Hello, who is this? The relationship between linguistic and speaker-dependent information in the acoustics of consonants

Smorenburg, B.J.L.

CHAPTER 1

General introduction

1.1 Studying speaker variation

In abstractionist models of speech perception and production, speaker variation has generally been regarded as noise or an obstacle to overcome, focusing on experimental effects and excluding as much speaker variation as possible. Over the last decade, however, the increasingly experimental nature of phonetics and phonology, where larger databases of speech are being used, has placed a particular focus on speaker variation in speech and on exemplar and episodic (as well as

hybrid) models of speech (cf. Bürki, 2018). In larger datasets, accounting for individual variation has become possible and forms an area of interest in itself. The change in focus from exclusively experimental effects to individual variation is very clearly observable in the popular statistical methods that are used in phonetics and phonology; mixed-effects models – where one can include individual speaker intercepts and slopes when modelling main effects – seem to have taken over from the formerly popular analysis of variance (ANOVA). Moreover, research fields for which speaker variation was considered a particular problem, such as automatic speech recognition, are now utilizing speaker variation to their advantage.

Looking at the history of automatic speech recognition, speaker variation has been one of the main sources of errors, especially when speakers have varying levels of fluency (e.g., Benzeghiba et al., 2007). The initial focus in this field was therefore to build speaker-*independent* speech recognition systems, which were built to function as accurately as possible despite speaker variation, much like abstractionist models of speech. Mirroring exemplar and episodic models of speech perception, later automatic speech recognition systems have integrated speaker information in so-called speaker-dependent systems and speaker-adaptive systems, with particular success for the latter (e.g., Saon, Soltau, Nahamoo, & Picheny, 2013; Rudzicz, 2007). Particularly in dysarthric speakers, who can benefit from automatic speech recognition to help with daily communication, the speaker-dependent and speaker-adaptive systems are preferred as the increased variability in these populations combined with the scarcity of data make the development of a well-performing speaker-independent (or abstractionist) system nearly impossible (cf. Shahamiri, 2021). These findings exemplify that speaker information does not necessarily have to be an obstacle, but can be taken advantage of, even in the speech recognition field where speaker variation has been the main source of errors.

These past developments in automatic speech recognition find some parallels in research on speech perception, where researchers have often questioned how learners acquire speech categories when there is so much talker variation (e.g., Weatherholtz & Jaeger, 2016). Normalization models of speech argue that talker variability, particularly the variability

associated with the shape and size of the vocal tract, is partially filtered out from perception by focusing on *relational* acoustic information (see Johnson, 2005 for a review on normalization models), whereas exemplar and episodic models argue that fine phonetic detail is stored and actively used in speech perception (e.g., Goldinger, 1998; Bradlow, Nygaard & Pisoni, 1999). Although listeners have been shown to store fine phonetic detail (Bradlow et al., 1999) along with relevant social context (Sumner, Kim, King, & McGowan, 2014), in exemplar and episodic models of speech, what makes a social context relevant to store and rely on in speech perception is often not quantified (cf. Kleinschmidt, 2019).

The 'ideal adapter' model of speech perception seeks to explain how speech perception depends on grouping talkers together to cope with talker variability (e.g. Kleinschmidt, 2019). In other words, observing that talker variability is somewhat structured, listeners must learn which groups of talkers can be treated as similar to improve inferences and predictions about speech input. An attempt at quantifying useful social groupings to speech perception for vowel contrasts (e.g., /ɛ/ versus /æ/) and stop voicing (e.g., /p/ versus /b/), Kleinschmidt (2019) found that talker variability was structured by some social variables but that these grouping were not necessarily *useful* in perception. Rather, at least when using *non*-normalized vowel formants, talker-specific cue-category mapping strongly outperformed any larger social groupings (age, gender, dialect, as well as the interaction between gender and dialect), although some of these social groupings did lead to small advantages (i.e. prediction accuracy) compared to not using any groupings at all. When using Lobanov normalization for formant values, the amount and structure of the variability changed; Kleinschmidt (2019) found an increase in dialectal variability and decreases in talker and gender variability, which was also reflected in their respective utility in phonemic predictions and suggests that there might be some role to play for normalization strategies in ideal adapter models of speech perception.

It is not yet clear how normalization in speech perception would interact with stored exemplars. So far, the different normalization methods and estimations of talker variability have led to different conclusions. For example, whereas Kleinschmidt (2019) reported a decrease in informativity and utility of talker variability after

normalizing vowel formants, a recent paper comparing 16 different normalization methods on vowel formants (Voeten, Heeringa & Van de Velde, 2022), reported that normalization methods had no great effect on the between-speaker variability as estimated by the explained variance of random speaker effects. The contributions of normalization versus – or in addition to – exemplars in speech production thus constitute an ongoing debate.

A research field that – by definition – takes advantage of talker-dependent variation is forensic speech science, where speakers in different speech samples are compared in forensic speaker comparisons. These comparative analyses serve to investigate the likelihood that speakers on different recordings are the same – or a different – individual. Formally, they provide strength of evidence for the likelihood of the evidence (i.e., the commonalities and differences between the speech recordings) under the hypothesis that the speakers are the same individual versus the hypothesis that the speakers are different individuals. The commonalities and differences between two speech samples are estimated with reference to the typicality of speech features found in a selected reference/background population (the selection of a reference population is discussed further in section 1.3). For any given reference population, what needs more attention in this field is the *structure* of spoken language. When comparing the speakers in two samples of speech, does the linguistic environment that speech features are sampled from affect the amount and the type of talker-dependent information that is present? For example, linguistic factors such as prosodic structure and phonetic context have been shown to affect the acoustics of speech sounds and syllables, but how these factors affect the talker-dependent information is largely unclear in forensic contexts.

1.2 Sources of speaker variation

Speaker variation can stem from variation in the metaphorical hardware of a speaker (i.e., the vocal tract) or from variation in the metaphorical

software of a speaker (i.e., a speaker's acquired language knowledge). What makes forensic linguistics and forensic phonetics somewhat more complex than other forensic disciplines such as fingerprint or DNA analysis, is that neither the vocal tract nor a speaker's language are invariant (cf. Nolan, 2001). The vocal tract is a highly flexible organ with multiple moving structures (including the lips, tongue, jaw, velum as the primary articulators but also secondary moving structures such as the glottis) and so has many degrees of freedom. It is therefore highly unlikely that speakers will – or rather are able to – produce the exact same speech more than once. This is opposed to fingerprints or DNA, which are generally considered invariant for any individual. Not only is the structure of the vocal tract highly flexible, but illness, smoking, and operations can also – temporarily or permanently – alter this structure. Furthermore, other factors such as a person's mood, anxiety levels, and fatigue can also affect speech. Anxiety, e.g., can cause an increase in muscular tension in the vocal tract and ribcage, which can result in a higher pitch (Pisanski, Nowak, & Sorokowski, 2016). Because of its flexibility within speakers, Nolan (1983, p. 27) argues that the vocal tract is not like a fingerprint or DNA in the sense that it only limits the *range* of achievable values. For example, a speaker's pitch range is limited by the minimum and maximum fundamental frequency that can be achieved by their vocal folds.

Although it can in some cases be a source of within-speaker variation, the vocal tract is also considered the 'purest' form of between-speaker variation. Similar to fingerprints and DNA, individual speakers' vocal tracts have different shapes and sizes, and these are relatively invariant compared to the other source of between-speaker variation: the software. The metaphorical software constitutes all the behavioral, acquired aspects of language (cf. Nolan, 2001). Speakers vary because they speak different languages and regional varieties and, within languages and dialects, also in different styles and registers (cf. Biber & Conrad, 2005; Schilling, 2004). The software is considered highly variant because speakers acquire language (in the broadest sense) throughout their lives. Not only do individuals learn entire *new* languages, dialects, and sociolects, but the languages themselves are subject to change in terms of lexicon, grammar, and pronunciation. Additionally, speakers

adapt their speech to the contextual situation, particularly to the addressee (cf. audience design: Bell, 1984). For example, someone might use a more formal speech register when speaking on the phone with their manager than they do when speaking with their friends at a bar. Similarly, sociolects are often only used, or more pronounced, when conversing with speakers from the same social group (e.g., see Nortier & Dorleijn, 2008 on the use of Moroccan Dutch). As a result, two recordings of the same person can show high degrees of within-speaker variation. To conclude, both idiosyncrasies in physiology and acquired language behavior are sources of between-speaker variation, with the former being considered a purer form because of its relative constancy compared to acquired speech behavior, which is highly adaptable to different sociolinguistic contexts.

1.3 Forensic speaker comparisons

In forensic speaker comparisons (FSC), there is often the question whether a disputed speech sample and a speech sample from a suspect were produced by the same individual. A disputed speech sample might be a telephone conversation that was wiretapped by the police as part of a police investigation where the identity of the speaker is "disputed" and a speech sample from the suspect can be a police interview with the suspect. The first type of speech sample is "disputed" because the speaker has not been verified, whereas in the police interview, the speaker's identity is known. The answer to the question whether the speech in the disputed and suspect samples come from the same individual may provide evidence for or against criminal conviction of the suspect. Given the potential consequences of this answer for the suspect, precise and accurate methods are required in forensic speaker comparisons.

There is a general consensus amongst forensic analysts that it is desirable to work with the Bayesian framework (Gold & French, 2011; 2019; Rose, 2002). In this framework, the similarity of features in speech samples is always estimated with reference to the typicality of features

in speech samples as represented by a reference population. For example, imagine that the speakers in the disputed recording and the suspect's recording are both Dutch adult males with an average pitch height of 100 Hz; given that this pitch height is very typical for adult male speakers in English and Dutch language populations (cf. Simpson, 2009), this is very weak evidence that these two male speakers are the same individual. Including a reference population gives information about what kind of between-speaker variation is present in a relevant group of speakers. For a fair comparison, this group of speakers that we call the reference or background population should be similar to the speaker in the suspect recording. Usually, they are matched on biological sex and language variety, but other social factors have been also shown to affect forensic speaker comparisons. For example, it has been shown that varying age and socio-economic class for the relevant reference population significantly affected the strength of evidence obtained for individual speaker comparisons, obtaining best system validity when the reference population matched the speaker in the suspect recording on age and social class (Hughes & Foulkes, 2015).

In the Bayesian framework, the probability of the speech evidence under the same-speaker hypothesis (i.e., the hypothesis that the speakers in the disputed and suspect recording are the same individual) is weighted against the probability of the speech evidence under the different-speaker hypothesis (i.e., the hypothesis that the speakers in the disputed and suspect recording are not the same individual). A conclusion of a forensic speaker comparison is then expressed in a likelihood ratio as follows (e.g., Nederlands Forensisch Instituut, 2017):

$$LR = \frac{P(E|H_{SS})}{P(E|H_{DS})}$$

A likelihood ratio (LR) expresses the probability (P) of the evidence (E) occurring under the same-speaker hypothesis (H_{SS}) against the probability of the evidence occurring under the different-speaker hypothesis (H_{DS}).

Although state-of-the-art methods in forensic speaker comparisons include automatic speaker recognition (ASR), which uses highly detailed and highly dimensional acoustic information such as Mel-frequency cepstral coefficients (MFCCs), fully automatic methods are used less than other methods (Morrison et al., 2016). Even though ASR systems' accuracy can be tested, practitioners cannot explain exactly how forensic speaker comparison conclusions are derived with ASR in the way that they can with auditory-acoustic analysis. This can be problematic, and therefore even prohibited, in the legal context. Many forensic analysts in Europe therefore still predominantly use auditory-acoustic analysis for casework (Morrison et al., 2016). In auditory-acoustic analysis, the forensic analyst can make use of perceptual observations and acoustic measurements, which can then be compared across speech samples with reference to a reference/background population.

Useful speech sounds for auditory-acoustic forensic speaker comparisons are the ones that are highly speaker-specific. A highly speaker-specific sound is one that has high between-speaker variation and low within-speaker variation. Vowels typically outperform consonants when it comes to speaker discrimination (e.g., Van den Heuvel, 1996, however, see Schindler & Draxler, 2013 who suggest using spectral characteristics from nasal and fricative consonants over vowel formants). However, combining different speech sounds results in stronger evidence and so consonants are often included in forensic speaker comparisons (Gold & French, 2011). Despite the use of consonants by forensic practitioners, previous research has largely focused on vowels, resulting in scarce population statistics for consonants. The work that has been done on consonants seems to show that nasal and fricative consonants are relatively speaker-specific compared to other consonants (e.g., Kavanagh, 2012; Van den Heuvel, 1996). Nasal consonants, because they involve the relatively rigid and highly individually-shaped nasal cavity, have been observed to show low within-speaker variation and high between-speaker variation (cf. Rose, 2002). In other words, nasals seem to be rather speaker-specific because they are a good reflection of a speaker's anatomy. Fricatives, with a special focus on /s/ in the literature, have often been associated with

social variables such as social class (Stuart-Smith, 2007), sexual orientation (Munson, McDonald, DeBoe & White, 2006), and ethnicity (Ditewig et al., 2021). Although fricative /s/ also reflects the size of the vocal tract (see e.g., the difference in /s/ acoustics between male and female speakers in Jongman, Wayland & Wong, 2000), it seems that this sound is more easily manipulated by speakers to convey aspects of social identity. Given previous reports on the forensic usefulness of fricative and nasal speech sounds compared to other consonants, they are the focus of this dissertation.

For both fricative and nasal consonants, two sounds with high frequencies of occurrence in Standard Dutch (cf. Luyckx, Kloots, Coussé, & Gillis, 2007) were selected for segmentation and analysis, to ensure that enough tokens would be available in the spontaneous speech data worked with in this dissertation. Within the class of fricatives, alveolar /s/ and velar or uvular (depending on region) /x/ were selected. Within the class of nasals, alveolar /n/ and bilabial /m/ were selected. One sound, fricative /s/, was also segmented and analyzed in an English speech corpus. From the four consonants analyzed in Dutch, /s/ was chosen for English because previous research shows clear differences in the acoustics of English and Dutch /s/ (e.g., Quené, Orr, & Van Leeuwen, 2017).

1.4 Research questions

One methodological issue with regards to estimating the within- and between-speaker variation in forensic speaker comparisons is the question of sampling. When looking at the available speech recordings, does it matter where in the speech signal speech is sampled from? Forensic speech science has partly answered this question by investigating the speaker-specificity of different speech sounds, but language is structured in several other ways which might be relevant to taking speech samples. Many linguistic factors are shown to (sometimes greatly) affect the acoustic realizations of speech sounds and it should

not be assumed that these acoustic differences do not affect the within- and/or between-speaker variation. In fact, some previous research has shown that several linguistic factors can affect the acoustics and speaker-specificity of segments. For example, acoustic-phonetic research has long shown that linguistic factors such as lexical stress and word class affect the acoustic realization of vowels (e.g., see Van Bergem, 1995 on vowels in Dutch). More recent work has shown that these acoustic differences also affect the speaker specificity: Stressed vowels are slightly more speaker-specific than unstressed vowels (McDougall, 2004) and vowels from content words are somewhat more speaker-specific than vowels from function words (Heeren, 2020a).

In this dissertation, I investigated how linguistic factors affect the acoustics and speaker-specificity of consonants. Consonants were selected because they are rather understudied compared to vowels in the forensic context. The main research question of this dissertation is therefore: How do linguistic factors affect the speaker-dependent variability of consonantal speech sounds? A set of consonants that has previously been shown to be rather speaker-specific was selected, namely a set of fricative and nasal consonants, to make results maximally useful to forensic speech comparisons. Given that, in forensic speaker comparisons, one does not always receive high-quality speech recordings for analysis, a qualitative question was raised with regards to the recording type of speech evidence, specifically the comparison between wiretapped telephone recordings and higher-quality recordings. Although there has been some work on the effect of the telephone band on vowels, its effect on consonants is not yet clear. Neither is it clear from the literature whether linguistic effects on speech sound acoustics are observable in lower-quality recordings such as wiretapped telephone recordings. To investigate whether linguistic factors are relevant in a medium that is often used in forensic speaker comparisons, i.e., telephone conversations, it is necessary to investigate the effects of linguistic factors on the acoustics and speaker information across different recording types. Lastly, it was asked whether obtained results would be applicable in forensic speaker comparisons, specifically whether the strength of evidence in forensic speaker comparison derived with

Bayesian likelihood-ratio analysis would be affected by sampling tokens from different linguistic environments.

To answer the main research question, first the linguistic factors that affect the acoustic realizations of fricative and nasal consonants need to be identified. The acoustics of consonants can generally be affected by two types of linguistic factors: structural factors and contextual factors. Linguistic structure is acoustically realized as prosody, resulting in identifiable prosodic constituents in the speech signal. Contextual factors are taken to refer to coarticulation, i.e., the assimilation of speech sound features in connected speech. With regards to prosodic structure, initial elements of prosodic constituents are strengthened, i.e., articulated with more effort (e.g., Cho & McQueen, 2005; Fougeron, 2001; Redford & Diehl, 1999). This is considered particularly important in language acquisition to help parse the different constituents in running speech (e.g., Hawthorne, Mazuka & Gerken, 2015). Conversely, the literature also speaks of domain-final weakening, which has often been observed for syllables (cf. coda reduction: Ohala & Kawasaki, 1984, Recasens, 2004). With regards to the speaker variation, previous research indicates that a syllable's intensity and formant contours contained more between-speaker variation in the second half of syllables, i.e., the mouth-closing gesture towards the coda, than in the mouth opening gesture in the first half of syllables (He & Dellwo, 2017; He, Zhang, & Dellwo, 2019). This was explained by the relative constraint to reach a linguistic target on the first part of syllables versus the relative articulatory freedom in the second half. Onset consonants are generally more constrained than coda consonants, although this does seem to be conditioned by the specific consonant and their production constraints in various contexts, i.e., not all consonants reduce equally in coda position (Recasens, 2004). In perceptual speaker identification, effects of syllabic position have also been found, with higher accuracy for syllables with onsets than for onset-less syllables (Amino, Arai, & Sugawara, 2007).

These findings indicate that the amount of speaker information in segment acoustics might be distributed in systematic ways with regards to prosodic structure. Specifically, positions where there is articulatory strengthening are articulated with more speech effort and therefore have less within-speaker variation and positions where there is articulatory

weakening are articulated with less speech effort and therefore have more between-speaker variation. This results in two competing hypotheses to the general research question: prosodic domain-initial elements such as syllable onsets might be more speaker-specific because they are associated with lower within-speaker variation, or prosodic domain-final elements such as syllable codas might be more speaker-specific because they are associated with more between-speaker variation. There are two competing hypotheses because there are two ways for a speech sound to increase in speaker-specificity: either the within-speaker variation has to decrease relative to the between-speaker variation, or the between-speaker variation has to increase relative to the within-speaker variation.

With regards to effects of phonetic context on consonant acoustics, individual speech sounds are affected by the surrounding speech sounds in connected speech as a result of having to move the articulators from one articulatory target to the next in quick succession. This coarticulation may contain speaker-specific information, due to speaker-dependent differences in the timing and degree of the coarticulatory movements (cf. Nolan, 1983, Ch. 3). These idiosyncrasies in coarticulation are partially explained by idiosyncrasies in anatomy; the shape of the palate, the dimensions of the vocal tract, and the biomechanics of the tongue all contribute to idiosyncrasies in articulation (Weirich, 2015). For both fricative and nasal consonants, previous literature was consulted to identify specific phonetic contexts that may be expected to affect the acoustic realizations of these sounds. It was then hypothesized that fricative and nasal consonants in contexts with high degrees of coarticulation would contain more between-speaker information – and be more speaker-specific – than fricative and nasal consonants in other phonetic contexts.

1.5 Practical considerations

Because this dissertation aims to investigate some within-speaker factors, namely linguistic environment, in forensic speaker comparisons, this

section discusses some practical considerations related to the forensic field.

First and foremost, in real forensic speaker comparisons, analysts would not limit themselves to analyzing only fricative and nasal consonants. Auditory-acoustic analysis often consists of both linguistic and suprasegmental voice and speech characteristics (e.g., the general nasality of a speaker, use of stop words, or disfluencies) and segmental features, which can be supplemented by acoustic measurements (cf. Nederlands Forensisch Instituut, 2020). Segmental features often include both vocalic and consonantal features (Gold & French, 2011), but consonantal features are relatively understudied. On top of that, auditory-acoustic analyses are nowadays often supplemented with automatic speech recognition (Gold & French, 2019). All that is to say, this dissertation is not suggesting that only considering nasal and fricative consonants is sufficient or a recommended course of action in forensic speaker comparisons. Rather, the question this dissertation aimed to answer is whether the linguistically-structured acoustic variation reported on in the phonetic literature influences the within- versus between-speaker variation in segments.

The linguistic factors that were considered for the linguistic environment in which these nasal and fricative consonants occur (phonetic context and syllabic position) were firstly derived from previous literature. There are, of course, many more linguistic factors which have been shown to affect these consonants. The first experiment on fricative consonants /s/ and /x/ therefore initially contained some secondary factors such as position in the phrase, lexical stress, word class, and articulation rate of the phrase. However, even in a relatively large speech corpus such as component-c of the Spoken Dutch Corpus (Oostdijk, 2000), which contains one to four ten-minute telephone conversations per speaker (M = 1.8), considering prosodic constituents larger than the syllable and word led to insufficient data per speaker to do any sort of speaker-classification tests. Given that scarcity of data is a common problem in forensic speaker comparisons, I decided that variables that are not possible to analyze in the Spoken Dutch Corpus would not be considered further. Additionally, previous research on consonant acoustics indicates that, with regards to linguistic structure in

within-speaker designs, the immediate phonetic context (e.g. Koenig et al., 2013; Su et al., 1974) and syllabic position (Byrd et al., 2009; Krakow, 1993; Ohala & Kawasaki, 1984) seemed particularly important factors to consider. There are other, non-linguistic, within-speaker factors that are known to affect forensic speaker comparisons, such as diachronic recordings (Audibert, Fougeron & Chardenon, 2021) and familiarity with the interlocuter (e.g., Kachkovskaia et al., 2022), but they fall outside of the scope of this dissertation, which only investigated linguistic structure synchronically. Although outside of the scope of this dissertation, some of the other works in the larger project group that this dissertation is part of have investigated factors such as lexical stress, word class, phrasal position, and diachronic recordings (see e.g., Heeren, 2020a on word class and De Boer, Quené, & Heeren, 2022 on the consistency of filled pauses in diachronic recordings).

With regards to the data used in this dissertation, I chose to limit this dissertation to two existing speech corpora. The first, component-c of the Spoken Dutch Corpus (Oostdijk, 2000), was selected for the nature of the speech task and the signal characteristics. What is rather exceptional about this speech corpus is that there was no speech task beyond having a telephone conversation with one to four different interlocuters that were previously known to the speaker (e.g., a parent or colleague) for around ten minutes in their own home environment. It therefore includes a lot of spontaneity and variation that is typically not present in speech corpora. For example, speakers make jokes and laugh, or they get annoyed, or even angry, when they feel that the interlocuter is not contributing enough to the conversation. Additionally, different background noises can be heard such as a crying baby or pet bird. Although these uncontrolled recording conditions are generally regarded as undesirable, it somewhat mimics the variability one might expect in real forensic data. Regarding the signal characteristics, these landline conversations were wiretapped from a switchboard. Forensic case work often has to deal with wiretapped telephone material. Although the landline signals can be considered somewhat old fashioned compared to the higher-quality 5G networks often in use today, they are not obsolete. Network selection depends on the technical specifications of the telephones used (the telephone with the lower specifications determines

the network selection, so a conversation between a newer 5G-compatible mobile phone and an older 3G-compatible phone will communicate on a 3G network). Criminals often use cheap prepaid mobile telephones, referred to as 'burner' phones, which are used temporarily with the intended purpose for more anonymous communication (cf. Bosma et al., 2020). These phones are likely to operate on earlier generation networks with specifications comparable to the landline signals. Dutch landline signals have a stable bandwidth of 300 – 3,400 Hz and 2G and 3G mobile networks have a 200 – 3,400 Hz bandwidth (Besette et al., 2002) with varying bitrates that can lower the upper band limit to 2,800 depending on signal conditions (Guillemin & Watson, 2006). A somewhat recent corpus for English, the West Yorkshire Regional English Database (WYRED: Gold, Ross & Earnshaw, 2018), that was gathered for the purpose of forensic phonetic research, still chose to use the landline signal for their telephone condition, suggesting its continued relevance. WYRED is the second corpus that was selected for this study.

Lastly, in section 1.1, I briefly discussed the role of talker variability in different models of speech. In a sociophonetic approach, one often considers social groupings in talker populations such as age, gender, and dialect. It has been shown that variability in acoustic phenomena such as coda reduction and coarticulation, which are studied in this dissertation, can be partially explained by social grouping variables. For example, word-final /n/ after schwa is less likely to be reduced/deleted in the Dutch speech from northern regions and in young males (Van de Velde & Van Hout, 2000) and seems to be rather speaker-specific due to different phonological processes – associated with word type and phonetic context – being used differently by different speakers (Van de Velde & Van Hout, 2001). However, from a forensic perspective, the focus is on individuals and not groups. Importantly, we want to be able to distinguish individuals within a population that, ideally, is similar in terms of these social groupings. Social grouping variables were therefore not considered beyond delimiting the datasets that were worked with in this dissertation. For the Spoken Dutch Corpus, only adult male speakers of Standard Dutch as their home, work and education language (aged 18 to 50) were considered. For the WYRED corpus, only adult male speakers from one area, namely Wakefield in Yorkshire, were

considered. In forensic speaker comparisons, social variables are considered when selecting the relevant background/reference population. As mentioned in section 1.3, speakers in the reference population should minimally be matched to the suspect recording on biological sex and language variety. Although it has been shown that other social grouping variables can have an effect on forensic speaker comparisons (e.g. a reference population matched on age and social class leads to somewhat better performance: Hughes & Foulkes, 2015), in practice this would mean that reference/background populations would need to become very specific for each case, which would mean that many forensic speaker comparisons would not be possible due to a lack of adequate specific reference populations. To make this even more complex, social groupings – or at least the expression of social identity in speech – also vary *within* speakers. For example, regarding sociolects as social group markers, speakers will display more standard language and pronunciation with speakers outside of those groups than with speakers of the same sociolect (e.g., Nortier & Dorleijn, 2008). If taken into account, analysts would need to select not only a sufficient number of speakers for the reference population that are matched to the suspect on a number of social grouping variables, but the speech task (read versus spontaneous, monologue versus dyad) and interlocuter (relative age, gender and dialect compared to speaker as well as the relationship between them) will then also need to be matched to the specific speech recordings of the suspect. Instead of collecting specific reference materials for each case, in practice, preliminary investigations are conducted to evaluate the degree to which the materials are estimated to be representative of a speaker (e.g., having only one versus multiple interlocuters for the speaker) and also the comparability of the disputed and reference materials with regards to, e.g., the communicative context (cf. Nederlands Forensisch Instituut, 2020). As comparability gets weaker, the strength-of-evidence of the features involved also decreases. Importantly, in the forensic framework, including social variables as predictors is referred to as speaker profiling, which is not the same as forensic speaker comparisons in its purpose.

To summarize, this section aimed to explain the methodological choices made in this dissertation by relating them to considerations in the forensic field.

1.6 Outline of the dissertation

This dissertation reports on a series of studies on the speaker-specificity of different consonantal speech sounds, particularly focusing on effects of two linguistic factors on the speaker specificity of Dutch fricatives (chapter 2) and Dutch nasals (chapter 3) in spontaneous Dutch telephone conversations. In chapter 4, we tested whether findings from chapter 2 generalize across languages and across recording types (studio versus telephone speech). Specifically, the effects of both linguistic information and the narrowband telephone filter on the acoustics and speaker-specificity of British English fricative /s/ was examined. Finally, in chapter 5, the forensic validity of findings is tested by using the state-of-the-art Bayesian likelihood-ratio framework. Chapter 6 provides a general summary of the dissertation, the discussion of the overall results, the limitations of these studies, and suggestions for future research on this topic. Chapters 2 to 5 were written as independent manuscripts with their own introductions and conclusions. As a result, there is some overlap between the information in these chapters.