant sex on accuracy found for the List B participants in Test 2 was not found here. Finally, a significant trial number effect was found in the present test for target-absent trials ($z = -2.46$, $p < .05$).

Considering the directionality of the found effects, the following simplified summary statements can be made:

1. **In target-present trials**, greater **f0 differences** between the speaker's study and test stimulus significantly increased the likelihood of a miss.
2. **In target-present trials**, faster **reaction times** significantly increased the likelihood of a miss.
3. **In target-absent trials**, faster **reaction times** significantly increased the likelihood of a correct rejection.
4. **In target-absent trials**, stimuli **presented later** within the test phase were correctly rejected more often than stimuli presented early in the test phase.

8.3.3 Impact of task awareness

To better assess the impact of task awareness, two further trials-based GLMMs were run on a merged test cohort, consisting of the List B participants from Test 2 and the Test 3 participants. Since Test 3 was based on Test 2 and made exclusive use of stimulus list B, the groups merged together for this analysis only differed in that Test 3 participants were not aware of the task in study. This new, merged cohort therefore describes a test with two different conditions, an "aware condition" (Test 2 List B participants), which allowed for intentional memorisation of the target voices, and an "unaware condition" (Test 3 participants), which only allowed for an incidental memorisation of the target voices. Separate GLMMs were run on the merged cohort for target-present and target-absent trials. The same independent variables (fixed effects) were fitted as for the equivalent models in Section 8.5.2 with the addition of the binary variable "task

awareness" in both models. This variable took the value of 1 for Test 2 List B participants and the value of 0 for Test 3 participants. Overall, this variable was not found to have a significant impact on the accuracy in either target-present ($z = -.13$, $p = .90$) or target-absent trials ($z = -.15$, $p = .88$).

8.4 Discussion

The present study adds to the very limited literature available on task awareness in voice processing contexts.

A study by Armstrong & McKelvie (1996) examined the effects of task awareness in conjunction with facial cues on voice recognition accuracy. Participants listened to ten unfamiliar voices, with half the participants informed they would later be tested on recognising the voices (intentional condition) and half not informed (incidental condition). During the study phase, each voice was paired with a face. Later, participants completed a two-alternative forced choice recognition test, with half the original face pairings reinstated and half new faces paired with the voices. Recognition accuracy was higher in the intentional versus incidental condition, and when original face pairings were reinstated versus new pairings. In each trial, reported confidence was higher for correct than for incorrect decisions, but more confident participants were not more accurate. The same confidence – accuracy correlation was also found in a similar study by Saslove & Yarmey (1980). Both studies suggest that voice recognition depends on task awareness.

A similar advantage of intentional memorisation could not be diagnosed when the Test 2 List B participant subset was compared to the present test's participants when accuracy means were compared. While Armstrong & McKelvie (1996) found a significant positive correlation between confidence and accuracy in the individual trials, such an effect was not found in the present Test 3 either. There are several possible explanations for this. First, Armstrong & McKelvie elicited a confidence judgement after each trial, while in the present study only one judgement was elicited at the end of the test. The present test might therefore have measured a different type of confidence. Crucially, however, a significant impact of overall confidence could be found in Test 2, where higher overall confidence was indicative of higher hit rates in target-present trials (cf. Table 11). It might be hypothesised that listeners were overall more cautious in their confidence ratings in Test 3 due to the misleading information given about the test's purpose in the study phase. However, participants used the confidence scale in a highly

comparable way in both Test 2 (mean = 2.95, range = 1 - 5, SD = 1.21) and Test 3 (mean = 3.02, range = 1 - 5, SD = 1.28). A t-test did not indicate a statistically significant difference in confidence means ($t(99.09) = .28$, $p = .78$). It is thus plausible that perceived confidence generally differed in this incidental memory task. This may have implications for categorical witness assessment in that participants who report having actively memorised the perpetrator's voice may not necessarily have an advantage at a subsequent recognition test.

Some effects were found in Test 3 that had not been found in Test 2, such as a significant RT effect in both target-present and target-absent trials, as well as a trial number effect in target-absent trials. Both effects cannot be evaluated against existing studies as – to my knowledge – no studies have considered these independent variables in relation to incidental voice memorisation. Surprisingly, shorter reaction times reduced the likelihood of a correct response in a target-present trial and increased the likelihood of a correct response in a target-absent trial. This pattern is different from Test 1, the discrimination test, where faster reaction times increased the likelihood in all types of trials. This could be the result of speakers taking more time in the present test to actively contemplate their decisions. In general, the effect's opposite direction between different trial types makes it difficult to apply this knowledge about RT to e.g. VPs, where the experimenter does not know whether a suspect is the target speaker or not.

The interpretation of the trial number effect is less problematic as this effect is in line with the findings of Test 1. In both cases, the appearance of a foil later in the test increased the likelihood of that foil eliciting a false alarm. This effect has implications for both test battery construction in voice processing tests, as well as for VP construction.

Tests 2 and 3, which only differed in task awareness elicited the same f0 effect in target-present trials, confirming the robustness of f0 as a measure of item difficulty in a test calibrated for normal distribution of speaker f0.

8.5 Limitations

The test shares all of Test 2's limitations as it is identical in design. These limitations were discussed in Section 7.5. Additionally, this test's implications are limited in that all participants were passively listening to the presented voices. An incidental voice memory test by Hammersley & Read (1985), however, found that active conversations with the target speaker significantly improved recognition after incidental memorisation and should therefore be considered for a more ecologically valid design.

8.6 Conclusion

The present test did not find significant differences in mean performance between an incidental voice memory test and an otherwise identical intentional voice memory test. Nonetheless, different independent variables were found to have an impact on participant accuracy in both tests in a trials-based analysis, indicating that latent ability differed. Crucially, an analysis of distribution shapes indicated that recognition after incidental memorisation elicited a leptokurtic distribution, which is more capable of finding extreme performances. Follow-up tests with larger participant samples could fully map out the differences between different task awareness conditions.

9. Conclusion

9.1 Summary

The present study brought together insights from linguistics, psychology, and jurisprudence. It characterised the extent to which the individual lay listener is a variable factor in legal voice identification procedures.

In Chapter 2, a summative framework for vocal identity was outlined, using the individual utterance as the fundamental unit of analysis. The central premise underlying this framework was the assumption that each utterance possesses unique qualities that inherently distinguish it from all other utterances, even those produced by the same speaker. An earwitness therefore faces a difficult task: They must assess whether two utterances are different because they were produced by different speakers or although they were produced by the same speaker. This is further complicated by the fact that these utterances are compared outside of the original contexts in which they were produced. The questioned utterance was produced by the perpetrator during the criminal event, while the comparison utterance was produced by a suspect at a later point. Establishing continuity or discontinuity between these decontextualised "snapshots" is challenging. It was further shown that vocal variability is highly complex, occurring at an organic, cultural, and habitual level.

Chapter 3 discussed the traditional classification of variables in witness cases and showed that the term "estimator variables" refers to a highly heterogeneous group of variables. Witness variables, which can be identified as a subset of estimator variables, might benefit from direct empirical testing. To this end, the theoretical benefits of a screening test for earwitnesses were discussed. Such a test could complement current voice parade procedures, which only produce one data point per witness and are partially based on circular reasoning. A screening test could provide more nuanced information on the individual witness's latent voice recognition ability. This would strengthen the validity of earwitness testimony.

Chapter 4 discussed voice perception through a psychological lens. In this context, neuroimaging studies have shown that familiar and unfamiliar voices are likely processed in distinct neurological pathways. This dissociation has important implications for the categorical assessment of earwitnesses, in that witnesses unfamiliar with the perpetrator may provide a different type of testimony than those familiar with the perpetrator. However, familiar/unfamiliar distinctions are likely more intricate from a functional point of view. A recent model (Lavan & McGettigan, 2023) hypothesises a functional integration of

predominantly feature-based unfamiliar voice recognition and predominantly Gestalt-based familiar voice recognition. Even unfamiliar voice recognition might therefore be informed by familiarity, depending on situational contexts and listener expectations.

Empirical contributions

A total of 250 phonetically untrained native speakers of English took part in three unfamiliar voice processing tests, assessing three distinct abilities: unfamiliar voice discrimination (Test 1, N = 100), unfamiliar voice recognition after intentional memorisation (Test 2, N = 100), and unfamiliar voice recognition after incidental memorisation (Test 3, N = 50). All of these abilities can be characterised as witness variables and were assumed to describe an implicational hierarchy - as complexity increased across tests, average performance was expected to decline. The results largely confirmed this hierarchy, which can be fully delineated by future research.

The results have implications for two different strands of research. First, they can complement earwitness research, as most existing research has focused on the interpretation of performance averages, rather than the individual witness. Second, the study expands on previous psychological voice processing experiments that used simplified stimuli. The use of naturalistic voice recordings in this study improves the ecological validity of the findings.

The specific outcomes of the individual tests can be found in the conclusion sections of the respective chapters (Chapters 6 – 8). However, the most important findings are summarised below:

1. All of the three tests found at least one potential 'super recogniser'.[31] In contrast, very few other tests have been able to find examples of super recognition for the vocal domain. The most common problem of existing tests are ceiling effects, which could be avoided in the present test. This is most likely due to the greater complexity of the employed naturalistic speech stimuli.
2. For the present tests reaction times were considered as an independent variable, which is a novel approach. It was found in this connection that RTs were highly predictive of accuracy in the discrimination test (Test 1), where faster judgements were generally more accurate. Significant, but less stringent RT effects were found in Tests 2 and 3.

[31] Note that the SR found by Test 2 was not part of the main testing cohort, cf. Section 7.3.3.

3. Participant confidence is generally assumed not to correlate with participant accuracy. However, the present findings indicate that a more nuanced approach to confidence ratings may be required. In Test 2 confidence ratings were significantly correlated with hits, but not with false alarms, indicating that listeners possess the ability to assess some aspects of their performance.
4. The tests showed that f0 is a key factor in voice similarity. It was a stable predictor for item difficulty and test difficulty, albeit less so when memory processes were involved. The conducted tests were the first of their kind in which direct connections could be drawn between participant accuracy and raw measurements of a test item's phonetic characteristics (f0).

9.2 Outlook

The present study has shown that individual listeners differ markedly in voice discrimination and voice recognition ability. Nonetheless, listener differences are barely considered in VP construction. Research on witness variables may provide much needed insight into the individual witness's objective reliability.

The purpose of the present test was to establish an empirical baseline for witness variable testing. This baseline character is directly linked to the limitations of this study. While earwitness scenarios are complex, the present tests presented listeners with comparably easy tasks, using high-quality stimuli, quiet listening environments, and short delays. Results do therefore not directly translate to real-world scenarios. Yet, the conducted tests showed a significant level of listener variability at this basic level. Future research may consequently map out the suggested implicational hierarchy by adding further independent or confounding variables to the experimental paradigm, thus increasing the ecological validity for earwitness scenarios. A sensible next step would, for example, be the inclusion of long-term memory processes.

Appendix

Appendix 1: DyViS speakers featured in Test 1

The subsequent three appendices (1.1–1.3) show the makeup of the three stimulus lists (A, B, C) used for Test 1. A detailed account of the stimulus design and stimulus list creation is provided in Sections 6.2.1 and 6.2.2, respectively. In each of the following lists, the rows define unique speaker pairs, which correspond to individual AX trials. To avoid ambiguity, the 96 unique speaker pairs are numbered consecutively across lists. Each list contains 16 same-speaker pairs (greyed out) and 16 different-speaker pairs. The three-digit speaker ID numbers are the original IDs used by the creators of DyViS (Nolan et al., 2009).

Appendix 1.1: Speakers featured in List A

Pair No.	Speaker presented in study condition (A)	Speaker presented in test condition (X)
1.	012 – study recording	012 – test recording
2.	120 – study recording	120 – test recording
3.	034 – study recording	050 – test recording
4.	060 – study recording	032 – test recording
5.	115 – study recording	115 – test recording
6.	073 – study recording	073 – test recording
7.	042 – study recording	020 – test recording
8.	081 – study recording	069 – test recording
9.	030 – study recording	030 – test recording
10.	029 – study recording	029 – test recording
11.	056 – study recording	111 – test recording
12.	064 – study recording	016 – test recording
13.	033 – study recording	033 – test recording
14.	090 – study recording	090 – test recording
15.	112 – study recording	105 – test recording
16.	103 – study recording	067 – test recording
17.	085 – study recording	085 – test recording
18.	087 – study recording	087 – test recording
19.	040 – study recording	072 – test recording
20.	046 – study recording	075 – test recording
21.	084 – study recording	084 – test recording
22.	052 – study recording	052 – test recording
23.	006 – study recording	049 – test recording
24.	044 – study recording	106 – test recording
25.	003 – study recording	003 – test recording
26.	019 – study recording	019 – test recording
27.	076 – study recording	045 – test recording
28.	010 – study recording	100 – test recording
29.	018 – study recording	018 – test recording
30.	022 – study recording	022 – test recording
31.	024 – study recording	074 – test recording
32.	053 – study recording	063 – test recording

Appendix 1.2: Speakers featured in List B

Pair No.	Speaker presented in study condition (A)	Speaker presented in test condition (X)
1.	012 – study recording	120 – test recording
2.	050 – study recording	034 – test recording
3.	032 – study recording	032 – test recording
4.	060 – study recording	060 – test recording
5.	115 – study recording	073 – test recording
6.	020 – study recording	042 – test recording
7.	069 – study recording	069 – test recording
8.	081 – study recording	081 – test recording
9.	030 – study recording	029 – test recording
10.	111 – study recording	056 – test recording
11.	016 – study recording	016 – test recording
12.	064 – study recording	064 – test recording
13.	033 – study recording	090 – test recording
14.	105 – study recording	112 – test recording
15.	067 – study recording	067 – test recording
16.	103 – study recording	103 – test recording
17.	085 – study recording	087 – test recording
18.	072 – study recording	040 – test recording
19.	075 – study recording	075 – test recording
20.	046 – study recording	046 – test recording
21.	084 – study recording	052 – test recording
22.	049 – study recording	006 – test recording
23.	106 – study recording	106 – test recording
24.	044 – study recording	044 – test recording
25.	003 – study recording	019 – test recording
26.	045 – study recording	076 – test recording
27.	100 – study recording	100 – test recording
28.	010 – study recording	010 – test recording
29.	018 – study recording	022 – test recording
30.	074 – study recording	024 – test recording
31.	063 – study recording	063 – test recording
32.	053 – study recording	053 – test recording

Appendix 1.3: Speakers featured in List C

Pair No.	Speaker presented in study condition (A)	Speaker presented in test condition (X)
1.	120 – study recording	012 – test recording
2.	034 – study recording	034 – test recording
3.	050 – study recording	050 – test recording
4.	032 – study recording	060 – test recording
5.	073 – study recording	115 – test recording
6.	042 – study recording	042 – test recording
7.	020 – study recording	020 – test recording
8.	069 – study recording	081 – test recording
9.	029 – study recording	030 – test recording
10.	056 – study recording	056 – test recording
11.	111 – study recording	111 – test recording
12.	016 – study recording	064 – test recording
13.	090 – study recording	033 – test recording
14.	112 – study recording	112 – test recording
15.	105 – study recording	105 – test recording
16.	067 – study recording	103 – test recording
17.	087 – study recording	085 – test recording
18.	040 – study recording	040 – test recording
19.	072 – study recording	072 – test recording
20.	075 – study recording	046 – test recording
21.	052 – study recording	084 – test recording
22.	006 – study recording	006 – test recording
23.	049 – study recording	049 – test recording
24.	106 – study recording	044 – test recording
25.	019 – study recording	003 – test recording
26.	076 – study recording	076 – test recording
27.	045 – study recording	045 – test recording
28.	100 – study recording	010 – test recording
29.	022 – study recording	018 – test recording
30.	024 – study recording	024 – test recording
31.	074 – study recording	074 – test recording
32.	063 – study recording	053 – test recording

Appendix 2: DyViS speakers featured in Tests 2 & 3

The subsequent two appendices (2.1 & 2.2) show the makeup of the two stimulus lists (A & B) used for Tests 2 and 3. NB that Test 2 made use of both List A and B, while Test 3 made use of List B exclusively (cf. Section 8.2.1 for further information). Tests 2 and 3 followed a study-then-test design, which is explained in detail in Section 7.2. Each of the subsequent lists therefore contains two parts: a speaker pool for the study phase (N = 8) and a speaker pool for the test phase (N = 16). All speakers featured in the study phase were also featured in the test phase, albeit with different recordings. The rows corresponding to these "old speakers" are greyed out. The three-digit speaker ID numbers are the original IDs used by the creators of DyViS (Nolan et al., 2009).

Appendix 2.1: Speakers featured in List A

Speaker pool for the study phase
004 – study recording
048 – study recording
054 – study recording
080 – study recording
035 – study recording
093 – study recording
068 – study recording
086 – study recording
Speaker pool for the test phase
004 – test recording
021 – test recording
048 – test recording
121 – test recording
054 – test recording
037 – test recording
080 – test recording
118 – test recording
035 – test recording
023 – test recording
093 – test recording
071 – test recording
068 – test recording
015 – test recording
086 – test recording
108 – test recording

Appendix 2.2: Speakers featured in List B

Speaker pool for the study phase
097 – study recording
078 – study recording
047 – study recording
051 – study recording
059 – study recording
002 – study recording
077 – study recording
066 – study recording
Speaker pool for the test phase
097 – test recording
102 – test recording
078 – test recording
094 – test recording
047 – test recording
001 – test recording
051 – test recording
043 – test recording
059 – test recording
114 – test recording
002 – test recording
013 – test recording
077 – test recording
028 – test recording
066 – test recording
088 – test recording

Appendix 3: DyViS speakers not featured in the study

A total of twenty (out of 100) DyViS speakers were not used for any of the three tests conducted within the framework of this study. The exclusion of these speakers served the purpose of homogenising the speaker pool. The following table lists all omitted speakers and states for each case, why the decision was made not to include the speaker. NB that the reasons in the right column apply to the two 10-second-long passages (study recording, test recording) selected for each of the speakers for the purpose of the conducted tests. They are not necessarily representative of the entire material available for that speaker in the DyViS database. The three-digit speaker ID numbers are the original IDs used by the creators of DyViS (Nolan et al., 2009).

Omitted speaker	**Reason**
058	Accent inconsistencies; features of a non-SSBE accent
079	Outlier in terms of audio quality
096	Outlier in terms of audio quality
113	Outlier in terms of audio quality
011	Noteworthy features that may be perceived as disordered speech
017	Noteworthy features that may be perceived as disordered speech
026	Noteworthy features that may be perceived as disordered speech
027	Noteworthy features that may be perceived as disordered speech
036	Noteworthy features that may be perceived as disordered speech
039	Noteworthy features that may be perceived as disordered speech
062	Noteworthy features that may be perceived as disordered speech
107	Noteworthy features that may be perceived as disordered speech
008	Outlier in terms of within-speaker variability (mean f0-difference between the selected study and test recording > 10 Hz) (Tests 2 & 3)
009	Outlier in terms of within-speaker variability (mean f0-difference between the selected study and test recording > 10 Hz) (Tests 2 & 3)
025	Outlier in terms of within-speaker variability (mean f0-difference between the selected study and test recording > 10 Hz) (Tests 2 & 3)
031	Outlier in terms of within-speaker variability (mean f0-difference between the selected study and test recording > 10 Hz) (Tests 2 & 3)
038	Outlier in terms of within-speaker variability (mean f0-difference between the selected study and test recording > 10 Hz) (Tests 2 & 3)
065	Outlier in terms of within-speaker variability (mean f0-difference between the selected study and test recording > 10 Hz) (Tests 2 & 3)
095	Outlier in terms of within-speaker variability (mean f0-difference between the selected study and test recording > 10 Hz) (Tests 2 & 3)
099	Outlier in terms of within-speaker variability (mean f0-difference between the selected study and test recording > 10 Hz) (Tests 2 & 3)

List of Abbreviations

|	given
3AFC	three-alternative forced choice
AVI	audio-visual integration
BKA	Bundeskriminalamt (German Federal Criminal Police)
BVMT	Bangor Voice Matching Test
CFMT	Cambridge Face Memory Test
CFPT	Cambridge Face Perception Test
CL	Common Law
CrimPR	Criminal Procedure Rules
CVC	consonant-vowel-consonant
d′	D prime
DA	District Attorney
DyViS	Dynamic Variability in Speech (Corpus)
ED	Euclidean distance
f0	Fundamental frequency
F1	first formant
F2	second formant
F3	third formant
fMRI	functional magnetic resonance imaging
FRU	face recognition unit
FSS	forensic speech science
FVRT	Famous Voice Recognition Test
GLMM	Generalised Linear Mixed Model
GVMT	Glasgow Voice Memory Test
Hd	Defence hypothesis
HNR	Harmonics to noise ratio
Hp	Prosecution hypothesis

Hz	Hertz (unit)
IFC	inferior frontal cortex
IRT	Item Response Theory
IVIP	Improving Voice Identification Procedures (research project)
JVLMT	Jena Voice Learning and Memory Test
KS	known sample
LFE	language-familiarity effect
LTF(D)	long-term formant (distribution)
MDS	multidimensional scaling
NB	nota bene
NPV	negative predictive value
PAC	primary auditory cortex
PACE	Police and Criminal Evidence Act (1984)
PIN(s)	person identity node(s)
PPV (a)	positive predictive value
PPV (b)	Person Perception from Voices model (Lavan & McGettigan, 2023)
PVC	primary visual cortex
QS	questioned sample
RP	Received Pronunciation
RT	reaction time
s	second(s)
SDT	Signal Detection Theory
SE	Standard English
SR(s)	super recogniser(s)
SSBE	Standard Southern British English
STC	superior temporal cortex
STS	superior temporal sulcus
TP(s)	temporal pole(s)
TVA(s)	temporal voice area(s)
UK	United Kingdom (of Great Britain and Northern Ireland)

VCV	vowel-consonant-vowel
VOT	voice onset time
VPA	vocal profile analysis
VQ	voice quality
VRU	voice recognition unit
vs	versus

Bibliography

Abberton, E., & Fourcin, A. J. (1978). Intonation and speaker identification. *Language and Speech*, *21*(4), 305–318. https://doi.org/10.1177/002383097802100405

Abercrombie, D. (1967). *Elements of general phonetics*. Edinburgh UP.

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*(1), 97–110. https://doi.org/10.3758/s13428-015-0689-6

Alexander, A., Dessimoz, D., Botti, F., & Drygajlo, A. (2005). Aural and automatic forensic speaker recognition in mismatched conditions. *International Journal of Speech, Language and the Law*, *12*(2), 214–234. https://doi.org/10.1558/sll.2005.12.2.214

Andics, A., McQueen, J. M., Petersson, K. M., Gál, V., Rudas, G., & Vidnyánszky, Z. (2010). Neural mechanisms for voice recognition. *NeuroImage*, *52*(4), 1528–1540. https://doi.org/10.1016/j.neuroimage.2010.05.048

Andrews, S., Jenkins, R., Cursiter, H., & Burton, A. M. (2015). Telling faces together: Learning new faces through exposure to multiple instances. *Quarterly Journal of Experimental Psychology*, *68*(10), 2041–2050. https://doi.org/10.1080/17470218.2014.1003949

ANSI. (1994). Psychoacoustic terminology: Timbre. In American National Standards Institute (Ed.), *American National Standard Psychoacoustical Terminology*. https://books.google.co.uk/books?id=5Fo0GQAACAAJ

Anwyl-Irvine, A., Dalmaijer, E. S., Hodges, N., & Evershed, J. K. (2021). Realistic precision and accuracy of online experiment platforms, web browsers, and devices. *Behavior Research Methods*, *53*(4), 1407–1425. https://doi.org/10.3758/S13428-020-01501-5/FIGURES/17

Armstrong, H. A., & McKelvie, S. J. (1996). Effect of face context on recognition memory for voices. *The Journal of General Psychology*, *123*(3), 259–270. https://doi.org/10.1080/00221309.1996.9921278

Assmann, J. (1992). *Das kulturelle Gedächtnis: Schrift, Erinnerung und politische Identität in frühen Hochkulturen*. C.H. Beck.

Atkinson, N. (2015). *Variable factors affecting voice identification in forensic contexts* [PhD Thesis]. University of York.

Baken, R. J., & Orlikoff, R. F. (2000). *Clinical measurement of speech and voice*. Singular Thomson Learning.

Baldwin, J., & French, J. P. (1990). *Forensic phonetics*. Pinter.

Basu, N., Bali, A. S., Weber, P., Rosas-Aguilar, C., Edmond, G., Martire, K. A., & Morrison, G. S. (2022). Speaker identification in courtroom contexts – Part I: Individual listeners compared to forensic voice comparison based on automatic-speaker-recognition technology. *Forensic Science International*, *341*, 111499. https://doi.org/10.1016/j.forsciint.2022.111499

Bate, S., Frowd, C., Bennetts, R., Hasshim, N., Murray, E., Bobak, A. K., Wills, H., & Richards, S. (2018). Applied screening tests for the detection of superior face recognition. *Cognitive Research: Principles and Implications*, *3*(1). https://doi.org/10.1186/s41235-018-0116-5

Bate, S., Portch, E., & Mestry, N. (2021). When two fields collide: Identifying "super-recognisers" for neuropsychological and forensic face recognition research. *Quarterly Journal of Experimental Psychology*, *74*(12), 2154–2164. https://doi.org/10.1177/17470218211027695

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Baumann, O., & Belin, P. (2010). Perceptual scaling of voice identity: Common dimensions for different vowels and speakers. *Psychological Research*, *74*(1), 110–120. https://doi.org/10.1007/S00426-008-0185-Z

Baumann, S., & Grice, M. (2006). The intonation of accessibility. *Journal of Pragmatics*, *38*(10), 1636–1657. https://doi.org/10.1016/j.pragma.2005.03.017

Beaudry, J. L., Bullard, C. L., & Dolin, J. R. (2014). Estimator variables and eyewitness identification. In G. Bruinsma & D. Weisburd (Eds.), *Encyclopedia of criminology and criminal justice* (pp. 1384–1394). Springer New York. https://doi.org/10.1007/978-1-4614-5690-2_668

Beck, J. M. (2010). Organic variation of the vocal apparatus. In W. J. Hardcastle, J. Laver & F. E. Gibbon (Eds.), *The handbook of phonetic sciences* (2nd ed., pp. 153–201). Wiley. https://doi.org/10.1002/9781444317251.ch5

Beinhoff, B. (2013). *Perceiving identity through accent*. Peter Lang.

Belin, P., Bestelmeyer, P. E. G., Latinus, M., & Watson, R. (2011). Understanding voice perception. *British Journal of Psychology*, *102*(4), 711–725. https://doi.org/https://doi.org/10.1111/j.2044-8295.2011.02041.x

Belin, P., Boehme, B., & McAleer, P. (2017). The sound of trustworthiness: Acoustic-based modulation of perceived voice personality. *PLOS ONE*, *14*(1), https://doi.org/10.1371/journal.pone.0211282

Belin, P., Fecteau, S., & Bédard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences 8*(3), 129–135. https://doi.org/10.1016/j.tics.2004.01.008

Belin, P., & Zatorre, R. J. (2003). Adaptation to speaker's voice in right anterior temporal lobe. *NeuroReport*, *14*(16), 2105–2109. https://doi.org/10.1097/00001756-200311140-00019

Belin, P., Zatorre, R. J., & Ahad, P. (2002). Human temporal-lobe response to vocal sounds. *Cognitive Brain Research*, *13*(1), 17–26. https://doi.org/10.1016/S0926-6410(01)00084-2

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature 403*(6767), 309–312. https://doi.org/10.1038/35002078

Bertillon, A. (1893). *Instructions Signalétiques*. Imprimerie Administrative.

Blatchford, H., & Foulkes, P. (2006). Idenfication of voices in shouting. *International Journal of Speech Language and the Law*, *13*(2). https://doi.org/10.1558/ijsll.2006.13.2.241

Bobak, A. K., Pampoulov, P., & Bate, S. (2016). Detecting superior face recognition skills in a large sample of young British adults. *Frontiers in Psychology*, *7*(SEP). https://doi.org/10.3389/fpsyg.2016.01378

Boersma, P., & Weenink, D. (2022). Praat: Doing phonetics by computer (6.1.15) [Computer software]. Retrieved from http://cogent.psyc.bbk.ac.uk/

Bowie, D. (2011). Aging and sociolinguistic variation. In A. Duszak & U. Okulska (Eds.), *Language, culture and the dynamics of age* (pp. 29–52). De Gruyter Mouton. https://doi.org/doi:10.1515/9783110238112.2.29

Bradshaw, L., Chodroff, E., Jäger, L., & Dellwo, V. (2022). Fundamental frequency variability over time in telephone interactions. *Proceedings of Interspeech 2022*, 101–105. https://doi.org/10.21437/Interspeech.2022-10669

Braun, A. (1995). Fundamental frequency: How speaker-specific is it? In A. Braun & J.-P. Köster (Eds.), *Studies in forensic phonetics* (pp. 9–23). Wissenschaftlicher Verlag Trier.

Braun, A. (1996). Age estimation by different listener groups. *International Journal of Speech Language and the Law*, *3*(1), 65–73. https://doi.org/10.1558/ijsll.v3i1.65

Braun, A. (2006). Stimmverstellung und Stimmenimitation in der forensischen Sprechererkennung. In T. Kopfermann (Ed.), *Das Phänomen Stimme: Imitation und Identität. 5. Stuttgarter Stimmtage 2004. Hellmut K. Geissner zum 80. Geburtstag* (pp. 177–181). Röhrig.

Braun, A., Llamas, C., Watt, D., French, P., & Robertson, D. (2018). Sub-regional 'other-accent' effects on lay listeners' speaker identification abilities. *International Journal of Speech Language and the Law*, *25*(2), 231–255. https://doi.org/10.1558/ijsll.37340

Bricker, P. D., & Pruzansky, S. (1966). Effects of stimulus content and duration on talker identification. *The Journal of the Acoustical Society of America*, *40*(6), 1441–1449. https://doi.org/10.1121/1.1910246

Bridges, D., Pitiot, A., MacAskill, M. R., & Peirce, J. W. (2020). The timing mega-study: comparing a range of experiment generators, both lab-based and online. *PeerJ*, *8*. https://doi.org/10.7717/PEERJ.9414

Broeders, A. P. A., & Rietveld, A. C. M. (1995). Speaker identification by earwitness. In A. Braun & J.-P. Koester (Eds.), *Studies in forensic phonetics* (pp. 24–40). Wissenschaftlicher Verlag Trier.

Broeders, A. P. A., & Van Amelsvoort, A. G. (1999). Lineup construction for forensic earwitness identification: A practical approach. *Proceedings of the 14th International Congress of Phonetic Sciences*, 1373–1376. https://www.researchgate.net/publication/262824196

Broeders, A. P. A., & van Amelsvoort, A. G. (2001). A practical approach to forensic earwitness identification: Constructing a voice line-up. *Problems of Forensic Science, 47*, 237–245.

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, *77*(3), 305–327. https://doi.org/10.1111/j.2044-8295.1986.tb02199.x

Bruckert, L., Bestelmeyer, P., Latinus, M., Rouger, J., Charest, I., Rousselet, G. A., Kawahara, H., & Belin, P. (2010). Vocal attractiveness increases by averaging. *Current Biology*, *20*(2), 116–120. https://doi.org/10.1016/j.cub.2009.11.034

Brumm, H., & Zollinger, S. A. (2011). The evolution of the Lombard effect: 100 years of psychoacoustic research. *Behaviour*, *148*(11–13), 1173–1198. https://doi.org/10.1163/000579511X605759

Bull, R., & Clifford, B. R. (1999). Earwitness testimony. *Medicine, Science and the Law*, *39*(2), 120–127. https://doi.org/10.1177/002580249903900206

Bull, R., Rathborn, H., & Clifford, B. R. (1983). The voice-recognition accuracy of blind listeners. *Perception*, *12*(2), 223–226. https://doi.org/10.1068/p120223

Burton, A. M., Bruce, V., & Johnston, R. A. (1990). Understanding face recognition with an interactive activation model. *British Journal of Psychology*, *81*(3), 361–380. https://doi.org/10.1111/j.2044-8295.1990.tb02367.x

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow Face Matching Test. *Behavior Research Methods, 42*(1), 286–291. https://doi.org/10.3758/BRM.42.1.286

Byrne, C., & Foulkes, P. (2004). The "Mobile Phone Effect" on vowel formants. *International Journal of Speech Language and the Law*, *11*(1), 83–102. https://doi.org/10.1558/ijsll.v11i1.83

Calvert, G. A., Brammer, M. J., & Iversen, S. D. (1998). Crossmodal identification. *Trends in Cognitive Sciences*, *2*(7), 247–253. https://doi.org/10.1016/S1364-6613(98)01189-9

Cambier-Langeveld, T. (2007). Current methods in forensic speaker identification: Results of a collaborative exercise. *International Journal of Speech Language and the Law*, *14*(2). https://doi.org/10.1558/ijsll.v14i2.223

Cambier-Langeveld, T., van Rossum, M., & Vermeulen, J. (2014). Whose voice is that? Challenges in forensic phonetics. In J. Caspers, Y. Chen, W. Heeren, J. Pacilly, N. O. Schiller & E. van Zanten (Eds.), *Above and beyond the segments* (pp. 14–27). John Benjamins Publishing Company. https://doi.org/10.1075/z.189.02cam

Carterette, E. C., & Barnebey, A. (1975). Recognition memory for voices. In A. Cohen & S. G. Nooteboom (Eds.), *Structure and process in speech perception* (pp. 246–265). Springer Berlin Heidelberg.

Chartrand, J.-P., & Belin, P. (2006). Superior voice timbre processing in musicians. *Neuroscience Letters*, *405*(3), 164–167. https://doi.org/10.1016/j.neulet.2006.06.053

Clark, J., & Foulkes, P. (2007). Identification of voices in electronically disguised speech. *International Journal of Speech Language and the Law*, *14*(2). https://doi.org/10.1558/ijsll.v14i2.195

Clarke, E. F. (2005). *Ways of listening: An ecological approach to the perception of musical meaning*. Oxford University Press.

Clifford, B. R. (1980). Voice identification by human listeners: On earwitness reliability. *Law and Human Behavior, 4(*4). https://www.jstor.org/stable/1393857

Collins, E., Robinson, A. K., & Behrmann, M. (2018). Distinct neural processes for the perception of familiar versus unfamiliar faces along the visual hierarchy revealed by EEG. *NeuroImage*, *181*, 120–131. https://doi.org/10.1016/j.neuroimage.2018.06.080

Compton, A. J. (1963). Effects of filtering and vocal duration upon the identification of speakers, aurally. *Journal of the Acoustical Society of America*, *35*, 1748–1752.

Conway, M. A., & Howe, M. L. (2022). Memory construction: a brief and selective history. *Memory*, *30*(1), 2–4. https://doi.org/10.1080/09658211.2021.1964795

Cook, S., & Wilding, J. (1997). Earwitness testimony: Never mind the variety, hear the length. *Applied Cognitive Psychology*, *11*(2), 95–111. https://doi.org/https://doi.org/10.1002/(SICI)1099-0720(199704)11:2<95::AID-ACP429>3.0.CO;2-O

Cook, S., & Wilding, J. (2001). Earwitness testimony: Effects of exposure and attention on the Face Overshadowing Effect. *British Journal of Psychology*, *92*(4), 617–629. https://doi.org/10.1348/000712601162374

Cox, R. V, Neto, S. F. D. C., Lamblin, C., & Sherif, M. H. (2009). ITU-T coders for wideband, superwideband, and fullband speech communication. *IEEE Communications Magazine*, *47*(10), 106–109. https://doi.org/10.1109/MCOM.2009.5273816

(The) Criminal Procedure Rules, (2020). https://www.legislation.gov.uk/uksi/2020/759/contents/made

Crown Office and Procurator Fiscal Service. (2007). *Lord Advocate's Guidelines on the Conduct of Visual Identification Procedures*. https://www.copfs.gov.uk/publications/lord-advocate-s-guidelines-visual-identification-procedures/html/

Crystal, D. (2008). *A dictionary of linguistics and phonetics*. Blackwell Publishing Ltd. https://doi.org/10.1002/9781444302776

Cutler, B., & Wells, G. (2009). Expert testimony regarding eyewitness identification. In J. Skeem, K. Douglas, & S. Lilenfeld (Eds.), *Psychological science in the courtroom: Consensus and controversy* (pp. 100–123). Guilford.

Dallaston, K., & Docherty, G. (2019). Estimating the prevalence of creaky voice: A fundamental frequency-based approach. *Proceedings of the 19th International Congress of Phonetic Sciences*, 532–536.

Damborenea Tajada, J., Fernández Liesa, R., Llorente Arenas, E., Mj, N. G., Marín Garrido, C., Rueda Gormedino, P., & Ortiz-García, A. (1999). The effect of tobacco consumption on acoustic voice analysis. *Acta Otorrinolaringológica Española*, *50*, 448.

de Jong, G., Nolan, F., McDougall, K., & Hudson, T. (2015). Voice lineups: A practical guide. *Proceedings of the 18th International Congress of Phonetic Sciences*, n.p.

DeCasper, A. J., & Fifer, W. P. (1980). Of human bonding: Newborns prefer their mothers' voices. *Science*, *208*(4448), 1174–1176. https://doi.org/10.1126/science.7375928

Deffenbacher, K. (1980). Eyewitness accuracy and confidence: Can we infer anything about their relationship? *Law and Human Behavior*, *4*(4), 243–260. https://doi.org/10.1007/BF01040617

Deffenbacher, K. (1983). The influence of arousal on reliability of testimony. In B. R. Clifford & S. M. A. Lloyd-Bostock (Eds.), *Evaluating witness evidence* (pp. 235–251). Wiley.

Deffenbacher, K. A., Bornstein, B. H., Penrod, S. D., & McGorty, E. K. (2004). A meta-analytic review of the effects of high stress on eyewitness memory. *Law and Human Behavior*, *28*(6), 687–706. https://doi.org/10.1007/s10979-004-0565-x

Devlin Committee Report. (1976). *Report to the Secretary of State for the Home Department of the Departmental Committee in Evidence of Identification in Criminal Cases*, Cmnd 338 134/135, 42.

Duchaine, B., Germine, L., & Nakayama, K. (2007). Family resemblance: Ten family members with prosopagnosia and within-class object agnosia. *Cognitive Neuropsychology*, *24*(4), 419–430. https://doi.org/10.1080/02643290701380491

Duchaine, B., & Nakayama, K. (2006). The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia*, *44*(4), 576–585. https://doi.org/10.1016/J.NEUROPSYCHOLOGIA.2005.07.001

Duchaine, B., Yovel, G., Butterworth, E. J., & Nakayama, K. (2006). Prosopagnosia as an impairment to face-specific mechanisms: Elimination of the alternative hypotheses in a developmental case. *Cognitive Neuropsychology*, *23*(5), 714–747. https://doi.org/10.1080/02643290500441296

Dunn, J. D., Summersby, S., Towler, A., Davis, J. P., & White, D. (2020). UNSW Face Test: A screening tool for super-recognizers. *PLOS ONE*, *15*(11), e0241747. https://doi.org/10.1371/journal.pone.0241747

Edmond, G., Martire, K., & San Roque, M. (2011). "Mere guesswork": Cross-lingual voice comparisons and the jury. *The Sydney Law Review*, *33*(3), 395–425. https://search.informit.org/doi/10.3316/informit.532339284270660

Elaad, E., Segev, S., & Tobin, Y. (1998). Long-term working memory in voice identification. *Psychology, Crime & Law*, *4*(2), 73–88. https://doi.org/10.1080/10683169808401750

Ellis, H. D., Jones, D. M., & Mosdell, N. (1997). Intra- and inter-modal repetition priming of familiar faces and voices. *British Journal of Psychology*, *88*(1), 143–156. https://doi.org/10.1111/j.2044-8295.1997.tb02625.x

Emson, R. (2004). *Evidence* (2nd ed.). Palgrave Macmillan.

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, *102*(2), 211–245. https://doi.org/10.1037/0033-295X.102.2.211

Eriksson, A., & Wretling, P. (1997). How flexible is the human voice? A case study of mimicry. *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech 1997)*, 1043–1046. https://doi.org/10.21437/Eurospeech.1997-363

Evans, N., Kinnunen, T., Yamagishi, J., Wu, Z., Alegre, F., & De Leon, P. (2014). Speaker recognition anti-spoofing. In S. Marcel, M. S. Nixon, S. Z. Li(Eds.), *Handbook of biometric anti-spoofing: Trusted biometrics under spoofing attacks* (p. 125-146). Springer.

Evett, I. W. (1996). Expert evidence and forensic misconceptions of the nature of exact science. *Science & Justice*, *36*(2), 118–122. https://doi.org/10.1016/S1355-0306(96)72576-5

Fant, G. (1960). *Acoustic theory of speech production*. Mouton & Co.

Fecher, N. (2014). *Effects of forensically-relevant facial concealment on acoustic and perceptual properties of consonants* [PhD Thesis]. University of York.

Fecher, N., & Watt, D. (2013). Effects of forensically-realistic facial concealment on auditory-visual consonant recognition in quiet and noise conditions. *Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP 2013)*, 81-86.

Fenn, K. M., Shintel, H., Atkins, A. S., Skipper, J. I., Bond, V. C., & Nusbaum, H. C. (2011). When less is heard than meets the ear: Change deafness in a telephone conversation. *Quarterly Journal of Experimental Psychology*, *64*(7), 1442–1456. https://doi.org/10.1080/17470218.2011.570353

Fisher, J. (1987). *The Lindbergh case: A story of two lives*. Rutgers University Press.

Fitch, W. T. (2000). The evolution of speech: A comparative review. *Trends in Cognitive Sciences*, *4*(7), 258–267. https://doi.org/10.1016/S1364-6613(00)01494-7

Fitzgerald, R. J., Oriet, C., & Price, H. L. (2015). Suspect filler similarity in eyewitness lineups: A literature review and a novel methodology. *Law and Human Behavior*, *39*(1), 62–74. https://doi.org/10.1037/lhb0000095

Fleming, D., Giordano, B. L., Caldara, R., & Belin, P. (2014). A language-familiarity effect for speaker discrimination without comprehension. *Proceedings of the National Academy of Sciences*, *111*(38), 13795-13798. https://doi.org/10.1073/pnas.1401383111

Foulkes, P. (2020). Phonological variation: A global perspective [revised]. In B. Aarts & A. M. S. McMahon (Eds.), *Handbook of English Linguistics* (2nd ed., pp. 407–440). Wiley.

Foulkes, P., & Barron, A. (2000). Telephone speaker recognition amongst members of a close social network. *Forensic Linguistics*, *7*(2), 198.

Foulkes, P., & French, P. (2012). Forensic speaker comparison: A linguistic-acoustic perspective. In L. M. Solan & P. M. Tiersma (Eds.), *The Oxford Handbook of Language and Law* (pp. 558–572). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199572120.013.0041

Foulkes, P., French, P., Wilson, K. (2019). LADO as forensic speaker profiling. In: P.L. Patrick, M.S. Schmid, K. Zwaan (Eds.), *Language Analysis for the Determination of Origin*. Language Policy (16). Springer. https://doi.org/10.1007/978-3-319-79003-9_6

Gainotti, G. (2013). Laterality effects in normal subjects' recognition of familiar faces, voices and names. Perceptual and representational components. *Neuropsychologia*, *51*(7), 1151–1160. https://doi.org/10.1016/J.NEUROPSYCHOLOGIA.2013.03.009

Gerlach, L., McDougall, K., Kelly, F., & Alexander, A. (2023). Automatic assessment of voice similarity within and across speaker groups with different accents. In R. Skarnitzl & J. Volín (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences,* 3785–3789. Guarant International.

Gerlach, L., McDougall, K., Kelly, F., Alexander, A., & Nolan, F. (2020). Exploring the relationship between voice similarity estimates by listeners and by an automatic speaker recognition system incorporating phonetic features. *Speech Communication*, *124*, 85–95. https://doi.org/10.1016/j.specom.2020.08.003

Gfroerer, S. (1994). *Häufigkeit und Art forensischer Stimmverstellungen (Manuscript)*. Bundeskriminalamt.

Gold, E., & French, P. (2011). International practices in forensic speaker comparison. *International Journal of Speech Language and the Law*, *18*(2). https://doi.org/10.1558/ijsll.v18i2.293

Goldman-Eisler, F. (1968). *Psycholinguistics: Experiments in spontaneous speech*. Academic Press.

González Hautamäki, R., Kinnunen, T., Hautamäki, V., & Laukkanen, A.-M. (2015). Automatic versus human speaker verification: The case of voice mimicry. *Speech Communication*, *72*, 13–31. https://doi.org/10.1016/j.specom.2015.05.002

Guberman, S. (2017). Gestalt theory rearranged: Back to Wertheimer. *Frontiers in Psychology*, *8*(OCT), 1782. https://doi.org/10.3389/FPSYG.2017.01782/

Hammarström, G. (1980). Idiolekt. In P. Althaus, H. Henne, & H. E. Wiegand (Eds.), *Lexikon der Germanischen Linguistik* (2nd ed., pp. 428–430). Niemeyer.

Hammersley, R., & Read, J. D. (1985). The effect of participation in a conversation on recognition and identification of the speakers' voices. *Law and Human Behavior*, *9*(1), 71–81.

Hammersley, R., & Read, J. D. (1996). Voice identification by humans and computers. In S. L. Sporer, R. S. Malpass, & G. Koehnken (Eds.), *Psychological issues in eyewitness identification* (pp. 117–152). Lawrence Erlbaum Associates, Inc.

Handel, S., & Erickson, M. L. (2001). A rule of thumb: The bandwidth for timbre invariance is one octave. *Music Perception*, *19*(1), 121–126. https://doi.org/10.1525/mp.2001.19.1.121

Hanley, J. R., Smith, S. T., & Hadfield, J. (1998). I recognise you but I can't place you: An investigation of familiar-only experiences during tests of voice and face recognition. *The Quarterly Journal of Experimental Psychology Section A*, *51*(1), 179–195. https://doi.org/10.1080/713755751

Hanley, J., & Turner, J. M. (2000). Why are familiar-only experiences more frequent for voices than for faces? *The Quarterly Journal of Experimental Psychology Section A*, *53*(4), 1105–1116. https://doi.org/10.1080/713755942

Harrison, P. (2013). *Making accurate formant measurements: An empirical investigation of the influence of the measurement tool, analysis settings and speaker on formant measurements* [PhD Thesis]. University of York.

Hecker, M. H. L. (1971). *Speaker recognition: An interpretive survey of the literature: Vol. ASHA Monographs 16*. 1-103.

Henton, C., & Bladon, A. (1988). Creak as a sociophonetic marker. In L. Hyman & C. N. Li (Eds.), *Language, Speech, and Mind* (pp. 3–29). Routledge.

Henze, R. (1953). Experimentelle Untersuchungen zur Phänomenologie der Sprechgeschwindigkeit. *Zeitschrift Für Experimentelle Und Angewandte Psychologie*, *1*, 214–243.

Hill, R. A., & Dunbar, R. I. M. (2003). Social network size in humans. *Human Nature*, *14*(1), 53–72. https://doi.org/10.1007/s12110-003-1016-y

Hirson, A., & Duckworth, M. (1993). Glottal fry and voice disguise: A case study in forensic phonetics. *Journal of Biomedical Engineering*, *15*(3), 193–200. https://doi.org/10.1016/0141-5425(93)90115-F

Hirson, A., French, J. P., & Howard, D. (1994). Speech fundamental frequency over the telephone and face-to-face: some implications for forensic phonetics. In J. Windsor Lewis (Ed.), *Studies in general and English phonetics* (pp. 230–240). Routledge.

Hogg, M. A., & Vaughan, G. M. (2002). *Social psychology* (3rd ed.). Prentice Hall.

Hollien, H. (1996). Consideration of guidelines for earwitness lineups. *International Journal of Speech Language and the Law*, *3*(1), 14–23. https://doi.org/10.1558/ijsll.v3i1.14

Hollien, H. (2002). *Forensic voice identification*. Academic Press.

Hollien, H., Majewski, W., & Doherty, E. T. (1982). Perceptual identification of voices under normal, stress and disguise speaking conditions. *Journal of Phonetics*, *10*(2), 139–148. https://doi.org/10.1016/S0095-4470(19)30953-2

Home Office. (2003). *Home Office circular 057/2003: Advice on the use of voice identification parades.* https://webarchive.nationalarchives.gov.uk/ukgwa/20130125153221/http://www.homeoffice.gov.uk/about-us/corporate-publications-strategy/home-office-circulars/circulars-2003/057-2003/

Home Office. (2017). *Police and Criminal Evidence Act 1984 (PACE): Code D revised - Code of Practice for the identification of persons by police officers*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/903812/pace-code-d-2017.pdf

Honorof, D. N., & Whalen, D. H. (2005). Perception of pitch location within a speaker's f0 range. *The Journal of the Acoustical Society of America*, *117*(4), 2193–2200. https://doi.org/10.1121/1.1841751

Hughes, A., Trudgill, P., & Watt, D. (2012). *English accents & dialects* (5th ed.). Hodder Education.

Hughes, V., Harrison, P., Foulkes, P., French, J. P., & Gully, A. (2019). Effects of formant analysis settings and channel mismatch on semi-automatic forensic voice comparison. *Proceedings of the 19th International Congress of Phonetic Sciences*, 3080–3084.

Hughes, V., Wood, S., & Foulkes, P. (2016). Strength of forensic voice comparison evidence from the acoustics of filled pauses. *International Journal of Speech, Language and the Law*, *23*(1), 99–132. https://doi.org/10.1558/ijsll.v23i1.29874

Humble, D., Schweinberger, S. R., Mayer, A., Jesgarzewsky, T. L., Dobel, C., & Zäske, R. (2022). The Jena Voice Learning and Memory Test (JVLMT): A standardized tool for assessing the ability to learn and recognize voices. *Behavior Research Methods, 55*, 1352–1371. https://doi.org/10.3758/s13428-022-01818-3

Jarque, C. M., & Bera, A. K. (1987). A test for normality of observations and regression residuals. *International Statistical Review / Revue Internationale de Statistique*, *55*(2), 163. https://doi.org/10.2307/1403192

Jenkins, J., & Setter, J. (2005). Teaching English pronunciation: A state of the art review. *Language Teaching*, *38*(1).

Jenkins, R. E., Tsermentseli, S., Monks, C. P., Robertson, D. J., Stevenage, S. V., Symons, A. E., & Davis, J. P. (2021). Are super-face-recognisers also super-voice-recognisers? Evidence from cross-modal identification tasks. *Applied Cognitive Psychology*, *35*(3), 590–605. https://doi.org/10.1002/acp.3813

Jenkins, R., White, D., Van Montfort, X., & Mike Burton, A. (2011). Variability in photos of the same face. *Cognition*, *121*(3), 313–323. https://doi.org/10.1016/j.cognition.2011.08.001

Jessen, M. (2007). Forensic reference data on articulation rate in German. *Science & Justice*, *47*(2), 50–67. https://doi.org/10.1016/j.scijus.2007.03.003

Jessen, M. (2012). *Phonetische und linguistische Prinzipien des forensischen Stimmenvergleichs*. Lincom Europa.

Jessen, M., Köster, O., & Gfroerer, S. (2005). Influence of vocal effort on average and variability of fundamental frequency. *International Journal of Speech, Language and the Law, 12*(2), 174–213. https://doi.org/10.1558/sll.2005.12.2.174

Johnson, E. K., Westrek, E., Nazzi, T., & Cutler, A. (2011). Infant ability to tell voices apart rests on language experience. *Developmental Science*, *14*(5), 1002–1011. https://doi.org/https://doi.org/10.1111/j.1467-7687.2011.01052.x

Johnson, J., McGettigan, C., & Lavan, N. (2020). Comparing unfamiliar voice and face identity perception using identity sorting tasks. *Quarterly Journal of Experimental Psychology*, *73*(10), 1537–1545. https://doi.org/10.1177/1747021820938659

Judicial College. (2023). *The Crown Court Compendium - Part I: Jury and trial management and summing up*. https://www.judiciary.uk/wp-content/uploads/2023/06/Crown-Court-Compendium-Part-I.pdf

Kennedy, L. (1985). *The airman and the carpenter: The Lindbergh kidnapping and the framing of Richard Hauptmann*. Viking.

Kerstholt, J. H., Jansen, N. J. M., Van Amelsvoort, A. G., & Broeders, A. P. A. (2004). Earwitnesses: Effects of speech duration, retention interval and acoustic environment. *Applied Cognitive Psychology*, *18*(3), 327–336. https://doi.org/https://doi.org/10.1002/acp.974

Kerstholt, J. H., Jansen, N. J. M., Van Amelsvoort, A. G., & Broeders, A. P. A. (2006). Earwitnesses: Effects of accent, retention and telephone. *Applied Cognitive Psychology*, *20*(2), 187–197. https://doi.org/10.1002/acp.1175

Kerswill, P. (2003). Dialect levelling and geographical diffusion in British English. In D. Britain, & J. Chesire (Eds.) *Social dialectology: In honour of Peter Trudgill* (pp. 223-243). Benjamins.

Kim, H.-H. (2008). *Accent disguise: Implications for forensic casework* [Unpublished MSc Dissertation]. University of York.

Kirchhübel, C., Howard, D. M., & Stedmon, A. W. (2011). Acoustic correlates of speech when under stress: Research, methods and future directions. *International Journal of Speech Language and the Law*, *18*(1). https://doi.org/10.1558/ijsll.v18i1.75

Kirk, P. L. (1963). The ontogeny of criminalistics. *The Journal of Criminal Law, Criminology and Police Science.*, *54*(2), 235–238.

Klingholz, F., Penning, R., & Liebhardt, E. (1988). Recognition of low-level alcohol intoxication from speech signal. *The Journal of the Acoustical Society of America*, *84*(3), 929–935. https://doi.org/10.1121/1.396661

Kraayeveld, J. (1997). *Idiosyncrasy in prosody: Speaker and speaker group identification in Dutch using melodic and temporal information* [PhD thesis]. Catholic University of Nijmegen.

Kreiman, J. (1997). Listening to voices: Theory and practice in voice perception research. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 85–108). Academic Press.

Kreiman, J., & Sidtis, D. (2011). *Foundations of voice studies*. Wiley. https://doi.org/10.1002/9781444395068

Kriegstein, K. von, Kleinschmidt, A., Sterzer, P., & Giraud, A.-L. (2005). Interaction of face and voice areas during speaker recognition. *Journal of Cognitive Neuroscience*, *17*(3), 367–376. https://doi.org/10.1162/0898929053279577

Künzel, H.J., Braun, A., Eysholdt, U. (1992). Einfluß von Alkohol auf Sprache und Stimme. *Rechtsmedizin*, *7*(2), 48–48. https://doi.org/10.1007/BF03042347

Künzel, H. J. (1994). Current approaches to forensic speaker recognition. *Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, 135–142.

Künzel, H. J. (1995). Field procedures in forensic speaker recognition. In J. Lewi (Ed.), *Festschrift for J. D. O'Connor* (pp. 68–84). Routledge.

Künzel, H. J. (1997). Some general phonetic and forensic aspects of speaking tempo. *Forensic Linguistics*, *4*, 48–83.

Künzel, H. J. (2000). Effects of voice disguise on speaking fundamental frequency. *Forensic Linguistics*, *7*(2), 149–179. 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. *International Journal of Speech Language and the Law*, *8*(1). https://doi.org/10.1558/ijsll.v8i1.80

Ladefoged, P., & Johnson, K. (2014). *A course in phonetics* (7th ed.). Cengage Learning.

Ladefoged, P., & Ladefoged, J. (1980). The ability of listeners to identify voices. *UCLA Working Papers in Phonetics*, *49*, 43–51.

Latinus, M., Crabbe, F., & Belin, P. (2011). Learning-induced changes in the cerebral processing of voice identity. *Cerebral Cortex*, *21*(12), 2820–2828. https://doi.org/10.1093/cercor/bhr077

Latinus, M., McAleer, P., Bestelmeyer, P. E. G., & Belin, P. (2013). Norm-based coding of voice identity in human auditory cortex. *Current Biology*, *23*(12), 1075–1080. https://doi.org/10.1016/J.CUB.2013.04.055

Lavan, N. (2023). How do we describe other people from voices and faces? *Cognition*, *230*, 105253. https://doi.org/10.1016/j.cognition.2022.105253

Lavan, N., Burston, L. F. K., & Garrido, L. (2019). How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices. *British Journal of Psychology*, *110*(3), 576–593. https://doi.org/10.1111/bjop.12348

Lavan, N., & McGettigan, C. (2023). A model for person perception from familiar and unfamiliar voices. *Communications Psychology*, *1*(1), 1. https://doi.org/10.1038/s44271-023-00001-4

Lavan, N., Mileva, M., & McGettigan, C. (2021). How does familiarity with a voice affect trait judgements? *British Journal of Psychology*, *112*(1), 282–300. https://doi.org/10.1111/bjop.12454

Lavan, N., Smith, H. M. J., & McGettigan, C. (2022). Unimodal and cross-modal identity judgements using an audio-visual sorting task: Evidence for independent processing of faces and voices. *Memory & Cognition*, *50*(1), 216–231. https://doi.org/10.3758/s13421-021-01198-7

Laver, J. (1994). *Principles of phonetics*. Cambridge University Press.

Laver, J. (2009). *The phonetic description of voice quality*. Cambridge University Press. https://books.google.co.uk/books?id=AqmAPgAACAAJ

Lavner, Y., Rosenhouse, J., & Gath, I. (2001). The prototype model in speaker identification by human listeners. *International Journal of Speech Technology*, *4*(1), 63–74. https://doi.org/10.1023/A:1009656816383

Law, L. N. C., & Zentner, M. (2012). Assessing musical abilities objectively: Construction and validation of the profile of music perception skills. *PLOS ONE*, *7*(12). https://doi.org/10.1371/JOURNAL.PONE.0052508

Lawrence, S., Nolan, F., & McDougall, K. (2009). Acoustic and perceptual effects of telephone transmission on vowel quality. *International Journal of Speech, Language and the Law*, *15*(2). https://doi.org/10.1558/ijsll.v15i2.161

Leemann, A. (2016). Analyzing geospatial variation in articulation rate using crowdsourced speech data. *Journal of Linguistic Geography*, *4*(2), 76–96. https://doi.org/10.1017/jlg.2016.11

Legge, G. E., Grosmann, C., & Pieper, C. M. (1984). Learning unfamiliar voices. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 298–303. https://doi.org/10.1037/0278-7393.10.2.298

Lindblom, B. (1963). Spectrographic study of vowel reduction. *The Journal of the Acoustical Society of America*, *35*(11), 1773–1781. https://doi.org/10.1121/1.1918816

Loakes, D., & McDougall, K. (2010). Individual variation in the frication of voiceless plosives in Australian English: A study of twins' speech. *Australian Journal of Linguistics*, *30*(2), 155–181. https://doi.org/10.1080/07268601003678601

Lüdecke, D. (2022). sjPlot: Data visualisation for statistics in social science (2.8.12). [Computer software]. Retrieved from https://CRAN.R-project.org/package=sjPlot

Macmillan, N. A. (2002). Signal detection theory. In H. Pashler & J. Wixted (Eds.), *Stevens' handbook of experimental psychology: Methodology in experimental psychology* (vol. 4, 3rd ed., pp. 43–90). John Wiley & Sons Inc.

Maguinness, C., Roswandowitz, C., & von Kriegstein, K. (2018). Understanding the mechanisms of familiar voice-identity recognition in the human brain. *Neuropsychologia*, *116*, 179–193. https://doi.org/10.1016/j.neuropsychologia.2018.03.039

Mahrholz, G., Belin, P., & McAleer, P. (2018). Judgements of a speaker's personality are correlated across differing content and stimulus type. *PLOS ONE*, *13*(10). https://doi.org/10.1371/journal.pone.0204991

Marozeau, J., & de Cheveigné, A. (2007). The effect of fundamental frequency on the brightness dimension of timbre. *The Journal of the Acoustical Society of America*, *121*(1), 383–387. https://doi.org/10.1121/1.2384910

Martin, K. D. (1999). *Sound-source recognition: A theory and computational model* [PhD Thesis]. Massachusetts Institute of Technology.

Masthoff, H. (1996). A report on a voice disguise experiment. *International Journal of Speech Language and the Law*, *3*(1), 160–167. https://doi.org/10.1558/ijsll.v3i1.160

Matthews, G., & Desmond, P. A. (2002). Task-induced fatigue states and simulated driving performance. *The Quarterly Journal of Experimental Psychology Section A*, *55*(2), 659–686. https://doi.org/10.1080/02724980143000505

Mayer, M., & Ramon, M. (2023). Improving forensic perpetrator identification with super-recognizers. *Proceedings of the National Academy of Sciences*, *120*(20). https://doi.org/10.1073/PNAS.2220580120

McAleer, P., Todorov, A., & Belin, P. (2014). How do you say 'Hello'? Personality impressions from brief novel voices. *PLoS ONE*, *9*(3). https://doi.org/10.1371/journal.pone.0090779

McAllister, H. A., Bregman, N. J., & Lipscomb, T. J. (1988). Speed estimates by eyewitnesses and earwitnesses: How vulnerable to postevent information? *Journal of General Psychology*, *115*, 25–35. https://doi.org/10.1080/00221309.1988.9711085

McAllister, H. A., Dale, R. H. I., Bregman, N. J., McCabe, A., & Cotton, C. R. (1993). When eyewitnesses are also earwitnesses: Effects on visual and voice identifications. *Basic and Applied Social Psychology*, *14*(2), 161–170. https://doi.org/10.1207/s15324834basp1402_3

McAllister, H. A., Dale, R. H. I., & Keay, C. E. (1993). Effects of lineup modality on witness credibility. *The Journal of Social Psychology*, *133*(3), 365–376. https://doi.org/10.1080/00224545.1993.9712155

McCarty, C., Killworth, P. D., Bernard, H. R., Johnsen, E. C., & Shelley, G. A. (2001). Comparing two methods for estimating network size. *Human Organization*, *60*(1), 28–39. https://doi.org/10.17730/humo.60.1.efx5t9gjtgmga73y

McDougall, K (2004). Speaker-specific formant dynamics: An experiment on Australian English /aɪ/. *International Journal of Speech, Language and the Law*, 11(1), 103–130. https://doi.org/10.1558/sll.2004.11.1.103

McDougall, K. (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.

McDougall, K. (2021). Ear-catching versus eye-catching? Some developments and current challenges in earwitness identification evidence. *Proceedings of AISV XVII*, 33–56. https://doi.org/10.17469/O2108AISV000002

McDougall, K., & Duckworth, M. (2017). Profiling fluency: An analysis of individual variation in disfluencies in adult males. *Speech Communication, 95,* 16–27. https://doi.org/10.1016/j.specom.2017.10.001

McDougall, K., & Duckworth, M. (2018). Individual patterns of disfluency across speaking styles. *International Journal of Speech, Language and the Law*, *25*(2), 205–230. https://doi.org/10.1558/ijsll.37241

McDougall, K., Nolan, F., & Hudson, T. (2015). Telephone transmission and earwitnesses: Performance on voice parades controlled for voice similarity. *Phonetica, 72*(4), 257–272. https://doi.org/10.1159/000439385

McEntee-Atalianis, L. (2019). *Identity in applied linguistic research*. Bloomsbury.

McGehee, F. (1937). The reliability of the identification of the human voice. *The Journal of General Psychology, 17*(2), 249–271. https://doi.org/10.1080/00221309.1937.9917999

McGehee, F. (1944). An experimental study of voice recognition. *The Journal of General Psychology, 31*(1), 53–65. https://doi.org/10.1080/00221309.1944.10545219

McGorrery, P. G., & McMahon, M. (2017). A fair 'hearing': Earwitness identifications and voice identification parades. *The International Journal of Evidence & Proof, 21*(3), 262–286. https://doi.org/10.1177/1365712717690753

McGurk, H., & Macdonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*(5588), 746–748. https://doi.org/10.1038/264746a0

Memon, A., & Yarmey, A. D. (1999). Earwitness recall and identification: Comparison of the cognitive interview and the structured interview. *Perceptual and Motor Skills, 88*(3), 797–807. https://doi.org/10.2466/pms.1999.88.3.797

Meuwly, D. (2001). *Reconnaisance de Locuteurs en Sciences Forensiques: L'apport d'une Approche Automatique* [PhD Thesis]. University of Lausanne.

Meuwly, D. (2006). Forensic individualisation from biometric data. *Science & Justice, 46*(4), 205–213. https://doi.org/10.1016/S1355-0306(06)71600-8

Mileva, M., & Lavan, N. (2023). Trait impressions from voices are formed rapidly within 400 ms of exposure. *Journal of Experimental Psychology: General, 152*(6), 1539–1550. https://doi.org/10.1037/xge0001325

Mishra, P., Pandey, C. M., Singh, U., Gupta, A., Sahu, C., & Keshri, A. (2019). Descriptive statistics and normality tests for statistical data. *Annals of Cardiac Anaesthesia*, *22*(1), 67. https://doi.org/10.4103/ACA.ACA_157_18

Moos, A. (2010). Long-term formant distribution as a measure of speaker characteristics in read and spontaneous speech. *The Phonetician, 101*, 7–24.

Mühl, C., Sheil, O., Jarutytė, L., & Bestelmeyer, P. E. G. (2018). The Bangor Voice Matching Test: A standardized test for the assessment of voice perception ability. *Behavior Research Methods, 50*(6), 2184–2192. https://doi.org/10.3758/s13428-017-0985-4

Mullennix, J. W., Ross, A., Smith, C., Kuykendall, K., Conard, J., & Barb, S. (2011). Typicality effects on memory for voice: Implications for earwitness testimony. *Applied Cognitive Psychology*, *25*(1), 29–34. https://doi.org/10.1002/acp.1635

Mullennix, J. W., Stern, S. E., Grounds, B., Kalas, R., Flaherty, M., Kowalok, S., May, E., & Tessmer, B. (2009). Earwitness memory: Distortions for voice pitch and speaking rate. *Applied Cognitive Psychology*, *24*(4), 513-526. https://doi.org/10.1002/acp.1566

Myers, D. G. (2001). *Psychology*. Worth Publishers.

Nador, J. D., Zoia, M., Pachai, M. V., & Ramon, M. (2021). Psychophysical profiles in super-recognizers. *Scientific Reports*, *11*(1). https://doi.org/10.1038/S41598-021-92549-6

Nakamura, K., Kawashima, R., Sugiura, M., Kato, T., Nakamura, A., Hatano, K., Nagumo, S., Kubota, K., Fukuda, H., Ito, K., & Kojima, S. (2001). Neural substrates for recognition of familiar voices: A PET study. *Neuropsychologia*, *39*(10), 1047–1054. https://doi.org/10.1016/S0028-3932(01)00037-9

Nechanský, T., Bořil, T., Houzar, A., & Skarnitzl, R. (2022). The impact of mismatched recordings on an automatic-speaker-recognition system and human listeners. *AUC PHILOLOGICA*, *2022*(1), 11–22. https://doi.org/10.14712/24646830.2022.25

Neisser, U. (1967). *Cognitive Psychology*. Appleton-Century-Crofts.

Neuhauser, S. (2011). Foreign accent imitation and variation of VOT and voicing in plosives. *Proceedings of the 17th International Congress of Phonetic Sciences*, 1462–1465.

Neuhauser, S. (2012). *Phonetische und linguistische Aspekte der Akzentimitation im forensischen Kontext: Produktion und Perzeption*. Narr.

Neuner, F., & Schweinberger, S. R. (2000). Neuropsychological impairments in the recognition of faces, voices, and personal names. *Brain and Cognition*, *44*(3), 342–366. https://doi.org/10.1006/brcg.1999.1196

Njie, S., Lavan, N., & McGettigan, C. (2022). Talker and accent familiarity yield advantages for voice identity perception: A voice sorting study. *Memory & Cognition, 51*, 175–187. https://doi.org/10.3758/s13421-022-01296-0

Nolan, F. (1991). Forensic phonetics. *Journal of Linguistics*, *27*(2), 483–493. https://doi.org/DOI: 10.1017/S0022226700012755

Nolan, F. (2003). A recent voice parade. *International Journal of Speech, Language and the Law*, *10*, 277–291.

Nolan, F. (2005). Forensic speaker identification and the phonetic description of voice quality. In W. J. Hardcastle & J. M. Beck (Eds.), *A figure of speech: A festschrift for John Laver* (pp. 385–411). Lawrence Erlbaum Associates.

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

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

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*. 1506–1509. http://www.icphs2011.hk/resources/OnlineProceedings/RegularSession/Nolan/Nolan.pdf.

Nolan, F., & Oh, T. (1996). Identical twins, different voices. *International Journal of Speech, Language and the Law*, *3*(1), 39–49. https://doi.org/10.1558/ijsll.v3i1.39

Nolan, F. (1983). *The phonetic bases of speaker recognition.* Cambridge University Press.

Noonan, H., & Curtis, B. (2022). Identity. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2022 ed., n.p.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2022/entries/identity/

OED. (2023). "memorable, adj. and n." In *OED Online*. Oxford UP. www.oed.com/view/Entry/116339

Öhman, L., Eriksson, A., & Granhag, P. A. (2011). Overhearing the planning of a crime: Do adults outperform children as earwitnesses? *Journal of Police and Criminal Psychology*, *26*(2), 118–127. https://doi.org/10.1007/s11896-010-9076-5

Öhman, L., Eriksson, A., & Granhag, P. A. (2013). Angry voices from the past and present: Effects on adults' and children's earwitness memory. *Journal of Investigative Psychology and Offender Profiling*, *10*(1), 57–70. https://doi.org/10.1002/jip.1381

Pautz, N., McDougall, K., Mueller-Johnson, K., Nolan, F., Paver, A., & Smith, H. M. J. (2023). Identifying unfamiliar voices: Examining the system variables of sample duration and parade size. *Quarterly Journal of Experimental Psychology*, 174702182311557. https://doi.org/10.1177/17470218231155738

Peretz, I., Kolinsky, R., Tramo, M., Labrecque, R., Hublet, C., Demeurisse, G., & Belleville, S. (1994). Functional dissociations following bilateral lesions of auditory cortex. *Brain*, *117*(6), 1283–1301. https://doi.org/10.1093/brain/117.6.1283

Perfect, T. J., Hunt, L. J., & Harris, C. M. (2002). Verbal overshadowing in voice recognition. *Applied Cognitive Psychology*, *16*(8), 973–980. https://doi.org/10.1002/acp.920

Philippon, A. C., Cherryman, J., Bull, R., & Vrij, A. (2007). Earwitness identification performance: The effect of language, target, deliberate strategies and indirect measures. *Applied Cognitive Psychology*, *21*(4), 539–550. https://doi.org/10.1002/acp.1296

Plante-Hébert, J., Boucher, V. J., & Jemel, B. (2021). The processing of intimately familiar and unfamiliar voices: Specific neural responses of speaker recognition and identification. *PloS One*, *16*(4), e0250214–e0250214. https://doi.org/10.1371/journal.pone.0250214

Police Scotland. (2021). *Identification Procedures - National Guidance, Version 1.00* (SCD - Crime Strategy, Ed.). https://www.scotland.police.uk/spa-media/i3tpzmsq/identification-procedures-national-guidance.doc

Pollack, I., & Ficks, L. (1954). Information of elementary multidimensional auditory displays. *The Journal of the Acoustical Society of America*, *26*(2), 155–158. https://doi.org/10.1121/1.1907300

Pollack, I., Pickett, J. M., & Sumby, W. H. (1954). On the identification of speakers by voice. *The Journal of the Acoustical Society of America*, *26*(3), 403–406. https://doi.org/10.1121/1.1907349

Ramon, M. (2019). Super-recognizers in criminal investigation: Hype or hope? *Journal of Vision*, *19*(10), 137a. https://doi.org/10.1167/19.10.137A

Ramon, M., Bobak, A. K., & White, D. (2019). Super-recognizers: From the lab to the world and back again. *British Journal of Psychology*, *110*(3), 461–479. https://doi.org/https://doi.org/10.1111/bjop.12368

Ramon, M., & Gobbini, M. I. (2018). Familiarity matters: A review on prioritized processing of personally familiar faces. *Visual Cognition*, *26*(3), 179–195. https://doi.org/10.1080/13506285.2017.1405134

Ramon, M., Miellet, S., Dzieciol, A. M., Konrad, B. N., Dresler, M., & Caldara, R. (2016). Super-memorizers are not super-recognizers. *PLoS ONE*, *11*(3). https://doi.org/10.1371/journal.pone.0150972

Rathborn, H. A., Bull, R. H., & Clifford, B. R. (1981). Voice recognition over the telephone. *Journal of Police Science and Administration*, *9*(3), 280–284.

Read, D., & Craik, F. I. M. (1995). Earwitness identification: Some influences on voice recognition. In *Journal of Experimental Psychology: Applied*, 1(1), 6–18. https://doi.org/10.1037/1076-898X.1.1.6

Reich, A. R. (1981). Detecting the presence of vocal disguise in the male voice. *The Journal of the Acoustical Society of America*, *69*(5), 1458–1461. https://doi.org/10.1121/1.385778

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), 1023–1028. https://doi.org/10.1121/1.383321

Remez, R. E., Fellowes, J. M., & Rubin, P. E. (1997). Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(3), 651–666. https://doi.org/10.1037/0096-1523.23.3.651

Rhodes, R. W. (2012). *Assessing the strength of non-contemporaneous forensic speech evidence* [PhD Thesis]. University of York.

Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity and Analogical Reasoning* (pp. 21–59). Cambridge University Press. https://doi.org/10.1017/CBO9780511529863.004

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and Item response theory analyses. *Journal of Statistical Software*, *17*(5). https://doi.org/10.18637/jss.v017.i05

Roach, P. (2009). *English phonetics and phonology: A practical course* (4th ed.). Cambridge UP.

Robertson, B., Vignaux, G. A., & Berger, C. E. H. (2016). *Interpreting evidence*. Wiley. https://doi.org/10.1002/9781118492475

Robson, J. (2017). A fair hearing? The use of voice identification parades in criminal investigations in England and Wales. *Criminal Law Review*, *1*, 36–50.

Robson, J. (2018). 'Lend me your ears'. *The International Journal of Evidence & Proof*, *22*(3), 218–238. https://doi.org/10.1177/1365712718782989

Rodman, R. D. (1998). Speaker recognition of disguised voices: A program for research. In H. M. Demirekler & A. Saranli (Eds.), *Proceedings of the Consortium on Speech Technology Conference on Speaker Recognition by Man and Machine: Directions for Forensic Applications* (pp. 9–22).

Roebuck, R., & Wilding, J. (1993). Effects of vowel variety and sample length on identification of a speaker in a line-up. *Applied Cognitive Psychology*, *7*(6), 475–481. https://doi.org/10.1002/acp.2350070603

Rosch, E. H. (1973). Natural categories. *Cognitive Psychology*, *4*(3), 328–350. https://doi.org/10.1016/0010-0285(73)90017-0

Rose, P. (1996). Between- and within-speaker variation in the fundamental frequency of Cantonese citation tones. In P. Davis & N. Fletcher (Eds.), *Vocal Fold Physiology: Controlling Complexity and Chaos* (pp. 307–324). Singular Publishing Group.

Rose, P. (2002). *Forensic speaker identification*. Taylor & Francis.

Rose, P., & Duncan, S. (1995). Naive auditory identification and discrimination of similar voices by familiar listeners. *International Journal of Speech, Language and the Law*, *2*(1), 1–17. https://doi.org/10.1558/ijsll.v2i1.1

Rostolland, D. (1982). Acoustic features of shouted voice. *Acustica*, *50*, 118–125.

Roswandowitz, C., Mathias, S. R., Hintz, F., Kreitewolf, J., Schelinski, S., & Von Kriegstein, K. (2014). Two cases of selective developmental voice-recognition impairments. *Current Biology*, *24*(19), 2348–2353. https://doi.org/10.1016/J.CUB.2014.08.048

Roswandowitz, C., Schelinski, S., & von Kriegstein, K. (2017). Developmental phonagnosia: Linking neural mechanisms with the behavioural phenotype. *NeuroImage*, *155*, 97–112. https://doi.org/https://doi.org/10.1016/j.neuroimage.2017.02.064

Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: People with extraordinary face recognition ability. *Psychonomic Bulletin & Review*, *16*(2), 252–257. https://doi.org/10.3758/PBR.16.2.252

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),* n.p.

Saslove, H., & Yarmey, A. D. (1980). Long-term auditory memory: Speaker identification. *Journal of Applied Psychology*, *65*(1), 111–116. https://doi.org/10.1037/0021-9010.65.1.111

Schäfer, S. (2017). *An empirical approach to the relevance of intonation to negotiating givenness* [Unpublished postgraduate dissertation]. Universität des Saarlandes.

Schall, S., Kiebel, S. J., Maess, B., & von Kriegstein, K. (2015). Voice identity recognition: Functional division of the right STS and its behavioral relevance. *Journal of Cognitive Neuroscience*, *27*(2), 280–291. https://doi.org/10.1162/jocn_a_00707

Schiller, N. O., & Köster, O. (1998). The ability of expert witnesses to identify voices: A comparison between trained and untrained listeners. *International Journal of Speech, Language and the Law*, *5*(1), 1–9. https://doi.org/10.1558/ijsll.v5i1.1

Schlichting, F., & Sullivan, K. P. H. (1997). The imitated voice: A problem for voice line-ups? *International Journal of Speech Language and the Law*, *4*(1), 148–165. https://doi.org/10.1558/ijsll.v4i1.148

Schötz, S. (2006). *Perception, analysis and synthesis of speaker age* [PhD dissertation]. Lund University.

Schweinberger, S. R., Herholz, A., & Sommer, W. (1997). Recognizing famous voices. *Journal of Speech, Language, and Hearing Research*, *40*(2), 453–463. https://doi.org/10.1044/jslhr.4002.453

Schweinberger, S. R., & Zäske, R. (2018). Perceiving speaker identity from the voice. In S. Frühholz & P. Belin (Eds.), *The Oxford Handbook of Voice Perception* (pp. 539–560). Oxford University Press.

Seale-Carlisle, T. M., & Mickes, L. (2016). US line-ups outperform UK line-ups. *Royal Society Open Science*, *3*(9), 160300. https://doi.org/10.1098/rsos.160300

Semmler, C., Mickes, L., Dunn, J., & Wixted, J. T. (2018). The role of estimator variables in eyewitness identification. *Journal of Experimental Psychology: Applied*, *24*(3), 400–415. https://doi.org/10.1037/xap0000157

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3/4), 591. https://doi.org/10.2307/2333709

Shen, C., & Watt, D. (2015). Accent categorisation by lay listeners: Which type of "native ear" works better? *York Papers in Linguistics Series 2*, *14*, 106–131.

Sherrin, C. (2015). Earwitness Evidence: The reliability of voice identifications. *Osgoode Legal Studies Research Paper Series*, *11*(6), 2–44. http://digitalcommons.osgoode.yorku.ca/olsrpshttp://digitalcommons.osgoode.yorku.ca/olsrps/101

Shirt, M. (1984). An auditory speaker-recognition experiment. *Proceedings of the Institute of Acoustics*, *6*(1), 101–104.

Siedenburg, K., & McAdams, S. (2017). Four distinctions for the auditory "wastebasket" of timbre. *Frontiers in Psychology*, *8(1747), 1-4*. https://doi.org/10.3389/fpsyg.2017.01747

Simon, B., & Klandermans, B. (2001). Politicized collective identity: A social psychological analysis. *American Psychologist*, *56*(4), 319–331. https://doi.org/10.1037/0003-066X.56.4.319

Skuk, V. G., & Schweinberger, S. R. (2013). Gender differences in familiar voice identification. *Hearing Research*, *296*, 131–140. https://doi.org/10.1016/j.heares.2012.11.004

Smith, H. M. J., Baguley, T. S., Robson, J., Dunn, A. K., & Stacey, P. C. (2019). 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

Smith, H. M. J., Bird, K., Roeser, J., Robson, J., Braber, N., Wright, D., & Stacey, P. C. (2020). Voice parade procedures: Optimising witness performance. *Memory*, *28*(1), 2–17. https://doi.org/10.1080/09658211.2019.1673427

Smith, H. M. J., Roeser, J., Pautz, N., Davis, J. P., Robson, J., Wright, D., Braber, N., & Stacey, P. C. (2022). Evaluating earwitness identification procedures: Adapting pre-parade instructions and parade procedure. https://doi.org/10.1080/09658211.2022.2129065

Smith, I., Foulkes, P., & Sóskuthy, M. (2017). Speaker identification in whisper. *Letras de Hoje*, *52*(1), 5–14. https://doi.org/10.15448/1984-7726.2017.1.26659

Solan, L. M., & Tiersma, P. M. (2003). Hearing voices: Speaker identification in court. *Hastings Law Journal*, *54*(2), 373–436.

Sorensen, D., & Horii, Y. (1982). Cigarette smoking and voice fundamental frequency. *Journal of Communication Disorders*, *15*(2), 135–144. https://doi.org/10.1016/0021-9924(82)90027-2

Sørensen, M. H. (2012). Voice line-ups: Speakers' f0 values influence the reliability of voice recognitions. *International Journal of Speech, Language and the Law*, *19*(2), 145–158. https://doi.org/10.1558/ijsll.v19i2.145

Stevenage, S. (2018a). Voice processing implications for earwitness testimony. In S. Frühholz & P. Belin (Eds.), *The Oxford Handbook of Voice Perception* (pp. 626–644). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198743187.013.28

Stevenage, S. V. (2018b). Drawing a distinction between familiar and unfamiliar voice processing: A review of neuropsychological, clinical and empirical findings. *Neuropsychologia*, *116*(Pt B), 162–178. https://doi.org/10.1016/J.NEUROPSYCHOLOGIA.2017.07.005

Stevenage, S. V., Clarke, G., & McNeill, A. (2012). The "other-accent" effect in voice recognition. *Journal of Cognitive Psychology*, *24*(6), 647–653. https://doi.org/10.1080/20445911.2012.675321

Stevenage, S. V., Tomlin, R., Neil, G. J., & Symons, A. E. (2021). May I speak freely? The difficulty in vocal identity processing across free and scripted speech. *Journal of Nonverbal Behavior*, *45*(1), 149–163. https://doi.org/10.1007/s10919-020-00348-w

Stowe, L. M., & Golob, E. J. (2013). Evidence that the Lombard effect is frequency-specific in humans. *The Journal of the Acoustical Society of America*, *134*(1), 640–647. https://doi.org/10.1121/1.4807645

Szczepek Reed, B. (2001). *Prosodic orientation in spoken interaction*. University of Bayreuth Press.

Szczepek Reed, B. (2003). *Prosodic orientation in English conversation* [PhD Dissertation]. University of Potsdam.

Szczepek Reed, B. (2011). *Analysing conversation: An introduction to prosody*. Palgrave Macmillan.

Tajfel, H. (1978). Social categorization, social identity and social comparison. In H. Tajfel (Ed.), *Differentiation between Social Groups: Studies in the Social Psychology of Intergroup Relations* (pp. 61–76). Academic Press.

Tajfel, H. (1981). *Human groups and social categories: Studies in social psychology*. Cambridge University Press.

Tardif, J., Morin Duchesne, X., Cohan, S., Royer, J., Blais, C., Fiset, D., Duchaine, B., & Gosselin, F. (2019). Use of face information varies systematically from developmental prosopagnosics to super-secognizers. *Psychological Science*, *30*(2), 300–308. https://doi.org/10.1177/0956797618811338

t'Hart, J., Collier, R., & Cohen, A. (1990). *A perceptual study of intonation: An experimental-phonetic approach*. Cambridge UP.

Thompson, M. B., Tangen, J. M., & Searston, R. A. (2014). Understanding expertise and non-analytic cognition in fingerprint discriminations made by humans. *Frontiers in Psychology*, *5*(737). https://doi.org/10.3389/fpsyg.2014.00737

Tomlin, R. J., Stevenage, S. v, & Hammond, S. (2017). Putting the pieces together: Revealing face–voice integration through the facial overshadowing effect. *Visual Cognition*, *25*(4–6), 629–643. https://doi.org/10.1080/13506285.2016.1245230

Tompkinson, J., & Watt, D. (2018). Assessing the abilities of phonetically untrained listeners to determine pitch and speaker accent in unfamiliar voices. *Language and Law*, *5*(1), 19–37.

Trouvain, J. (2014). Laughing, breathing, clicking: The prosody of nonverbal vocalisations. *Proceedings of Speech Prosody*, 598–602.

Trouvain, J., & Truong, K. P. (2012). Comparing non-verbal vocalisations in conversational speech corpora. *Proceedings of the LREC Workshop on Corpora for Research on Emotion Sentiment and Social Signals*, 36–39.

Unfamiliar voice identification: Effect of post-event information on accuracy and voice ratings. (2014). *Journal of European Psychology Students*, *5*(1), 59–68. https://doi.org/10.5334/jeps.bs

Vaissière, J. (2005). Perception of intonation. In D. Pisoni & R. Remez (Eds.), *The handbook of speech and perception* (pp. 236–289). Blackwell.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology - Section A*, *43*(2), 161–204. https://doi.org/10.1080/14640749108400966

Valentine, T., & Bruce, V. (1986). The effects of distinctiveness in recognising and classifying faces. *Perception*, *15*(5), 525–535. https://doi.org/10.1068/p150525

Van Lancker, D., & Canter, G. J. (1982). Impairment of voice and face recognition in patients with hemispheric damage. *Brain and Cognition*, *1*(2), 185–195. https://doi.org/https://doi.org/10.1016/0278-2626(82)90016-1

Van Lancker, D., & Kreiman, J. (1986). Familiar voice recognition and unfamiliar voice discrimination are independent and unordered abilities. *The Journal of the Acoustical Society of America*, *79*(S1), S8–S8. https://doi.org/10.1121/1.2023449

Van Lancker, D., & Kreiman, J. (1987). Voice discrimination and recognition are separate abilities. *Neuropsychologia*, *25*(5), 829–834. https://doi.org/10.1016/0028-3932(87)90120-5

Van Lancker, D., Kreiman, J., & Cummings, J. (1989). Voice perception deficits: Neuroanatomical correlates of phonagnosia. *Journal of Clinical and Experimental Neuropsychology*, *11*(5), 665–674. https://doi.org/10.1080/01688638908400923

Van Lancker, D., Kreiman, J., & Emmorey, K. (1983). Recognition of famous voices forwards and backwards. *The Journal of the Acoustical Society of America*, *74*(S1), S50. https://doi.org/10.1121/1.2021007

Van Lancker, D., Kreiman, J., & Wickens, T. D. (1985). Familiar voice recognition: Patterns and parameters: Recognition of rate-altered voices. *Journal of Phonetics*, *13*, 39–52.

von Kriegstein, K., Eger, E., Kleinschmidt, A., & Giraud, A. L. (2003). Modulation of neural responses to speech by directing attention to voices or verbal content. *Cognitive Brain Research*, *17*(1), 48–55. https://doi.org/10.1016/S0926-6410(03)00079-X

von Kriegstein, K., & Giraud, A.-L. (2004). Distinct functional substrates along the right superior temporal sulcus for the processing of voices. *NeuroImage*, *22*(2), 948–955. https://doi.org/10.1016/j.neuroimage.2004.02.020

Wagner, I., & Köster, O. (1999). Perceptual recognition of familiar voices using falsetto as a type of voice disguise. *ICPhS-14*, 1381–1384.

Walter, T. J. (1992). Voice identification: Levels-of-processing and the relationship between prior description accuracy and recognition accuracy. *Proceedings of the Annual Meeting of the American Psychological Association*, 1–51.

Warren, J. D., Scott, S. K., Price, C. J., & Griffiths, T. D. (2006). Human brain mechanisms for the early analysis of voices. *NeuroImage*, *31*(3), 1389–1397. https://doi.org/10.1016/j.neuroimage.2006.01.034

Watkinson, J. (2002). *An introduction to digital audio* (2nd ed.). Focal Press.

Watt, D. (2010). The identification of the individual through speech. In C. Llamas & D. Watt (Eds.), *Language and identities* (pp. 76–85). Edinburgh University Press.

Watt, D., & Burns, J. (2012). Verbal descriptions of voice quality differences among untrained listeners. *York Papers in Linguistics Series 2*(12a), *1-28*.

Wells, G. L. (1978). Applied eyewitness-testimony research: System variables and estimator variables. *Journal of Personality and Social Psychology*, 36, (12), 1546–1557. https://doi.org/10.1037/0022-3514.36.12.1546

Wertheimer, M. (1925). Über Gestalttheorie. *Philosophische Zeitschrift Für Forschung Und Aussprache*, *1*, 39–60.

Wester, M. (2012). Talker discrimination across languages. *Speech Communication*, *54*(6), 781–790. https://doi.org/https://doi.org/10.1016/j.specom.2012.01.006

Whitmore, J., & Fisher, S. (1996). Speech during sustained operations. *Speech Communication*, *20*(1–2), 55–70. https://doi.org/10.1016/S0167-6393(96)00044-1

Wilding, J., & Cook, S. (2000). Sex differences and individual consistency in voice identification. *Perceptual and Motor Skills*, *91*(2), 535–538. https://doi.org/10.2466/pms.2000.91.2.535

Winters, S. J., Levi, S. v, & Pisoni, D. B. (2008). Identification and discrimination of bilingual talkers across languages. *The Journal of the Acoustical Society of America*, *123*(6), 4524–4538. https://doi.org/10.1121/1.2913046

Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest*, *18*(1), 10–65. https://doi.org/10.1177/1529100616686966

Wolfram, W. (2004). Social varieties of American English. In E. Finegan & J. R. Rickford (Eds.), *Language in the USA: Themes for the twenty-first century*. Cambridge UP.

Yarmey, A. D. (1991). Voice identification over the telephone. *Journal of Applied Social Psychology*, *21*(22), 1868–1876. https://doi.org/https://doi.org/10.1111/j.1559-1816.1991.tb00510.x

Yarmey, A. D. (1995). Earwitness speaker identification. *Psychology, Public Policy, and Law*, *1*, 792–816. https://doi.org/10.1037/1076-8971.1.4.792

Yarmey, A. D., & Matthys, E. (1992). Voice identification of an abductor. *Applied Cognitive Psychology*, *6*(5), 367–377. https://doi.org/10.1002/acp.2350060502

Yarmey, A. D., Yarmey, A. L., Yarmey, M. J., & Parliament, L. (2001). Commonsense beliefs and the identification of familiar voices. *Applied Cognitive Psychology*, *15*(3), 283–299. https://doi.org/https://doi.org/10.1002/acp.702

Yonelinas, A.P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, *46*(3), 441-517.

Young, A. W., Frühholz, S., & Schweinberger, S. R. (2020). Face and voice perception: Understanding commonalities and differences. *Trends in Cognitive Sciences 24*(5), 398–410. https://doi.org/10.1016/j.tics.2020.02.001

Yovel, G., & Belin, P. (2013). A unified coding strategy for processing faces and voices. *Trends in Cognitive Sciences*, *17*(6), 263–271. https://doi.org/10.1016/j.tics.2013.04.004

Zetterholm, E. (2003). The same but different – three impersonators imitate the same target voices. *Proceedings of the 15th International Congress of Phonetic Sciences*, 2205–2208.

Zetterholm, E., Sarwar, F., Thorvaldsson, V., & Allwood, C. M. (2012). Earwitnesses: the effect of type of vocal differences on correct identification and confidence accuracy. *International Journal of Speech, Language and the Law*, *19*(2), 219–237. https://doi.org/10.1558/ijsll.v19i2.219

Zhang, W., Xie, Y., Lin, B., Wang, L., & Zhang, J. (2022). Estimation of the underlying f0 range of a speaker from the spectral features of a brief speech input. *Applied Sciences*, *12*(13), 6494. https://doi.org/10.3390/app12136494

Zhou, X., & Mondloch, C. J. (2016). Recognizing "Bella Swan" and "Hermione Granger": No own-race advantage in recognizing photos of famous faces. *Perception*, *45*(12), 1426–1429. https://doi.org/10.1177/0301006616662046