Universiteit
Leiden
The Netherlands

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

Smorenburg, B.J.L.

Citation
Smorenburg, B. J. L. (2023, June 28). *Hello, who is this?: The relationship between linguistic and speaker-dependent information in the acoustics of consonants*. *LOT dissertation series*. LOT, Amsterdam. Retrieved from https://hdl.handle.net/1887/3627840

Version:	Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/3627840

Note: To cite this publication please use the final published version (if applicable).

Hello, who is this?

The relationship between linguistic and speaker-dependent information in the acoustics of consonants

Published by
LOT
Binnengasthuisstraat 9
1012 ZA Amsterdam
The Netherlands

phone: +31 20 525 2461
e-mail: lot@uva.nl
http://www.lotschool.nl

ISBN: 978-94-6093-432-2
DOI: https://dx.medra.org/10.48273/LOT0647
NUR: 616

Copyright © 2023: Laura Smorenburg. All rights reserved.

Hello, who is this?

The relationship between linguistic and speaker-dependent information in the acoustics of consonants

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden
op gezag van rector magnificus prof. dr. ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op woensdag 28 juni 2023
klokke 10:00 uur

door

Laura Smorenburg

geboren te Tiel, Nederland
op 3 februari 1992

Promotor: Prof. dr. Niels O. Schiller

Copromotor: Dr. Willemijn Heeren

Promotiecommissie: Dr. Tina Cambier-Langeveld
(Immigratie- en Naturalisatiedienst,
Nederlands Forensisch Instituut)

Prof. dr. Yiya Chen

Prof. dr. Paul Foulkes
(University of York)

Prof. dr. Hans Van de Velde
(Fryske Akademy, Universiteit Utrecht)

Funding statement: This PhD project is part of the VIDI project *The speaker in speech* (principle investigator: dr. Willemijn Heeren) and was funded by the Dutch Research Council (276-75-010)

Contents

Acknowledgements

First and foremost, I'd like to thank my promotor and supervisor, Niels Schiller and Willemijn Heeren, for their thoughtful feedback and support throughout these four (plus) years. Willemijn, when asked about you throughout these years by other PhDs, I always said I felt really lucky to have you as a supervisor and still feel that way. I aspire to have your professionalism, time management and planning skills one day.

Also thanks to my wonderful colleagues at LUCL, in particular my PhD colleagues: Astrid van Alem, Meike de Boer, Hang Cheng, Lis Kerr, Rasmus Puggaard-Rode, Menghui Shi, Xander Vertegaal, Cesko Voeten, Andrew Wigman, Sarah von Grebmer zu Wolfsthurn, Jiang Wu, and everyone else. I really enjoyed our discussions at the lunch table, which very predictably ended up being about linguistics in all its shapes and forms. You have made me feel really at home at LUCL. Special thanks also to our weekly statistics meeting group with Willemijn, Cesko, Meike and Sanne Ditewig. Cesko, thanks for all your statistical advice and thanks for writing and publishing your R package buildmer, which has saved me many hours of (rather boring) work in R!

Most chapters in this dissertation have gone through the peer review system from journals and conferences, which has greatly improved their content and readability. I'd like to thank all peer reviewers who have provided their professional and detailed feedback to these manuscripts and abstracts, even when their recommendation was to

'reject'. I'd also like to thank all those who have asked questions and provided feedback at conferences; I really appreciate it.

Due to the pandemic, this PhD did not quite turn out as expected. In March 2020, I had just arrived in York for a research visit, which ended up being briefer than anticipated! Nevertheless, I'd still like to give my thanks to Erica Gold and Vince Hughes for their willingness to receive me at the University of Huddersfield and University of York. Upon returning home, my apartment also became my office, lunch-table discussions stopped and basically all conferences were cancelled and then moved online. My thanks to the members of the LUCL PhD Council, who invited me to be a member and organized online lunch meetings regularly. Willemijn and Meike, thanks for hosting project meetings at your homes (and for the baked goods). Thanks also to the YouTube channel Yoga with Adriene and the service point at Leiden University for bringing me my office chair when we had to work from home.

I'd also like to thank Aoju Chen. You (and admittedly several others) forwarded me the vacancy for this PhD position and encouraged me to apply.

To my book club members in The Hague, Arre, Lars and Linde, thanks for making sure I read some fiction while there were so many academic works to read instead.

Lastly, thanks to my friends and family, who I think still haven't quite grasped what I've been working on in this PhD – especially my family who, at one point, just seemed to inaccurately simplify it to 'she investigates wiretapped telephone conversations for the police, she's basically a detective'. Thanks for making my PhD project sound cool and thanks for letting me complain about it every now and then.

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.

CHAPTER 2

Linguistic effects on the speaker-dependent variability in fricatives

Abstract

Although previous work has shown that some speech sounds are more speaker-specific than others, not much is known about the speaker information of the same segment in different linguistic contexts. The present study therefore investigated whether Dutch fricatives /s/ and /x/ from telephone dialogues contain differential speaker information as a function of syllabic position and labial co-articulation. These linguistic effects, established in earlier work on read broadband speech, were firstly

investigated. Using a corpus of Dutch telephone speech, results showed that the telephone bandwidth captures the expected effects of perseverative and anticipatory labialization for dorsal fricative /x/, for which spectral peaks fall within the telephone band, but not for coronal fricative /s/, for which the spectral peak falls outside the telephone band. Multinomial logistic regression shows that /s/ contains slightly more speaker information than /x/ in telephone speech and that speaker information is distributed across the speech signal in a systematic way; even though differences in classification accuracy were small, codas and tokens with labial neighbors yielded higher scores than onsets and tokens with non-labial neighbors for both /s/ and /x/. These findings indicate that speaker information in the same speech sound is not the same across linguistic contexts.

This chapter was published:

Smorenburg, L., & Heeren, W. (2020). The distribution of speaker information in Dutch fricatives /s/ and /x/ from telephone dialogues. *Journal of the Acoustical Society of America, 147*(2), 949-960. doi: 10.1121/10.0000674

2.1 Introduction

Speakers' voices convey idiosyncratic information. In everyday communication, listeners make use of this information while interpreting what they hear and, in forensic phonetics, speech analyst use this information to acoustically characterize speakers. Although previous research has already shown that some speech sounds convey more speaker information than others (e.g., Kavanagh, 2012; Van den Heuvel, 1996), not much is known about how speaker information in the same speech sound interacts with its linguistic environment. The present study investigated the speaker-dependency of the same speech sound in different linguistic contexts. Specifically, we examined whether the speaker-dependency of Dutch fricatives varied as a function of syllabic position and labial co-articulation. Additionally, the aim was to determine which segment and which specific (combinations of) acoustic features are most successful in characterizing speakers. Contrary to many previous studies that used read speech, the present study used spontaneous telephone dialogues to investigate speaker variation.

Investigating the distribution of speaker information is relevant for forensic speech science because the role of the speaker in speech production is still largely unclear. It is known that speaker-dependent information conveys all kinds of meanings (e.g., gender identity) and that these meanings are also perceived by listeners. However, it is not clear where in the speech signal speakers have the articulatory freedom to convey speaker information, or if there are such distributional limitations. Additionally, this study may be particularly relevant for forensic speaker comparisons, where often low-quality speech samples are assessed in terms of the typicality and similarity of the speaker-dependent features they contain. In forensic phonetics, speaker-specificity is defined as the ratio of between- to within-speaker variation. The present work contributes to both fields by checking whether previously reported linguistic effects for fricatives are present in spontaneous telephone dialogues, which is a relevant speech style and channel both for everyday communication and forensic speaker comparisons, and whether these linguistic effects interact with the

amount of speaker information for two highly frequent fricatives in Dutch.

2.1.1 Within-speaker variability in fricative production

2.1.1.1 Labialization

Within speakers, it has been shown that fricative acoustics vary systematically as a function of phonetic context. Predominantly, anticipatory lip-rounding has repeatedly been shown to lower resonance frequencies in fricatives (e.g., Bell-Berti & Harris, 1979; Koenig et al., 2013). Anticipatory lip-rounding lowers the resonance frequencies in fricatives because the lip protrusion associated with the lip movement lengthens the anterior cavity. Notably, neighboring labial consonants such as English bilabial /w/ and /p/ also seem to display a lowering effect on /s/ spectra (Munson, 2004), even though the lip movement for /p/ is better described as lip closure rather than lip-rounding. This implies that labial closure also lengthens the anterior cavity to some extent.

Regarding within-speaker variation, Munson (2004) hypothesized that variability in degree and timing of the labial co-articulation in /s/ would result in increased within-speaker variation. Replicating earlier research, Munson (2004) reported that /s/ has lower resonance frequencies when followed by rounded /u/ versus non-rounded /a/ and when followed by rounded /w/ versus vowels /a, u/, with labial – but not rounded – /p/ falling in-between. The results for the within-speaker variation, however, only showed increased within-speaker variation for /s/ followed by /w/ and not for /s/ followed by /u/ compared to when it is followed by /a/. It is probable that the lip-movements for /w/ versus /u/ and /p/ constitute different labial movements. Other work has shown that there are different types of labialization, e.g., different lip-area size involved in labialization for postalveolar fricatives /ʃ, ʒ/ versus approximant /w/ (Toda et al., 2003). It is therefore possible that the labial movement for /w/ is more sensitive to within-speaker variation than the

labial movements for /u/ and /p/. Alternatively, /s/ followed by /w/ may display more within-speaker variation due to differences in articulatory timing between /s/ from consonant clusters versus consonant-vowel sequences. Munson (2004) did not report on the between-speaker variation, therefore, no information on the speaker-specificity of fricatives in labialized context is available. Given that the degree and timing of labial co-articulation in fricatives might vary between speakers (Perkell & Matthies, 1992), fricatives with labialized context might also constitute relatively speaker-specific locations.

2.1.1.2 Speech effort

Articulatory strengthening (hyperarticulation) or weakening (hypoarticulation) also affect fricative acoustics within speakers. Generally speaking, it has been shown that there are articulatory strong and weak locations in speech. Whereas the initial edges of prosodic domains such as phrases and words are generally found to be locations of articulatory strengthening (Cho & McQueen, 2005; Fougeron, 2001), the final edges of syllables, i.e., codas, are generally found to be locations of articulatory weakening compared to syllable onsets (Ohala & Kawasaki, 1984). For fricatives as a group, American English coda fricatives are found to be less identifiable (Redford & Diehl, 1999), and to have a lower intensity and a delayed and lower air pressure peak than onset fricatives (Solé, 2003). However, studies that consider different fricatives separately show inconsistent results with regards to coda reduction for /s/ specifically; Redford & Diehl (1999) found coda reduction in duration in American English /s/, but not in intensity or spectral mean. Furthermore, they reported that, whereas consonant classification using linear discriminant analysis overall showed more accurately classified onsets than codas, this was not the case for /s/, where there was a reverse tendency. This lack of coda reduction for /s/ was replicated for German, where spectral mean for codas was not lower, but slightly higher than for onsets (Cunha & Reubold, 2015). Although there was no reduction effect for German /s/ in coda position, Cunha & Reubold (2015) found that codas display higher variability than onsets and that /s/ in de-accented syllables displays higher variability than /s/ in

accented syllables. In other words, they reported more variability, but no reduction, in articulatory weak locations. Overall, reports on reduction in fricative acoustics are inconsistent, particularly with regards to /s/, but studies generally report more variability for articulatory weak positions. It is unclear whether that variability is within- or between-speakers.

2.1.1.3 Segmental effects

From the somewhat conflicting results reported above, it seems that not all fricatives reduce in the same manner or to the same extent. Rather, reduction seems to be constrained by specific production requirements (Recasens, 2004). This means that features that have high production requirements for a particular speech sound are more resistant to co-articulation and reduction than features that have low production requirements for a particular speech sound. For example, in fricatives /s/ and /x/, the resistance to anticipatory labialization might be low because there are no production requirements for the lips in /s/ and /x/. Tongue front and dorsum in the production of /s/, on the other hand, are relatively resistant to co-articulation and reduction due to the production necessity of tongue front raising and dorsum lowering for this fricative (Recasens & Dolorspallarè, 2001). Speakers might vary in their articulatory timing, degree of co-articulation, and their reduction of specific features. This means that some speakers may be more sensitive to certain co-articulatory effects than others. As a result, the acoustic realizations of /s/ and /x/ might be more context-dependent in some speakers than others. It is therefore possible that highly context-dependent realizations, such as /s/ and /x/ in labialized context, display high between-speaker variability.

2.1.1.4 Other linguistic effects

Speech style can also affect fricative acoustics within speakers. Maniwa et al. (2009) compared clearly spoken fricatives to fricatives in a conversational speech style in American English and found that clearly spoken fricatives had longer duration, higher resonance frequencies, and – surprisingly – lower relative amplitude. Moreover, individual speakers

used different strategies for producing clear speech, which were not related to speaker sex/gender. This implies that different patterns of within- and between-speaker variation may be expected in clearly spoken speech versus conversational speech. It therefore seems important to extend research on speaker variation to include conversational speech styles.

2.1.2 Between-speaker variability in fricative production

Between speakers, anatomical/physiological and social effects have been observed in fricative acoustics. Regarding anatomical/physiological variation, fricative acoustics can vary as a function of the shapes and sizes of the articulators and cavities (Stevens, 2000, pp. 411–412). In practice, this type of variation in fricative acoustics has often been observed between males and females; fricatives produced by females have higher resonance frequencies than by males, which is often explained as resulting from anatomical differences between female and male speakers (e.g., Jongman et al., 2000; Schwartz, 1968). This difference in production is perceivable and meaningful to listeners, as speaker sex can be perceived from isolated voiceless fricatives (Ingemann, 1968; Schwartz, 1968).

From sociolinguistics, there are known between-speaker factors that affect fricative acoustics. For example, there are well-attested effects of gender identity and sexual orientation on /s/ spectra that are not associated with anatomical/physiological differences but rather with production strategies, i.e., learned behavior (e.g., Bang et al., 2017; Fuchs & Toda, 2010; Munson et al., 2006). Social class may also affect fricative spectra; Stuart-Smith (2007) found that English working-class females could be grouped with working-class males, rather than with higher-class females, on several spectral features from /s/. When looking at social identity on a larger scale, such as ethnolect, dialect, and language communities, variation in fricative spectra is also observed. For example, the so-called 'Moroccan flavored Dutch' ethnolect is known for a retracted [s] realization that resembles [ʃ], i.e., sibilant palatalization,

in certain phonetic contexts (Mourigh, 2017). Another example is the regional variation for Dutch fricative /x/, which is produced with velar place of articulation (and thus higher resonance frequencies) in Flanders and Southern regions of the Netherlands, and with uvular place of articulation – often accompanied by uvular scrape, i.e., uvular trill – in Northern regions of the Netherlands (Van der Harst et al., 2007).

Given that group-level speaker characteristics such as sex/gender and ethnolect are associated with shared acoustic features, it seems important to eliminate as much group-level variation as possible when focused on characterizing individual speakers. Moreover, in forensic casework, it is deemed necessary to compare speakers amongst a reference population of similar speakers, i.e., speakers of the same sex/gender and dialect. This work therefore chose to limit itself to speakers from the same sex/gender and dialect.

2.1.3 Speaker-specificity and linguistic context

It is currently unclear how speaker-specificity is dependent on linguistic context. Given that speaker-specificity is a ratio of between-speaker to within-speaker variation, speech samples need high between-speaker variation and low within-speaker variation to be speaker-specific. There are some linguistic contexts that might facilitate such environments, and thus help listeners extract speaker information.

2.1.3.1 Segmental effects on speaker-specificity

Previous work has shown that some individual speech sounds are more speaker-specific than others. For example, vowels are found to be more speaker-specific than consonants (Van den Heuvel, 1996, pp. 145-146). Within the class of consonants, fricative /s/ – one of the speech sounds investigated in the present work – is found to be relatively speaker-specific. In Dutch read speech, /s/ was ranked below vowels and nasals, but above /r/ and plosives in terms of speaker-specificity (Van den Heuvel, 1996, pp. 72). In English read speech, /s/ along with nasal /m/

are ranked above nasals /n/ and /ŋ/, and liquid /l/ (Kavanagh, 2012, pp. 387-388). Studies on the speaker-specificity of fricatives that are not /s/ – such as the dorsal fricative /x/ also examined in the present work – are rare. Perceptually, differences in the amount of speaker-dependent information have also been observed. Comparing speaker sex identification between fricative sounds, Ingemann (1968) found that listeners can identify speaker sex from isolated back fricatives [h, χ, x] but not from isolated front fricatives [θ, f, ɸ]. Front fricatives [s, ʃ] broke this pattern; speaker sex identification from these sounds was also above chance.

2.1.3.2 Speech effort and speaker-specificity

Articulatory strong locations are locations in speech that are produced with more vocal effort, e.g., onsets and stressed syllables. They are often argued to constitute canonical speech, and might therefore be characterized by low within-speaker variation. If these locations are not also characterized by low between-speaker variation, they might be relatively speaker-specific. Evidence supporting this hypothesis comes from a finding that speakers were characterized more accurately using vowels receiving sentence stress – which are generally considered to be articulatory strong locations – than vowels without sentence stress (McDougall, 2004). Other evidence that suggests that articulatory strong locations contain more speaker-dependent information can be found in Heeren (2018), who showed that the vowel /a/ sampled from spontaneous speech gave higher speaker classification scores in content than in function words. Content words are generally also found to be articulatory strong locations, which is evidenced by studies that found reduction in vowels sampled from function words relative to content words (Shi et al., 2005; Van Bergem, 1993, pp. 38–39).

Alternatively to articulatory strong locations displaying high speaker-specificity, articulatory weak locations such as codas and highly context-dependent segments, e.g., fricatives with labial neighbors, might be characterized by high between-speaker variation and may therefore also display high speaker-specificity. Based on their work on formant and intensity dynamics, He et al. (2017; 2019) hypothesize that speakers may

have more articulatory freedom in speech locations that are less constrained by articulatory targets, resulting in higher between-speaker variation in these locations. This is sometimes also referred to as variation due to target undershoot. They showed that both intensity dynamics (He & Dellwo, 2017) and formant dynamics (He et al., 2019) show more between-speaker variation in negative than in positive dynamics. Negative dynamics were defined as the intensity and formant slopes from the syllable's peak to the following trough, which are the parts of syllables associated with mouth-closing gestures. They suggest that the mouth-opening gestures (positive dynamics) might be more restricted by articulatory targets.

Previous studies thus indicate that some linguistic contexts affect the amount of within- and between-speaker variation. Namely, articulatory strong locations seem to have relatively low within-speaker variation and articulatory weak locations seem to have relatively high between-speaker variation. However, for fricatives, it is unclear how articulatory weak versus strong positions affect the speaker-specificity.

2.1.4 Fricatives in Dutch telephone speech

2.1.4.1 Dutch fricatives

The Standard Dutch fricative inventory contains eight fricatives (see Table 2.1). The present study focuses on two voiceless fricatives: the laminal alveolar /s/ and the dorsal fricative /x/ (for notation sake, the dorsal fricative – which can have a velar [x] or uvular [χ] place of articulation, will be denoted with symbol 'x'). Fricatives /s/ and /x/ were selected because they are highly frequent in syllable onsets and to a slightly lesser extent in coda position in Dutch (Baayen et al., 1993), which makes them suitable speech sounds to analyze in spontaneous speech samples.

Table 2.1: *Standard Dutch fricative inventory (cf. Gussenhoven, 1999). Fricatives in parentheses are restricted to loanwords and to alveolar fricatives with place assimilation from a following* [j] *(e.g., jas 'coat'* [jɑs]; *jasje 'little coat'* [jɑʃə]*).*

	Voiceless	Voiced
Labiodental	f	v
Alveolar	s	z
Post-alveolar	(ʃ)	(ʒ)
Dorsal	x/χ	
Glottal		ɦ

Fricative sounds are produced with a narrow constriction which results in noise generated by turbulence (Stevens, 2000, p. 379). The resonance frequencies of fricatives are mainly determined by the size of the cavity anterior to the narrow constriction (Stevens, 2000, pp. 398-403). Whereas the Dutch laminal alveolar fricative /s/ has a relatively small anterior cavity and therefore high resonance frequencies, Dutch dorsal fricative /x/ has a medium to large anterior cavity (depending on a velar or uvular place of articulation) and therefore much lower resonance frequencies. Fricative /s/ is reported to have a spectral center of gravity of around 4.8 kHz in Standard Dutch read speech (Ditewig et al., 2019) and fricative /x/ is reported to have a spectral peak of around 1.7 kHz in Standard Dutch read speech (Van der Harst et al., 2007).

2.1.4.2 Telephone filter

Most acoustic reports on /s/ and /x/ are based on studio-recorded read speech. However, this speech style is not representative of everyday communication nor of forensic speaker comparisons. It is unclear

whether acoustic-phonetic and indexical information in /s/ and /x/ can be captured in spontaneous telephone dialogues. Particularly in the context of forensic speech comparisons, telephone speech is highly relevant compared to studio-recorded (read) speech, as wiretapping telephone conversations from criminal suspects is common in police investigations in the Netherlands (Odinot et al., 2010, p. 82). Using higher-quality, non-telephone speech may misrepresent what listeners may use in speech perception in daily conversation as well as what is possible for forensic speaker comparisons.

Telephone signals have a limited frequency bandwidth. For example, the landline telephone dialogues worked with in this study have a bandwidth of 340 - 3400 Hz. Given that the spectral energy for Dutch /s/ is concentrated around 4.8 kHz (Ditewig et al., 2019), this means that the spectral energy for fricative /s/ mostly resides above the upper limit of this bandwidth (see Figure 2.1a). It is therefore possible that both linguistic information and speaker information from /s/ are (partly) lost in telephone speech. The spectral energy for back fricative /x/, on the other hand, falls mostly within the telephone bandwidth (see Figure 2.1b).

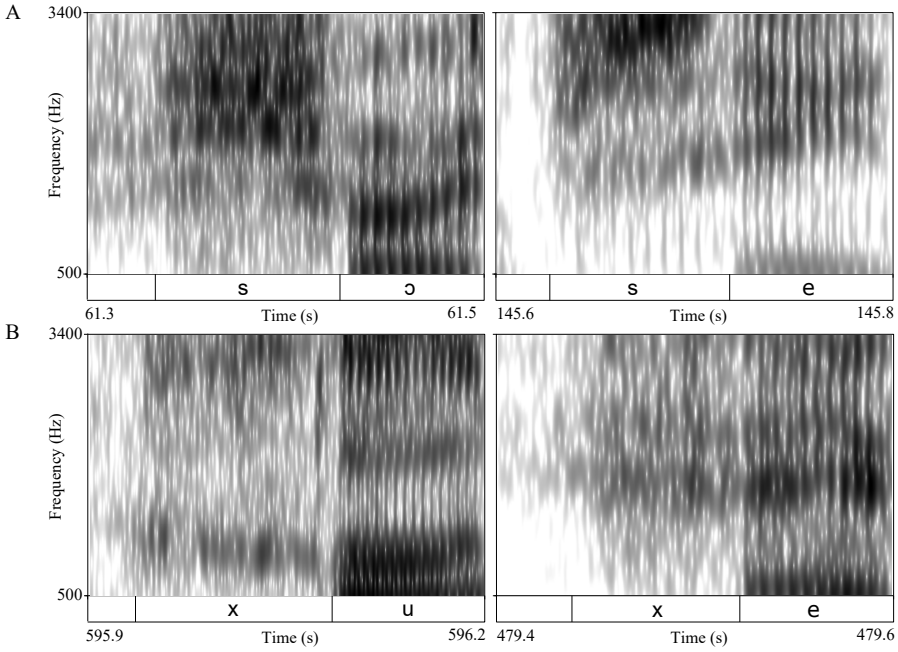

Figure 2.1: *Spectrograms for onset fricatives in labial and non-labial contexts spoken by a male speaker of Standard Dutch over a 500-3400 Hz bandwidth. A: Onset /s/ from words* soort *('sort',* /sɔːrt/) *and* cd *('cd',* /sede/*). B: Onset /x/ from words* goed *('good',* /xut/) *and* geen *('no',* /xen/*).*

Telephone speech also has other limitations that have to be considered in an acoustic analysis. Regarding signal-related transformative qualities, the lower formants may display an upward shift. Particularly F1 values display a large shift of around 14% on average, whereas higher formants generally remain unaffected (Künzel, 2001; for mobile signals, this number is 29% on average, with some F1 values rising by up to 60%: Byrne & Foulkes, 2004). Moreover, when this signal-related shift is paired with speaker-behavior such as holding the phone between the cheek and shoulder, these upwards shifts are amplified

(Jovičić et al., 2015). Additionally, the signal-related qualities of telephone speech are accompanied by distinct speech behavior. For example, speakers often increase their vocal effort, possibly to adjust for increased background noises from variable environments. This effect is generally described as the Lombard effect (e.g., Junqua, 1993).

2.1.5 Research questions and hypotheses

The main research question of the present study is whether the amount of speaker information in speech sounds is dependent on their linguistic context. Two fricatives were selected based on their frequency of occurrence in Dutch: alveolar /s/ and dorsal /x/. These fricatives were sampled from spontaneous telephone dialogues, which are representative of everyday communication as well as forensic voice comparisons. To answer the main research question, we first tested whether linguistic context factors (syllabic position, perseverative and anticipatory labialization) – which have been shown to affect fricative acoustics in read broadband speech – can be replicated in spontaneous telephone speech. Secondly, we examined whether speaker-classification models for the two fricatives show effects of linguistic context. In this second step, the effect of the speech sound (/s/ versus /x/) and the contribution of individual acoustic measurements on speaker-classification were also examined.

2.1.5.1 Linguistic effects

Based on previous research on read broadband speech (e.g., Bell-Berti & Harris, 1979; Koenig et al., 2013), we hypothesized that perseverative and anticipatory labialization would lower fricative spectra, but that this might not be measurable for /s/ because the spectrally-defining characteristics for /s/ mostly reside over the upper limit of the telephone bandwidth. Spectrally-defining characteristics for dorsal fricative /x/, on the other hand, should fall within the telephone bandwidth. The literature is not clear on the effect of syllabic position, particularly for /s/.

2.1.5.2 Speaker classification

In the second step, we hypothesized that there might be a segmental effect in speaker classification. Namely, /x/ might be more speaker-specific than /s/, because the telephone channel cuts off most spectral energy for /s/ but not /x/. Regarding the performance of acoustic measures, previous findings report that spectral center of gravity and standard deviation were the most speaker-discriminating features (e.g., Kavanagh, 2012). We therefore predicted that most speaker-specific information might be found in spectral as opposed to temporal or amplitudinal measures. Regarding the speaker variation as a function of linguistic context, we hypothesized that articulatory strong locations (onsets and fricatives with non-labial neighbors) are characterized by low within-speaker variation and that articulatory weak locations (codas and fricatives with labial neighbors) are characterized by high between-speaker variation. However, there were no clear expectations for speaker-specificity, which equals the ratio of between- to within-speaker variation.

2.2 Methodology

2.2.1 Materials

Spontaneous telephone dialogues available in the Spoken Dutch Corpus (Oostdijk, 2000) were used to investigate the speaker-specificity in the realization of fricatives /s/ and /x/. The telephone dialogues were obtained via a switchboard. No information on the task is available, but from the recordings' content it was inferred that speakers were located in their home environment (deduced from background noises such as a crying baby or a barking dog) and were asked to converse for around ten minutes on any topic of their choosing. One to four telephone conversations (M = 1.88, SD = 0.96) – with different interlocutors – are available for each speaker in the corpus. All available conversations for a speaker were included.

Given the overrepresentation of male speakers in forensic voice comparisons, only male speakers were analyzed in this study[1]. Male speakers were included if the metadata from the corpus indicated that they were between 18 and 50 years old and if they were reported to be speakers of Standard Dutch. Speakers were excluded if the first author judged them to be speakers of non-standard Dutch. For the remaining 66 male speakers of Standard Dutch (age range = 21 - 50, M = 36.5, SD = 7.3), a total of 3,331 /s/ tokens and their adjacent contexts as well as 3,491 /x/ tokens with their adjacent contexts were first automatically segmented and provided with a broad phonetic transcription using the orthographic transcript available with the corpus. These were then manually validated by the first author. When interference such as laughter, overlapping speech from the interlocutor, or background noise showed up in the signal, tokens were excluded. Fricative tokens occurring in context with a creaky phonation were not excluded, as previous research has shown that /s/ spectra are relatively stable against creakiness (Hirson & Duckworth, 1993). Tokens were labelled as onsets (/s/: N = 1,359; /x/: N = 1,657), codas (/s/: N = 1,532; /x/: N = 1,453), or ambisyllabic (/s/: N = 440; /x/: N = 380). The latter category, containing tokens that cannot be categorized as either onsets or codas (e.g., was ook 'was also' [wɑsoːk]), was excluded from analysis.

As reviewed above, labialization of adjacent context affects fricative spectra. To test whether the measures extracted from telephone speech are sensitive to contextual labialization, preceding and following context was furthermore labelled as labial or non-labial. Rounded vowels /u, ɔ, o, ø, y, ʏ/, (partially) rounded diphthongs /œy, ɑu/ (cf. temporal patterns of lip-rounding: Bell-Berti & Harris, 1982), and bilabial consonants /p, b, m/, were considered to be labial. Labiodental consonants /f, v, ʋ/ were not coded as labial because the teeth-to-lip movement in these sounds does not involve lip-rounding or closure, but rather eliminates the anterior cavity and can therefore not be assumed to

[1] It is unclear from the metadata from the Spoken Dutch Corpus how the label 'male' was assigned to speakers. It is assumed here that 'male' refers to biological sex.

have the same lowering effect on the spectrum. Speakers with fewer than 25 tokens per fricative sound were excluded, which excluded 23 speakers and left a total of 43 speakers with a sufficient number of tokens for both /s/ and /x/. The resulting numbers of tokens per factor level are presented in Table 2.2.

Table 2.2: *Totals, and means, standard deviations, and ranges for numbers of /s/ and /x/ tokens by speaker (N = 43) and by linguistic context factor level.*

			Syllabic Position		Left Context		Right context	
		Total	Onset	Coda	Non-labial	Labial	Non-labial	Labial
/s/	Total	2,346	1,066	1,280	1,846	500	1,903	443
	M	55	25	30	43	12	44	10
	SD	19	11	11	16	5	15	7
	range	25-108	9-63	15-78	20-88	3-22	24-88	1-35
/x/	Total	2,820	1,460	1,360	2,336	484	2,250	570
	M	66	34	32	54	11	52	13
	SD	26	13	15	23	6	22	7
	range	27-124	11-67	9-73	20-106	3-29	23-100	3-31

2.2.2 Acoustic analysis

The telephone dialogues available in the Spoken Dutch Corpus have a sampling frequency of 8 kHz with an 8-bit resolution and were originally filtered at a bandwidth of 340 – 3,400 Hz. There are separate channels

for the two speakers in each telephone conversation. A low-frequency cut-off of 500 Hz was used to reduce the influence of background noise and (partial) voicing. For each fricative token, seven measures were taken in Praat version 6.0.46 (Boersma & Weenink, 2020). First, duration (DUR; in milliseconds, ms) was computed from fricative onset to fricative offset as characterized by the presence of aperiodic fricative noise, which was then used to establish the middle 50% of each fricative over which the static spectral measures were taken. The static spectral measures consisted of two spectral moments – spectral center of gravity (CoG) and standard deviation (SD) – and spectral tilt. After filtering the fricative tokens to the 0.5 - 3.4 kHz band (band pass Hann filter, smoothing = 100 Hz), the center of gravity and the standard deviation (CoG and SD; in Hertz, Hz) were computed from the spectrum determined over the mid-50% of the fricative, using power spectrum weighting. Although the formant-like structure of spectral energy for Dutch /x/ (see Figure 2.1b) might be captured better by more complex measures such as discrete cosine transforms (DCT), the relatively simple measure CoG has been shown to capture between-speaker variation such as regional variation (Harst et al., 2007)[2].

Spectral tilt (TILT) was measured to reflect vocal effort as an alternative to absolute amplitudinal measures, and computed from the long-term average spectrum determined over the mid-50% of the fricative (bin = 1 Hz) on a logarithmic frequency scale (dB/decade), using a least-squares fit. A decade is a step on the frequency scale with the power of 10, i.e., 1 Hz, 10 Hz, 100 Hz, etc. Mean amplitude (AMP; in dB) was measured over the full fricative's duration and normalized by speaker through Z-transformation.

Additional to the static measures, dynamic spectral measures were computed by measuring spectral CoG in non-overlapping 20%-portions of the entire fricative's duration. Coefficients from quadratic

[2] To pilot our data and acoustic measures, all /x/ tokens were auditorily labelled on place of articulation (velar versus uvular) and CoG was shown to predict place of articulation with a cross-validated accuracy of 83.9% in a linear-discriminant analysis (LDA). We therefore expect CoG to adequately capture the linguistic effects and speaker-dependent spectral characteristics in fricative acoustics.

polynomial equations over the five resulting data points per fricative token constituted our dynamic measures for analysis. Both cubic and quadratic models to the data were estimated; likelihood-ratio tests showed no significant difference between these two models (/s/: χ^2 (1) = 0.96, p = .33; /x/: χ^2 (1) = 0.11, p = .74). The simpler quadratic function ($\gamma = \beta_0 + \beta_1 x + \beta_2 x^2$) was chosen as the fewer coefficients reduced the number of predictors in further modelling. The intercept (β_0) was excluded because it correlated highly with the static CoG measure (/s/: r = .95, N = 2,346, p < .001; /x/: r = .96, N = 2,820, p < .001), resulting in only a linear (CoG^{linear}) and quadratic ($CoG^{quadratic}$) coefficient.

2.2.3 Statistical analysis

The statistical analysis consisted of two parts: (1) linear mixed-effect modelling was used to check whether linguistic factors affected /s/ and /x/ acoustics in spontaneous telephone speech, and (2) multinomial logistic regression was used to investigate whether the amount of speaker information in /s/ and /x/ varied as a function of syllabic position and labial co-articulation. Additionally, segmental effects as well as the relative importance of acoustic measures in speaker classification were estimated from the regression model. A more traditional measure for speaker-specificity, called the Speaker-Specificity Index (SSI), was also computed for all acoustic variables to assess its relationship with the regression modelling results. The SSI relates the between-speaker variance to the within-speaker variance (Van den Heuvel, 1996).

2.2.3.1 Linear mixed-effect modelling: Linguistic effects

In the first part of the analysis, the effects of linguistic context factors on acoustic measures were investigated by means of linear mixed-effect modelling (LMM) in R version 3.5.1. (R Core Team, 2018). First, a model with maximal fixed and random structure was built for each dependent variable, i.e., each acoustic measure (CoG, SD, TILT, DUR, and AMP). This maximal model contained six fixed factors: three main factors for Syllabic Position (CODA, ONSET; sum coded), Left Context (NON-LABIAL, LABIAL; dummy coded), and Right Context (NON-

LABIAL, LABIAL; dummy coded) and three one-way interactions between these main factors. One-way interaction terms were included because Right Context for factor level CODA contained only consonants and pauses coded for labialization (see section 2.2.1). Because labial consonants possibly produce attenuated coarticulation effects on neighboring fricatives compared to labial vowels (Munson, 2004), an interaction between the Left and Right Context factors and Syllabic Position might be expected. The random structure of the maximal model contained random intercepts for Word and Speaker, as well as random slopes by Speaker over all three fixed factors. This means that Syllabic Position and Left and Right Context were added to the model as both within-speaker and between-speaker factors.

All fixed and random terms in the maximal model were tested via model comparisons. First, a full model with maximal random structure was built by restricted maximum likelihood (REML) estimation (Barr et al., 2013). Next, stepwise deletion was used to reduce the random structure of the model, given this led to a better-fitting model as estimated by the Bayesian information criterion (Bates et al., 2015). Model fit was assessed through inspection of the residuals and duration was log-transformed (base = 10) for a better model fit. The *p*-values were generated empirically with bootstrapping using function *mixed()* from R package 'afex' (Singmann, 2019). This function derives a mean *p*-value for a fixed effect by comparing the optimal model with a model without the fixed effect in question for a specified number of data simulations (N = 10,000). The significance level (α = .05) of fixed effects was adjusted via Bonferroni correction (α = .05/(5*2)), to account for the fact that the different acoustic measures (N = 5) and fricative sounds (N = 2) were extracted from the same dataset of speakers.

Lastly, the results were tested in the presence of two prosodic factors that would possibly confound results obtained by previous modelling. Models were rebuilt including factors for Phrasal Position (INITIAL, MEDIAL, FINAL; sum coded) and Word Stress (NON-STRESSED, STRESSED; sum coded) to see if results were maintained. For Word Stress, only tokens from content words (nouns, verbs, adjectives, and adverbs) were labelled for word stress, as function words can have stressed syllables only in special circumstances (Selkirk, 1996). This

resulted in the exclusion of 16% of the data for /s/ and 12% of the data for /x/. Results from these latter models are not presented because these extended models did not change the results obtained by earlier models, although exact statistics were slightly different.

2.2.3.2 Multinomial logistic regression: Speaker classification

Multinomial logistic regression (MLR) was used to test which linguistic context factors and acoustic measures significantly predicted the dependent variable Speaker. Function *buildmultinom()* from R Package 'buildmer' (Voeten, 2020) was used to automatically build and then reduce the maximal MLR model by estimating each predictor with backward stepwise selection using likelihood-ratio tests. Highly correlating predictors ($r > .70$) were excluded, which resulted in the exclusion of TILT because it correlated highly with CoG (/s/: $r = .76$, $N = 2{,}346$, $p < .001$; /x/: $r = .91$, $N = 2{,}820$, $p < .001$). This means the maximal MLR model to predict SPEAKER contained 27 predictors: six acoustic measures (CoG, SD, AMP, DUR, CoG^{linear}, and $CoG^{quadratic}$), three linguistic factors (Syllabic Position, Left Context, and Right Context), and 18 one-way interactions between the acoustic measures and linguistic factors.

In a second step, the optimal model obtained by function *buildmultinom()* was inspected to see which fricative contained more speaker-dependent information and which combinations of acoustic measures and linguistic context factors affect speaker classification predictions. The predicted speaker classification of factor levels was compared, i.e., for Syllabic Position, speaker classification of codas is compared to onsets. This was achieved by splitting the data on factor level and then predicting speaker classification on the resulting two datasets using the best-fitting model acquired in the previous step. This was done for factor levels from all linguistic context factors that were included in the best-fitting models. Secondly, acoustic measures and their interactions with linguistic context factors were excluded from the best-fitting model one at a time to assess the relative importance of each acoustic measure.

2.3 Results

2.3.1 Linguistic effects

2.3.1.1 Labialization

Linear mixed-effect modelling results for /s/ and /x/ are summarized in Table 2.3. For /s/, there were no effects for Left Context. However, /s/ tokens with labial Right Context have lower SD, shorter duration in codas, and – opposite to what we hypothesized – higher CoG. Contrary to results for /s/, there is a clear labialization effect for /x/. When Left Context is labial, /x/ CoG lowers and spectral tilt decreases, i.e., there is less energy at higher frequencies. When Right Context is labial, CoG lowers (although this effect is larger for onsets), spectral tilt decreases, and amplitude decreases. The interaction between Left and Right Context for spectral tilt indicates that spectral lowering is attenuated by 4.7 dB per decade when both Left and Right Context are labial.

2.3.1.2 Syllabic position

Results in Table 2.3 show that /s/ onsets have higher CoG, higher positive tilt, i.e., more high-frequency energy, higher amplitude and shorter duration than codas. In other words, all measures from /s/ except duration show coda reduction. Note also that the spectral tilt intercept in Table 2.3 is a positive value, i.e., there is no energy drop-off but an increase in higher frequencies. This is expected for /s/ because, within the telephone band, all the spectral energy is expected to reside in the higher frequencies.

Fricative /x/ also showed coda reduction; onsets have higher amplitude than codas. Contrasting our data for /s/, tilt for /x/ shows a negative value. This shows that, whereas there is no energy drop-off for high-frequency /s/, there is an average energy drop-off of 7.8 dB per decade for /x/.

Table 2.3. *Summary of fixed effects from linear mixed-effect modelling for /s/ (N = 2,346) and /x/ (N = 2,820) with Kenward-Roger degrees of freedom approximation. Reference values are CODA for Syllabic Position and NON-LABIAL for Left and Right Context. Empty cells indicate that the factor was not included in the best-fitting model. The p-values for fixed effects were obtained empirically by bootstrapping (N simulations = 10,000). Non-significant effects are in italic.*

		/s/			/x/		
	Fixed effects	*Est.*	*SE*	*t*	*Est.*	*SE*	*t*
CoG	(intercept)	2541	37	68.2	1648	34	48.6
	SyllPos: ONSET				-5	13	*-0.4*
	Left Context: LABIAL	-15	28	*-0.5*	-192	25	-7.8
	Right Context: LABIAL	86	19	4.6	-281	39	-7.3
	SyllPos × Right Context				-103	30	-3.5
SD	(intercept)	603	18	32.7	599	14	43.0
	SyllPos: ONSET				-6	4	*-1.6*
	Left Context: LABIAL				27	9	3.0
	Right Context: LABIAL	-42	9	-4.7	-7	20	*-0.4*
	SyllPos × Left Context				-54	9	-6.1
	SyllPos × Right Context				-57	9	-6.3
TILT	(intercept)	17.3	1.5	11.8	-7.8	1.3	-6.2
	SyllPos: ONSET				-0.6	0.4	*-1.4*
	Left Context: LABIAL				-7.4	1.1	-6.6
	Right Context: LABIAL	2.1	0.5	4.6	-8.6	1.3	-6.7
	SyllPos × Right Context				-3.9	1.0	-4.0
	Left × Right Context				4.7	1.8	2.5
AMP	(intercept)	0.04	0.03	*1.5*	0.01	0.03	*0.3*
	SyllPos: ONSET	0.15	0.03	5.5	0.24	0.03	8.1
	Right Context: LABIAL				-0.26	0.07	-3.5
DUR	(intercept)	1.95	0.01	235	1.92	0.01	212.8
	SyllPos: ONSET	-0.03	0.01	-5.0	0.01	0.01	*1.1*
	Right Context: LABIAL	-0.07	0.01	-6.2	-0.02	0.01	*-1.2*
	SyllPos × Right Context	0.06	0.01	5.2	0.09	0.01	6.5

2.3.1.3 Intermediate discussion

Linear mixed-effect modelling has indicated that both /s/ and /x/ are affected by our fixed factors for several measures, but not in the same way. Figure 2.2 illustrates the differences in effects of context labialization on CoG between the two fricatives under study. Whereas /x/ CoG lowers when context is labial, this is clearly not the case for /s/. As hypothesized, this may be due to the telephone bandwidth. If the speaker-specificity is sensitive to linguistic context factors, the acoustic results would predict stronger effects for /x/ than for /s/ in the speaker-classification analysis, since /x/ shows more context-dependent acoustic variation than /s/.

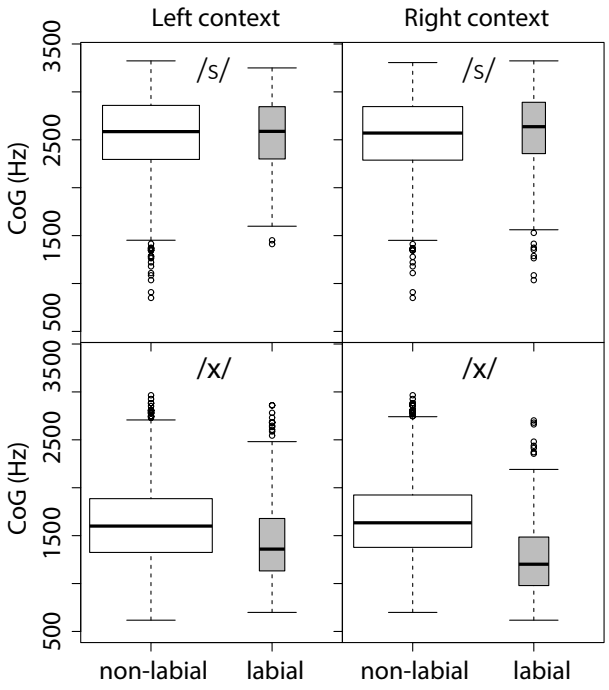

Figure 2.2: *Boxplots for CoG by fricative sound, Syllabic Position, and Left and Right Context labialization. The width of the box represents the number of cases included in the MLE and MLR analyses (see Table 2.2 for exact numbers).*

2.3.2 Speaker classification

2.3.2.1 Segmental effects

For both /s/ and /x/, the best-fitting model to predict Speaker (/s/: $N = 43$, $n = 2{,}346$; /x/: $N = 43$, $n = 2{,}820$) included all acoustic measures and all linguistic context factors as significant predictors. The interactions that were included as predictors are indicated in Table 2.4. The model for /s/ had a speaker-classification accuracy of 19.5% against a chance level of 2.3%. The model for /x/ had a speaker-classification accuracy of 18.4% (chance = 2.3%). This means that, despite the limited telephone band, speaker classification from fricative /s/ acoustics was better than from fricative /x/ acoustics.

Table 2.4: *Included one-way interactions in the optimal MLR models for /s/ and /x/.*

Predictor	Syllabic position		Left context		Right context	
	/s/	/x/	/s/	/x/	/s/	/x/
CoG	✓	✓	✓	✓	✓	✓
SD		✓	✓	✓		✓
CoG^{linear}	✓	✓			✓	✓
$CoG^{quadratic}$	✓				✓	✓
AMP	✓	✓			✓	✓
DUR	✓	✓	✓		✓	✓

2.3.2.2 Contribution of individual acoustic measures

The decreases in speaker-classification accuracy when a single acoustic measure and its interactions with linguistic context factors were dropped from the model are presented in Table 2.5. For example, excluding CoG and the interactions between CoG and linguistic context factors from the best-fitting model for /s/ resulted in a decrease in speaker-classification accuracy from 19.5% (for the optimal model) to 13.9%, which makes a decrease of 5.6%. As can be seen in Table 2.5, CoG and SD were relatively important measures for speaker classification. Moreover, measures contributed to speaker classification in comparable ways across fricatives. The contribution of acoustic measures to the speaker classification from the MLR model is accompanied by the more traditional SSI measure; these more or less mirror the relative ranking from the MLR model.

Table 2.5: *Speaker-classification accuracy decreases (in %) per acoustic measure relative to the full models' speaker-classification accuracy of 19.5% for /s/ and 18.4% for /x/ and speaker-specificity index (SSI) per acoustic measure for /s/ and /x/.*

	/s/		/x/	
Excluded measure	*Δacc*	*SSI*	*Δacc*	*SSI*
CoG	5.6	0.56	4.5	0.26
SD	4.5	0.63	3.4	0.31
DUR	1.9	0.07	2.1	0.10
CoG^{linear}	0.9	0.07	1.6	0.06
$CoG^{quadratic}$	1.3	0.08	1.2	0.07
AMP	1.1	0.14	0.7	0.06

2.3.2.3 Linguistic effects

Per linguistic context, speaker-classification accuracies were similar (see Table 2.6), but there seems to be a small, yet systematic, advantage for articulatory weak locations, i.e., codas and tokens with labial co-articulation.

Table 2.6: *Speaker classification accuracies (in %) per fricative sound and per linguistic context factor level (chance level = 2.3%).*

	Linguistic context	/s/	/x/
Syllabic Position	Total	19.5	18.4
	Onset	19.5	18.2
	Coda	19.5	18.6
Left Context	Non-labial	18.3	18.5
	Labial	24.2	18.8
Right Context	Non-labial	18.5	17.6
	Labial	18.8	21.4

The small advantage in speaker classification for articulatory weak locations was examined to see whether it was due to an increase in between-speaker variation. The between- and within-speaker variances per linguistic context factors are presented for the most-contributing measure in speaker-classification for /x/, i.e., CoG (see Figure 2.3). Consistent with the SSIs reported in Table 2.5, Figure 2.3 shows that the within-speaker variance is consistently higher than the between-speaker variance across all linguistic contexts. Additionally, as hypothesized, the between-speaker variance seems to be increased in articulatory weak locations compared to strong locations. Against expectation, the within-speaker variation seems to be decreased in articulatory weak locations.

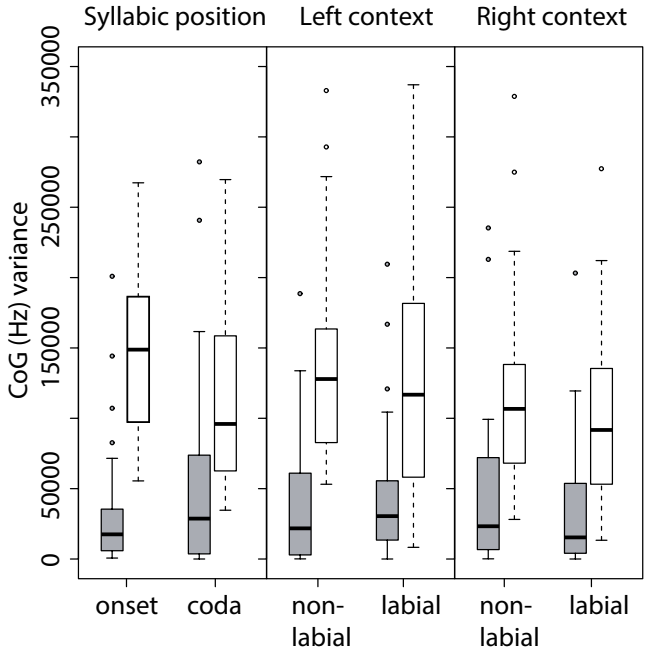

Figure. 2.3: *Boxplots of between- (grey bars) and within-speaker (white bars) variances per linguistic context factor level for /x/ CoG.*

2.4 Discussion

Previous work on read speech has shown that linguistic effects such as labial co-articulation and syllabic position have effects on fricative acoustics, and that some segments, such as /s/, are more speaker-specific than other segments. The present study wished to further investigate (1) whether linguistic effects on fricative spectra are present in speech materials that were not recorded in highly-controlled circumstances (in

this case, telephone dialogues), and (2) whether there is an interaction between segments' speaker-specificity and their linguistic context.

2.4.1 Linguistic effects

Regarding the first aim, linguistic effects were present in /x/, but were less prominent in /s/. The effect of syllabic position was present in both fricative sounds. Onsets showed higher intensity for both fricatives, which is consistent with results reported by Solé (2003) for American English fricatives. However, only for /s/ was there any indication for coda reduction in spectral measures, namely a higher center of gravity in /s/ onsets compared to codas.

As for labialization, the results confirmed the expected linguistic effects in /x/ acoustics; both left and right labial neighbors lower the resonance frequencies in /x/ by around 200 Hz and 300 Hz respectively. This is consistent with work on /s/ from read speech where anticipatory labialization lowered spectral energy by around 300 ~ 400 Hz (Koenig et al., 2013). Two significant interaction effects for center of gravity and spectral tilt furthermore indicated that spectral lowering is attenuated when both left and right context are labial and that the effect of anticipatory labialization is slightly larger in onsets. Regarding the first interaction, spectral lowering in these cases might be attenuated to not undershoot the articulatory target for /x/. The second interaction could be explained by more resistance to co-articulation across word boundaries; all onsets in this dataset were word-initial and all codas were word-final. This means that right context for onsets was part of the same syllable, whereas left context for onsets was part of the previous word. Previous work, however, found only minor effects of prosodic boundaries on co-articulation of consonant cluster [kl], and then predominantly when articulation rate was slow (Hardcastle, 1985). Regarding fricatives, work on s#CV versus #sCV clusters has shown no effects of word boundary on /s/ duration (Cho et al., 2014; Dumay et al., 1999). These findings suggest that word boundary effects may not explain why anticipatory labialization is larger for onsets than codas. Alternatively, this

interaction may reflect a qualitative difference in the type of lip-rounding; whereas right labial context for onsets consisted of rounded vowels, right labial context for codas consisted exclusively of bilabial consonants /b, p, m/ (because codas followed by vowels were labelled as ambisyllabic). Given that Munson (2004) has shown that the labialization effect in /s/ before /p/ was smaller than before /u/, the present result that anticipatory labialization lowers /x/ spectra more in onsets is therefore likely to stem from the specific labial segments that followed /x/ in onset versus coda position.

Contrary to /x/, the /s/ acoustics did not show the expected spectral lowering in labial contexts; in fact, when right context was labial, center of gravity showed a small but significant increase. The lack of spectral lowering in /s/ acoustics is likely a result of the speech channel used here, as much of the spectral energy for /s/ falls above the upper limit of the telephone bandwidth. In other words, given that the effect of labial co-articulation is well-attested for /s/, it is likely that labial co-articulation effects are not captured in these data. From the literature as well as the current results on /x/, the lowering due to labialization would be on the order of 300 Hz, which – relative to 4.8 kHz for a Dutch /s/ center of gravity – falls outside of the telephone band. This is supported by the mean CoG values; the mixed model's CoG intercept of 1.6 kHz for /x/ (CoG mean from the data was 1,586 Hz, *SD* = 421 Hz) was very similar to previously reported resonance frequencies for Dutch /x/ in broadband speech (Van der Harst et al., 2007). However, for /s/, the mixed model's CoG intercept of 2.5 kHz (*M* = 2,548 Hz, SD = 387 Hz) was around 2 kHz lower than what previous broadband studies have reported (Ditewig et al., 2019). In other words, we assume that the actual spectral peaks for /s/ were far over the upper limit of the landline telephone bandwidth used here, resulting in much lower CoG values in the present analysis with a lack of linguistic effects as a result.

2.4.2 Speaker classification

Regarding the dependence of speaker information on linguistic context in spontaneous telephone speech, the speaker-dependency of fricatives /s/ and /x/ seems to be distributed across linguistic contexts in a systematic way, but differences in speaker-classification accuracies were very small. In the current results, articulatory weak locations, i.e., codas and fricatives with labial neighbors, had slightly better speaker-classification scores than articulatory strong locations, i.e., onsets and fricatives with non-labial neighbors, for both /s/ and /x/. It seems that our data provides further evidence for the hypothesis proposed by He et al. (2017; 2019) that speech locations that may be less constrained by articulatory targets have more between-speaker variation. Moreover, the present study showed that these locations are more speaker-specific. Further examination of the between- and within-speaker variances showed that, for /x/ center of gravity, both between-speaker variance was increased and within-speaker variation was decreased in articulatory weak locations relative to articulatory strong locations.

Interestingly, speech features sampled from articulatory weak locations seemed to have more between-speaker variation even in the absence of clear acoustic differences. Fricative /x/ acoustics were altered by linguistic context within the telephone band and simultaneously showed differences in speaker-classification per linguistic context. However, /s/ also showed higher speaker-classification accuracies in articulatory weak locations, even though the expected acoustic effects for /s/ were minimal. The relative differences in speaker classification per linguistic context were very similar, but small, for both /s/ and /x/. Therefore, there is a possibility that these results are dependent on the specific sampling of the current dataset, which we assume to reflect distributional patterns of conversational Dutch; there are many more /s/ and /x/ tokens with non-labial context than with labial context (see Table 2.2). We cannot exclude that the lower number of labial contexts may have resulted in an under-estimation of speaker variance in that particular context. Given the minor differences between linguistic contexts, however, the results are expected to have no major implications for either listeners' perception of speaker information or for forensic speaker comparisons.

Comparing the contribution of the different acoustic measures to the speaker-classification accuracy of the multinomial logistic regression model, our results are similar to those reported by Kavanagh (2012) for English /s/ from read speech. Namely, spectral center of gravity and standard deviation are speaker-specific acoustic measures compared to temporal and amplitudinal measures. This might be because, whereas spectral measures reflect the size and shape of resonance cavities in the production of fricatives, this is not the case for temporal and amplitudinal measures. The same can be said for the lack of contribution of dynamic spectral measures; whereas static spectral measures reflect the shape and size of the resonance cavity, the dynamic measures reflect temporal patterns of articulation. Given the relatively static nature of fricatives, the lack of contribution of dynamic measures is not surprising. In addition, the short duration of fricatives in spontaneous telephone speech in combination with the large variation in phonetic context might also contribute to the lack of contribution for dynamic measures. Notably, the relative contributions of acoustic measures to speaker-specificity were very similar for the two fricative sounds examined here.

Interestingly, when using the same set of measures, fricative /s/ seems to be slightly more speaker-specific than /x/ even though the spectral peak of /s/ is not captured by the telephone bandwidth. In other words, /s/ retains some speaker-specificity even in limited bandwidths. Moreover, another highly frequent fricative in Dutch, /x/, contains comparable amounts of speaker-specificity in telephone speech. The correlation coefficient between the mean CoG values per speaker for /s/ and /x/ ($r = .46, N = 43, p < .01$) furthermore shows that the two fricative sounds carry partly complementary speaker information.

2.4.3 Limitations

It has to be noted that the current results only apply to male speakers and that it is possible that female speakers would display different behavior. Moreover, although studies have shown that sexual orientation and gender identity affect spectral measures such as CoG for /s/, the Spoken

Dutch Corpus only reports the speakers' sex (Oostdijk, 2000). Furthermore, the telephone dialogues from the Spoken Dutch Corpus were recorded almost two decades ago, which means that these results may not fully generalize to contemporary populations. With regards to Dutch fricatives, it has been shown that there is a general trend of devoicing, whereby /s/-/z/ and /f/-/v/ are merging (Gussenhoven, 1999; Pinget, Van de Velde, & Kager, 2014). In fricative realizations, this progressing merger may result in more variation. This means that it is possible that a contemporary population of speakers of Standard Dutch might show more between-speaker variation for /s/ than the set of male speakers in this study.

The use of the rather simple measures spectral CoG and SD might also be a possible limitation. These measures have been used often in previous work on fricatives, mostly with the goal of distinguishing the different fricative phonemes (e.g., Jongman et al., 2000). Much of this work focused on /s/ especially, which seems to be captured quite well by these measures. However, dorsal fricative /x/ seems to display a formant-like structure for most realizations, i.e., containing multiple spectral peaks. Although CoG seems to capture linguistic effects in /x/, such as contextual labialization, in the expected way, it is possible that some between-speaker variation is captured better by more complex measures such as discrete cosine transforms (DCT: Jannedy & Weirich, 2017). The spectral moments used in this study might thus underestimate the speaker-specificity for fricative /x/.

2.5 Conclusion

The present study investigated the distribution of speaker information in fricatives /s/ and /x/ as a function of syllabic position and labial co-articulation. Results have firstly shown that linguistic contexts affect fricative acoustics; whereas the linguistic-context effects reported in previous studies working with studio-recorded read speech can be replicated for dorsal fricative /x/ in spontaneous telephone speech, this is

less so the case for alveolar fricative /s/. We argue that the lack of effects for labial co-articulation for /s/ is a result of the telephone bandwidth used here. Secondly, for both /s/ and /x/, results showed somewhat more speaker-specificity for codas and for tokens with labial context. However, differences in speaker-specificity per linguistic context were small. These results support the hypothesis that the role of the speaker in speech is more explicit in parts of the speech signal where speakers may have more articulatory freedom, in this case, fricatives occurring in labial context and in coda positions.

CHAPTER 3

Linguistic effects on the speaker-dependent variability in nasals

Abstract

In forensic speech science, nasals are often reported to be particularly useful in characterizing speakers because of their low within-speaker and high between-speaker variability. However, empirical acoustic data from nasal consonants indicate that there is a somewhat larger role for the oral cavity on nasal consonant acoustics than is generally predicted by acoustic models. For example, in read speech, nasal consonant acoustics show lingual coarticulation that differs by nasal consonant, and syllabic position also seems to affect realizations of nasal consonants within

speakers. In the current exploratory study, the within and between-speaker variation in the most frequent nasals in Standard Dutch, /n/ and /m/, was investigated. Using 3,695 [n] and 3,291 [m] tokens sampled from 54 speakers' spontaneous telephone utterances, linear mixed-effects modelling of acoustic-phonetic features showed effects of phonetic context that differed by nasal consonant and by syllabic position. A following speaker-classification test using multinomial logistic regression on the acoustic-phonetic features seems to indicate that nasals displaying larger effects of phonetic context also perform slightly better in speaker classification, although differences were minor. This might be caused by between-speaker variation in the degree and timing of lingual coarticulatory gestures.

This chapter has been published:

Smorenburg, L., & Heeren, W. (2021). Acoustic and speaker variation in Dutch /n/ and /m/ as a function of phonetic context and syllabic position. *Journal of the Acoustical Society of America, 150*(2), 979-989. https://doi.org/10.1121/10.0005845

3.1 Introduction

Models of speech production and perception more and more consider the role of within- and between-speaker variation (cf. Bürki, 2018). Speaker variability is not only relevant for modelling speech, but also for the practice of speaker identification. In forensic speech science, researchers have been trying to establish acoustic-phonetic features that have low within-speaker variation and high between-speaker variation and are therefore effective in discriminating speakers. Among consonants, nasals are often reported to be highly speaker-specific (e.g., Amino & Arai, 2009; Glenn & Kleiner, 1968; Kavanagh, 2012; Su et al., 1974, van den Heuvel, 1996). Nasals' within-speaker variability is argued to be low, and the between-speaker variability to be high. Speaker variation comes from two sources: a speaker's anatomy, i.e., the shape and size of the vocal tract, and articulatory behavior, i.e., the timing and specific movements in articulation (e.g., Nolan, 1983, Chapter 3). Compared to the flexible oral cavity which contains many moving parts that may change its shape and size, the nasal cavity is a rigid resonator that is relatively fixed in shape and size between speakers and, apart from changes brought by nose colds, aging, and surgical procedures, stable within speakers (e.g., Rose, 2002, p. 135).

Acoustic modelling more or less agrees with this view of nasal consonants that exists in forensic speech science; the resonances in nasals are dependent mostly on the pharynx and nasal cavity, thus reflecting a speaker's anatomy, with relatively little influence of the oral cavity as the main vocal pathway runs from the glottis to the nostrils (cf. Johnson, 2003; Stevens, 2000). However, acoustic studies on nasal consonants seem to show a somewhat larger influence of the oral cavity on nasal consonant acoustics (e.g., Tabain, Butcher, Breen, & Beare, 2016) and also show that linguistic factors affect nasal acoustics within speakers. For example, nasal consonant acoustics show lingual coarticulation with the following vowel (Su et al., 1974) and the phonemic contrast between /n/ and /m/ is realized more clearly in onset than in coda position (Seitz et al., 1990), which is possibly related to findings that nasals in onset and coda positions have different articulatory timing mechanisms (Byrd,

Tobin, Bresch, & Narayanan, 2009; Krakow, 1993). One other aspect of nasal consonant acoustics that is not often mentioned in forensic speech science, is that nasals are acoustically weak, i.e., have very low amplitude compared to other speech sounds (e.g., Stevens, 2000). This might be problematic in forensic contexts as the speech in forensic case work often consists of low-quality (telephone) speech.

In the current exploratory study, we investigate the within- and between-speaker variation in Dutch nasal consonant acoustics in intercepted telephone conversations, which is similar to data in forensic case work. Our work has two aims: (1) test whether linguistic factors affect the acoustics of nasal consonants within and between speakers, focusing on lingual coarticulation and syllabic position, and (2) test to what extent speaker discrimination depends on the linguistic context from which tokens are sampled.

3.1.1 Nasal consonants

3.1.1.1 Dutch

In the language under investigation, Dutch, there are three nasal consonants: bilabial /m/, alveolar /n/, and velar /ŋ/, with the latter only occurring in intervocalic (/χɪŋə/ *gingen* 'went') or postvocalic position (/zɪŋ/ *zing* 'sing'). The bilabial and alveolar nasals also occur in prevocalic position across word classes and are therefore more frequent in Dutch (Luyckx, Kloots, Coussé, & Gillis, 2007). Although it does not occur in Standard Dutch, some dialects also have a syllabic nasal (e.g., [wetn̩] 'to know': Van Oostendorp, 2001). Standard Dutch does not have nasal vowels, but they may occur in loanwords (Gussenhoven, 1999, p. 75).

3.1.1.2 Acoustic models

Nasal sounds are articulated with a lowered velum, which opens the nasal cavity and makes sound produced at the vocal chords resonate in the nasal cavity (Stevens, 2000, pp. 187 - 194 and 487 - 513). In nasal consonants, air is blocked from passing through the oral cavity by the lips or a lingual constriction and is instead released through the nasal cavity. For the velar nasal consonant, the oral cavity is entirely closed off at the lingual constriction at the velum, which means that the air flows from the glottis to the nostrils. The simplest model for the velar nasal consonant is a simple tube model consisting of the pharynx and nasal cavity, with evenly spaced resonances reflecting the length of the tube which is estimated to be around 21.5 cm for an adult (9 cm pharynx plus 12.5 cm nasal cavity: Johnson, 2003, p. 152), with some models also considering varying tube widths along the length of this vocal tract which results in predicted resonances at slightly different frequencies (cf. Stevens, 2000; Fant, 1970). Acoustic models (Fant, 1970; Fujimura, 1962; Johnson, 2003; Stevens, 2000) generally predict the following resonances for the velar nasal consonant: a low first formant at around 200 ~ 400 Hz that arises from the pharynx with a relatively wide bandwidth, a second formant at around 750 ~ 1,100 Hz that arises from the nasal cavity, a third formant at around 1,700 ~ 2,200 Hz that also arises from the pharynx, and a fourth formant at around 2,300 ~ 3,000 Hz that arises from the nasal cavity. Because the coupling of the nasal cavity with the pharyngeal cavity lengthens and increases the surface area of the vocal tract, more sound is absorbed in nasal than in oral sounds (Fant, 1970). As a result, nasal sounds have relatively low amplitude, particularly in frequency regions above 500 Hz, and lower resonance frequencies than oral sounds.

Requiring different modelling than the velar, the bilabial and alveolar consonants have more anterior constrictions which result in a side-branch off the main pathway that is open at the uvula and closed at the bilabial or alveolar constriction. Johnson (2003) and Stevens (2000) describe this side-branch as a simple tube that is closed off at one end and absorbs energy from the main tube at certain frequency regions (around 1,000 Hz for /m/ and 1,600 - 1,900 Hz for /n/) depending on the length of the tube (around 8 - 9 cm for /m/ and 5 - 6 cm for /n/). In these models, the antiresonances, or antiformants, that arise from the oral

cavity and their effects on the resonances that arise from the pharyngeal-nasal tract provide the only cue to place of articulation in nasal consonants. Fant (1970), however, sees the oral cavity not as a side-tube but as a Helmholtz resonator with the neck at the velum, which, in addition to antiformants, also outputs oral formants at around 900 Hz for /m/ and 1,200 ~ 1,400 Hz for /n/ (p.145 - 146).

From these models, it is not clear what role the shape of the lips or tongue may play in nasal consonant acoustics. However, even the models that see the oral cavity as a simple side-tube predict that the length of the oral cavity affects the acoustics through, at least, the location of the antiformants (the more forward the constriction and therefore the longer the oral cavity side-tube, the lower the antiformant). These antiformants may neutralize or shift the resonance frequencies that arise from the pharynx and nasal cavity. When the frequency of an antiformant coincides with the frequency of a formant, the formant will be attenuated or neutralized (as the oral side-tube absorbs energy from the main tube at this frequency). When the antiformant is in the vicinity of the formant, the formant's lower or upper energy in attenuated or neutralized, thus shifting the formant. This ultimately results in different resonance frequencies for /m/ and /n/.

3.1.1.3 Empirical acoustic data

As nasal consonants are acoustically weak, i.e., have low amplitude, acoustically distinguishing between nasal phonemes is difficult, and much work on nasal consonant acoustics seems concerned with this problem (e.g., Kurowski & Blumstein, 1984; Mermelstein, 1977). Although the current work is not particularly concerned with distinguishing the nasal phonemes, but rather with observing how phonetic context and syllabic position affect the acoustics and the idiosyncratic speaker information in nasal consonants, the two research aims are not entirely unrelated, as both involve the acoustic measurement of variations in place of articulation.

Acoustic modelling generally attributes most of the resonance frequencies in nasal consonants to be associated with the pharynx and nasal cavities, with a relatively small role to play for the oral cavity in

the form of antiformants that may shift or neutralize the resonances produced by the pharynx and nasal cavities. Empirical acoustic data, on the other hand, seems to imply a somewhat larger role for the oral cavity. In acoustic data from nasal consonants from (mostly) female speakers of three Australian languages, /n/ and /m/ were distinguishable along each of the four nasal formants that were measured, with lower formant values for /m/ than for /n/ (Tabain et al., 2016), whereas acoustic models describe that only formants in the vicinity of antiformants, i.e., N2, N3 and possibly N4, should be affected by PoA. Considering the oral cavity as a simple side-tube of 7 - 8 cm in length for /m/ and 5 - 6 cm for /n/ to the main 19.5 cm long pharyngeal-nasal passage, antiresonances are predicted at 1,000 ~ 1,200 Hz for /m/ and 1,600~1,900 Hz for /n/ (Stevens, 2000, pp. 494 - 513). Grigorjevs (2012) also points out that there is some discrepancy between acoustic modelling and observations from real language data, where it seems to be the case that the antiformant for /m/ is generally observed to be lower than predicted and the antiformant for /n/ more or less equal as predicted by simple tube models, with some variation between languages. This indicates that tube models might not fully account for acoustic observations. The relation between articulatory variables and acoustic-phonetic features is therefore not entirely clear for nasal consonants. From acoustic modelling and previous empirical findings, it is plausible that variations in place of articulation caused by phonetic context might have a measurable effect on nasal consonant acoustics.

3.1.2 Within and between-speaker variability in nasals

As mentioned before, there are two sources of between-speaker variation: anatomy and articulatory behavior. Whereas the former is relatively stable, i.e., is not also a source of within-speaker variation (except for colds, surgeries, etc.), the latter concerns learned motor behavior and is dependent on, e.g., language, speech register, social factors, and linguistic structure. Regarding linguistic structure, there is a general hypothesis that parts of the speech signal that are less constrained to reach articulatory targets may display more between-speaker variation in

articulation (cf. He & Dellwo, 2017). Evidence for this hypothesis was found in intensity and first-formant dynamics from syllables, which reflect mouth-opening and closing gestures. Mouth opening gestures, such as during the articulation of onsets towards nuclei, are described as having more precise articulations than mouth closing gestures, such as during the articulation towards codas (Ohala & Kawasaki, 1984). Regarding the speaker variation in articulation, more between-speaker variation was found in the second than in the first part of syllables for both intensity (He & Dellwo, 2017) and first-formant dynamics (He, Zhang, & Dellwo, 2019). Earlier work on speaker variation in fricatives corroborates this hypothesis. Fricative acoustics are highly dependent on the labialization of neighboring segments (e.g., Koenig, Shadle, Preston, & Mooshammer, 2013) and the between-speaker variation in fricatives in labialized contexts was found to be slightly higher than fricatives in non-labialized contexts, assumedly because of between-speaker variability in the degree and timing of the lip-rounding movement (Smorenburg & Heeren, 2020).

In the following two subsections, previous research on the effects of phonetic context and syllabic structure are discussed for nasal consonants.

3.1.2.1 Phonetic context

In nasals, the lowering of the velum may carry over to neighboring speech sounds, resulting in distinct nasality in speech sounds that would otherwise be oral (e.g., Jang et al., 2018). How preceding and following context affect nasal consonants has not received as much attention in the literature. The few studies on this topic indicate that neighboring vowels may also affect nasals; nasal consonants may show lingual coarticulation with neighboring speech sounds (e.g., Fujimura, 1962; Su et al., 1974). These coarticulation patterns seem to vary by nasal consonant. For speakers of English, Su et al. (1974) founds that the Euclidean distance of filter bank spectra (using 25 filters from 250 - 3681 Hz) between nasal consonants followed by front vowels versus back vowels was three times larger for /m/ than for /n/. In other words, there was more anticipatory lingual coarticulation for /m/ than for /n/. This was attributed to the lack

of an articulatory tongue target for bilabial /m/ versus the alveolar tongue target for coronal /n/ (Su et al., 1974). The lack of an articulatory tongue target for /m/ seems to result in the tongue having more articulatory freedom to anticipate following speech sounds. Others have also observed that /m/ shows larger effects of phonetic context than /n/ (Fujimura, 1962, p. 1873; Tabain, 1994, cited in Tabain et al., 2016, p. 892). In Su et al. (1974), the degree of coarticulation, i.e., the Euclidean distance between front and back vowel contexts, was also used in a speaker-classification test. Results showed that the degree of coarticulation for /m/ was more predictive for speakers than for /n/. This means that there was more between-speaker variation in the acoustics dependent on the following vowel for /m/ than for /n/.

3.1.2.2 Syllabic position

Some speech styles and some positions in speech are articulated with more effort than others, which affects the acoustics. For example, spontaneous speech is generally articulated faster and with less effort than read speech and the comparison between the two speech styles is often used to investigate speech reduction (e.g., Van Bael et al., 2004). Like vowels, Dutch nasal consonants have shorter durations and lower center of gravity (CoG) in spontaneous speech than in read speech, but opposed to other speech sounds, nasals in spontaneous speech did not have reduced amplitude (Van Son & Van Santen, 2005).

Regarding positional effects of articulatory effort within one speech style, coda reduction is a well-known phenomenon, with codas being more 'sloppy' and reduced than onset consonants (e.g., Ohala & Kawasaki, 1984). The effect of syllabic structure on nasal consonants has mostly been investigated in terms of articulation. Real-time MRI research has shown that timing mechanisms for articulatory gestures in nasals vary by syllabic position; the alveolar nasal in onset position shows a timing synchrony in the tongue tip raising and velum lowering gestures, whereas in coda position there seems to be a time lag between gestures, with velum lowering occurring earlier in the preceding vowel (Byrd et al., 2009). Similar synchrony in onset nasals and lags in coda nasals were found for the lip-closing gesture and velum-lowering gesture in /m/

(Krakow, 1993). Regarding the acoustics, a direct comparison between onset and coda nasal consonants seems to be lacking in the literature, instead focusing on distinguishing the different nasal consonants. The transition between the murmur and the vowel has long been found useful in distinguishing place in nasal consonants (e.g., Kurowski & Blumstein, 1984; Mermelstein, 1977), but not equally useful across syllabic positions; measures of spectral change between the nasal murmur and vowel show a clearer differentiation between /n/ and /m/ in onset than in coda position (Seitz et al., 1990).

In perception, syllabic position also seems to affect speaker discrimination. In Japanese read speech, perceptual speaker identification by listeners showed better accuracy for syllables containing onset nasals than coda nasals (Amino, Arai, & Sugawara, 2007)[3]. Onset consonants are generally articulated more precisely than coda consonants (Ohala & Kawasaki, 1984) and often have longer durations and higher signal-to-noise ratios (SNR), both of which could potentially be causing this advantage in speaker classification from an acoustic perspective.

Given the different timing mechanisms in articulation of nasal consonants by syllabic position, the between-speaker information stemming from articulation might also vary by syllabic position.

3.1.3 Research questions

Nasal consonants have received much attention in forensic speech science for their usefulness in speaker discrimination. From acoustic models, it seems that the resonances in nasal consonant acoustics are mainly dependent on the pharynx and nasal cavity, with influence from the oral cavity only through the presence of antiformants. This would

[3] These results should be interpreted with caution; it is possible that nasal consonants in coda position, or moraic nasals, may be articulated differently in Japanese because they have fewer phonetic competitors.

mean that nasal acoustics are highly dependent on the anatomy of the speaker and therefore have high between- and low within-speaker variability. Empirical acoustic data however, shows a larger role for the oral cavity than acoustic models (cf. Tabain et al., 2016) and others have also shown that nasal acoustics are dependent on their phonetic context (e.g., Su et al., 1974) and on syllabic position (e.g., Seitz et al., 1990). Therefore, within- and between-speaker variability in nasal consonant acoustics may also be affected by articulation.

The current work aimed to investigate the variability in the acoustics of nasal consonants across linguistic factors and speakers. So far, Su et al. (1974) have shown that there seems to be anticipatory lingual coarticulation with the following vowel in the acoustics of /n/ and /m/ and that the degree of coarticulation is larger for /m/. The degree of coarticulation was also highly speaker-specific, i.e., there was between-speaker variation in the degree and/or timing of coarticulation of /m/ with the following vowel. This suggests that nasal consonant acoustics do not only contain anatomical idiosyncrasies, but also articulatory idiosyncrasies.

In the first part of this study, the effects of phonetic context and syllabic position on the acoustics of Dutch /n/ and /m/ were investigated, also looking at the between-speaker variation of these effects. Given some inconsistencies between acoustic modelling and empirical data, it is unclear which acoustic-phonetic features could be sensitive to phonetic context and syllabic position, but it is plausible that at least the formants (and their bandwidths) in the vicinity of antiformants could be affected. It is further expected that /m/ will show larger effects of phonetic context than /n/, because the lack of an articulatory target for the tongue in /m/ might allow for larger carry-over and anticipatory lingual gestures than in /n/. Some effects of syllabic position on /n/ and /m/ acoustics are also expected, given the articulatory timing differences by syllabic position (Byrd et al., 2009; Krakow, 1993) and the clearer place distinction in onset than in coda position (Seitz et al., 1990).

In the second part of this study, a speaker-classification test was performed to investigate to what extent speaker discrimination is dependent on linguistic factors. It was expected that, if /m/ showed larger

between-speaker variation of linguistic effects than /n/ in the first experiment, that this would be reflected in differences in speaker-classification accuracies.

3.2 EXPERIMENT I: Acoustics

3.2.1 Materials and speakers

Nasal consonants were sampled from telephone dialogues intercepted via a switchboard from the Spoken Dutch Corpus (Oostdijk, 2000). Speakers were recorded from their home landline telephone while conversing with a male or female speaker for around ten minutes on a topic of their choice. For each speaker, between one and four telephone conversations were available in the corpus (M = 1.8, SD = 1.1). We chose this component of the corpus because it seems to resemble natural speech most closely; speakers were in their home environment and conversed with speakers previously known to them. In addition to being representative for everyday natural speech, the speech from the selected part of the corpus is in ways comparable to speech found in forensic casework where experts often analyze conversational speech in low-quality telephone recordings.

Speakers were selected on their language variety and sex. Given the overrepresentation of this general population in police investigations and the possible relevance of this work to forensic speaker comparisons, we chose to further limit our dataset to male speakers between the ages of 18 and 50. To exclude dialect speakers, only speakers of Standard Dutch (home, work, and education language) were included. This means that this work focused itself on the variation present in a relatively homogeneous set of speakers. These exclusion criteria left 60 speakers from the relevant component of the corpus.

3.2.2 Segmentation

The orthographic transcription available in the Spoken Dutch Corpus was used to segment the speech signal in a forced-alignment protocol. Given the many reductions and deletions in spontaneous speech, the result of this segmentation was not very accurate. Therefore, the automatic segmentation functioned as a tool to locate the nasal consonants in the speech signal for manual segmentation of target tokens along with their immediate phonetic context. Tokens were excluded if (1) tokens were reduced to the extent that they were not auditorily identifiable, (2) the interlocutor or noise could be heard in the background, (3) the speaker put on a marked voice (such as in an accent imitation) or was laughing, (4) the tokens were shorter than 30 milliseconds, or (5) tokens were ambisyllabic (lexical codas followed by a vowel, e.g.: om een 'around a' [ɔm.ən]) and could not be classified as onsets or codas.

Each token was coded for syllabic position (onset versus coda) and neighboring segments to the left and right of each nasal were coded for place of articulation (PoA, non-back versus back)[4]. The non-back category included front vowels, consonants with a bilabial to palatal place of articulation, the schwa vowel, and pauses. The back category included back vowels and consonants with a velar to uvular place of articulation. The specific speech sounds included in these categories are presented in Table 3.1.

[4] The mixed-effects model analysis was also performed using factor levels 'front' and 'back' for factors Left PoA and Right PoA, which excluded pauses and mid-vowels. Although exact coefficients were different, the significant effects were similar. The non-back versus back distinction was then chosen because it included more tokens.

Table 3.1: *Non-back versus back categorization of Dutch phoneme context.*

	Vowels	Consonants
Non-back category	i, ɪ, y, ʏ, ø, e, ɛ, ə	p, b, m, f, v, ʋ, s, z, t, d, n, l, ʃ, ʒ, j
Back category	u, ɔ, o, a, ɑ	k, g, ŋ, x, χ

The rhotic did not receive a categorization because of its variable place of articulation in Dutch and the glottal consonant did not because there is no oral constriction for this sound. This coding scheme for phonetic context was selected for three reasons: firstly, this categorization could be applied to both vowels and consonants. Secondly, as /m/ does not have an articulatory tongue target and could therefore have a neutral, i.e., mid, tongue position when spoken in isolation, this categorization would capture effects of back articulation for both /n/ and /m/. Lastly, a binary categorization ensured sufficient token numbers per factor level.

The exclusion criteria resulted in some speakers having very low token numbers per factor level. It was therefore decided to only include speakers with at least eight tokens per factor level. This excluded six speakers. The remaining numbers of tokens for 54 speakers are presented in Table 3.2.

Table 3.2: *Numbers of tokens per factor level by speaker*

			Syllabic Position		Left context place of artic.		Right context place of artic.	
		Total	Onset	Coda	Non-back	Back	Non-back	Back
/n/	Total	3,695	2,265	1,430	2,417	1,278	2,694	1,001
	M	68	42	26	45	24	50	19
	(SD)	(23)	(18)	(10)	(18)	(8)	(16)	(9)
	Range	23-127	10-95	9-77	15-91	8-42	17-99	8-43
/m/	Total	3,291	2,357	934	2,189	1,102	1,916	1,375
	M	61	44	17	41	20	35	25
	(SD)	(19)	(17)	(8)	(14)	(8)	(13)	(8)
	Range	19-103	8-66	8-41	12-80	8-49	16-70	8-41

3.2.3 Acoustical analysis

As noted before, the relation between acoustic-phonetic features and the articulation of nasals is not entirely clear from the literature as the role of the oral cavity seems to play a somewhat larger role in empirical data than it does in acoustic models. The acoustical analysis was performed in Praat (Boersma & Weenink, 2020) and has been adapted from Tabain et al. (2016) to be suitable for male speakers and for the telephone bandwidth of 300 - 3,400 Hz. First, the duration was measured from the nasal onset to the offset as determined by low-amplitude and low-frequency spectral energy characteristic of nasal consonants. Second, the middle 50% of each consonant was used to estimate two spectral moments (center of gravity and standard deviation), the second (N2),

third (N3), and fourth nasal formants (N4), and their bandwidths (BW2, BW3, and BW4). The first formant was not included as it cannot be reliably measured in telephone speech because of the 300 - 3,400 Hz band pass. For the N4 and BW4, some undefined values were returned (N = 131), meaning that the N4 for some tokens probably exceeded the upper limit of the telephone band, but given this only concerned a relatively small number of tokens and the mean N4 was not too close to the upper frequency limit of 3,400 Hz, the N4 was still included in the analysis. Although the spectral moments are a very simplified estimation of the spectrum for speech sounds with formant structures like in /n/ and /m/, CoG is often highly correlated with formant values and might therefore be a very simple measurement to capture effects of phonetic context and syllabic position[5]. Formants and their bandwidths were measured over the 800 - 3,400 Hz band using the Burg method, querying three formants in that range. These metrics might vary by place of articulation; antiformants produced by the oral cavity (whose frequency varies by the length of the oral cavity and thus by place of articulation) may dampen or shift formants and their bandwidths.

3.2.4 Statistical analysis

Linear mixed-effects modelling (LME) was used to investigate effects of phonetic context and syllabic position on nasal consonant acoustics. Given previous findings showing larger anticipatory lingual coarticulation for /m/ than for /n/, we also tested whether the effects of context and syllabic position differed by nasal consonant. Again, we were not particularly concerned with distinguishing the two nasal consonants, but rather with testing whether linguistic effects differed by nasal consonant.

Linear mixed-effects modelling was performed in R version 3.6.3 (R Core Team, 2019). Fixed and random effects were estimated automatically with the Bayesian Information Criterion (BIC) and backward stepwise selection using function *buildmer()* from R package

[5] In the current data, Pearson correlation coefficients between CoG and formants were .58 for N2, .56 for N3 and –.13 for N4.

'buildmer' (Voeten, 2020). The user-specified maximal model included treatment-coded fixed factors Nasal (/n/, /m/), Syllabic Position (ONSET, CODA), Left Context (NON-BACK, BACK place of articulation), Right Context (NON-BACK, BACK), and interactions. Interactions between fixed factors were also tested because previous research has shown different gestural timing effects in nasals for onsets and codas (Byrd et al., 2009; Krakow, 1993) and larger coarticulatory effects in /m/ than in /n/ (Su et al., 1974). In the random structure of each model, by-speaker intercepts and slopes over fixed effects were estimated. The *p*-values for fixed effects were tested empirically by parametric bootstrapping using function *mixed()* from R package 'afex' (nsim = 10,000; Singmann, 2019). Additionally, the alpha level for significance was Bonferroni-corrected to 0.05 / (9 × 2), to account for the fact that the acoustic measures (N = 9) and nasal consonants (N = 2) were extracted from the same speakers in the same telephone recordings and therefore cannot be assumed to be entirely independent.

3.3 Results I

In Table 3.3 (/n/) and Table 3.4 (/m/), the means and standard deviations for the acoustic measures by factor level are presented.

Table 3.3: *Acoustic measures' mean and standard deviation by factor level for /n/ (all in Hz, duration in ms)*

			Syllabic position				Left context				Right context			
Measure	Total		Onset		Coda		Back		Non-back		Back		Non-back	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Dur	65	28	64	31	67	23	66	31	64	22	66	29	62	26
CoG	1,753	353	1,792	329	1,693	380	1,807	329	1,652	374	1,755	359	1,749	334
SD	580	135	571	125	595	150	569	123	600	154	580	138	579	129
N2	1,117	134	1,140	128	1,080	136	1,140	134	1,072	123	1,113	138	1,126	123
N3	2,037	187	2,038	187	2,035	186	2,043	181	2,026	197	2,039	186	2,033	190
N4	2,647	182	2,633	182	2,669	179	2,634	178	2,672	185	2,654	183	2,628	176
BW2	163	107	172	106	148	106	182	109	127	91	166	111	153	93
BW3	423	271	419	281	429	254	406	265	454	278	416	265	441	284
BW4	441	365	448	372	431	353	453	377	418	338	445	366	430	363

Table 3.4: Acoustic measures' mean and standard deviation by factor level for /m/ (all in Hz, duration in ms)

Measure	Total		Syllabic position				Left context PoA				Right context PoA			
			Onset		Coda		Non-back		Back		Non-back		Back	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Dur	75	43	69	21	92	70	79	50	68	23	79	53	70	22
CoG	1,584	340	1,577	341	1,602	337	1,607	330	1,538	355	1,617	326	1,538	353
SD	569	139	557	136	600	142	560	130	588	155	569	136	570	144
N2	1,067	105	1,068	104	1,066	106	1,085	106	1,033	93	1,089	107	1,037	93
N3	2,035	162	2,039	158	2,027	171	2,038	155	2,030	176	2,031	163	2,040	161
N4	2,717	221	2,714	232	2,723	190	2,720	228	2,711	206	2,736	219	2,690	220
BW2	113	71	109	68	123	78	116	73	106	67	121	74	102	66
BW3	329	220	310	217	377	218	306	208	375	233	327	222	332	217
BW4	637	412	676	423	540	366	665	417	580	397	649	426	619	391

Optimal LME models are shown in Table 3.5. One immediate observation is that there are many significant effects of nasal consonant, left and right phonetic context, and syllabic position, as well as many significant interactions between these factors.

Table 3.5: *Best-fitting linear mixed-effects models, N = 6,986, n = 54. Non-significant effects are highlighted in italic.*

	CoG [Hz]			SD [Hz]		
Effect	*Est*	*SE*	*t*	*Est*	*SE*	*t*
(intercept)	1,836	31	59.4	567	9	59.9
Nasal = /m/	–211	14	–14.8	*–24*	*8*	*–3.0*
Left = BACK	–102	11	–8.9	*–10*	*7*	*–1.5*
Right = BACK	–63	8	–8.2	*8*	*7*	*1.2*
SyllPos = CODA	*–43*	*16*	*–2.7*	*–3*	*7*	*–0.4*
Nasal×Left	54	15	3.7			
Nasal×Right				*21*	*7*	*3.1*
Nasal×SyllPos	96	15	6.2	*23*	*7*	*3.2*
SyllPos×Left	–89	15	–5.9	72	7	11
	N2 [Hz]			BW2 [Hz]		
(intercept)	1,146	12	99.6	191	7	26.1
Nasal = /m/	–32	7	–4.9	–70	5	–13.7
Left = BACK	*–12*	*5*	*–2.3*	–19	4	–4.2
Right = BACK	*–11*	*4*	*–2.7*	–33	3	–10.3
SyllPos = CODA	*–3*	*4*	*–0.7*	*2*	*4*	*0.6*
Nasal×Left	*–18*	*6*	*–2.8*	*11*	*5*	*2.1*
Nasal×Right	–52	5	–9.7	18	4	4.1
Nasal×SyllPos	*–6*	*8*	*–0.8*	*18*	*7*	*2.8*
SyllPos×Left	–104	7	–15.7	–73	6	–13.1
Nasal×Syll×Left	52	11	4.8	47	9	5.3

	N3 [Hz]			BW3 [Hz]		
Effect	*Est*	*SE*	*t*	*Est*	*SE*	*t*
(intercept)	2,041	13	151.4	376	17	22.5
Nasal = /m/				–95	11	–8.3
Left = BACK	–15	4	–3.8	36	8	4.7
Right = BACK				47	6	7.2
SyllPos = CODA				*17*	*9*	*2*
SyllPos×Left				47	12	3.8
	N4 [Hz]			BW4 [Hz]		
(intercept)	2,643	13	207.8	480	23	20.6
Nasal = /m/	120	13	9.4	275	22	12.4
Left = BACK	*–12*	*8*	*–1.5*	*1*	*12*	*0.1*
Right = BACK	*–7*	*7*	*–1*	*–26*	*15*	*–1.8*
SyllPos = CODA	*–8*	*11*	*–0.7*	*22*	*16*	*1.4*
Nasal×Left	*–6*	*12*	*–0.5*			
Nasal×Right	*–53*	*10*	*–5.3*	–100	21	–4.7
Nasal×SyllPos	*–6*	*15*	*–0.4*	–128	22	–5.8
SyllPos×Left	96	12	7.7	–115	20	–5.8
Nasal×Syll×Left	–85	20	–4.2			
	Dur [log(ms)]					
(intercept)	1.76	0.005	376.1			
Nasal = /m/	0.04	0.003	12.7			
Left = BACK	–0.02	0.004	–4.3			
Right = BACK						
SyllPos = CODA	0.03	0.006	5.3			

Cog and N2 were positively correlated (r = .58) and showed similar effects. CoG showed a lowering when right context had a back place of articulation. For left context, this lowering effect was mediated by nasal consonant (slightly less lowering in /m/) and by syllabic position (more lowering in codas). N2 showed a lowering when right context had a back place of articulation which differed by nasal consonant (more

lowering in /m/) and a lowering when left context had a back place of articulation which differed by nasal consonant and syllabic position (smaller lowering for /m/ than /n/ in codas, see Figure 3.1).

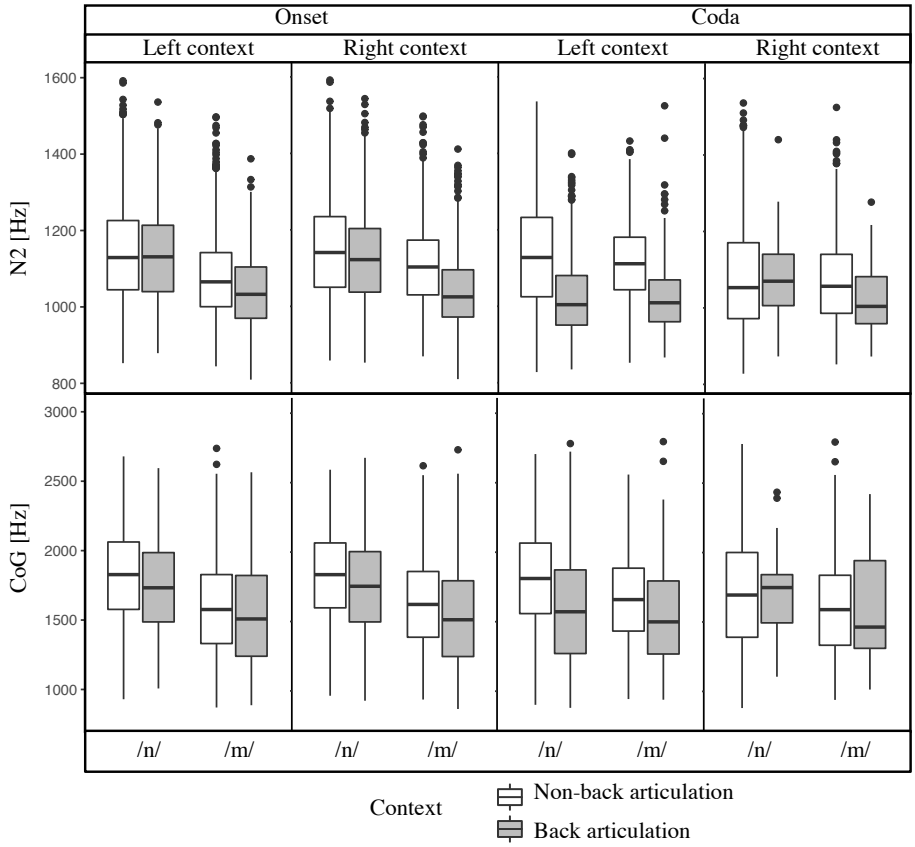

Figure 3.1. *Boxplots for N2 and CoG (Hz) by place of articulation of left and right context, nasal consonant, and syllabic position*

For N3 and N4, linguistic effects were generally smaller and less consistent than for CoG and N2. N3 only showed a small lowering (–15 Hz) effect when preceding context had a back place of articulation. N4 was lower for /m/ when following context had a back place of articulation. When preceding context had a back place of articulation, /n/ had a higher N4, but only in codas.

Linguistic effects on formant bandwidth measures seem to be less consistent than those on the nasal formants. BW2 is smaller when left context has a back place of articulation, more so in codas than in onsets, which further differs by nasal consonant (the lowering of BW2 when left context has a back place of articulation in codas is smaller for /m/ than for /n/). Whereas N3 only showed an effect of Left Context, BW3 also shows an effect of Right Context. BW3 is higher when left context has a back place of articulation, which differs by syllabic position (this effect is larger in codas than in onsets). BW3 is also higher when left context has a back place of articulation. Lastly, BW4 is lower when right context has a back place of articulation for /m/ and when left context has a back place of articulation for codas. Lastly, SD was larger when preceding and following context had a back place of articulation, but only for /m/, and log-transformed duration was longer for /m/ than /n/, shorter when preceding context had a back place of articulation, and longer in codas.

In summary, best-fitting models show effects of a lowering in resonance frequencies when preceding and following phonetic context had a back place of articulation. These phonetic context effects are most prominent in CoG and N2 (also see the change in N2 in the spectral slices from two randomly selected /m/ tokens in non-back versus back-articulated context in Figure 3.2) and interacted with nasal consonant and syllabic position (see Figure 3.1). Generally speaking, for onsets, there are larger effects of right context and larger effects for /m/. Whereas for codas, there are larger effects of left context and larger effects for /n/.

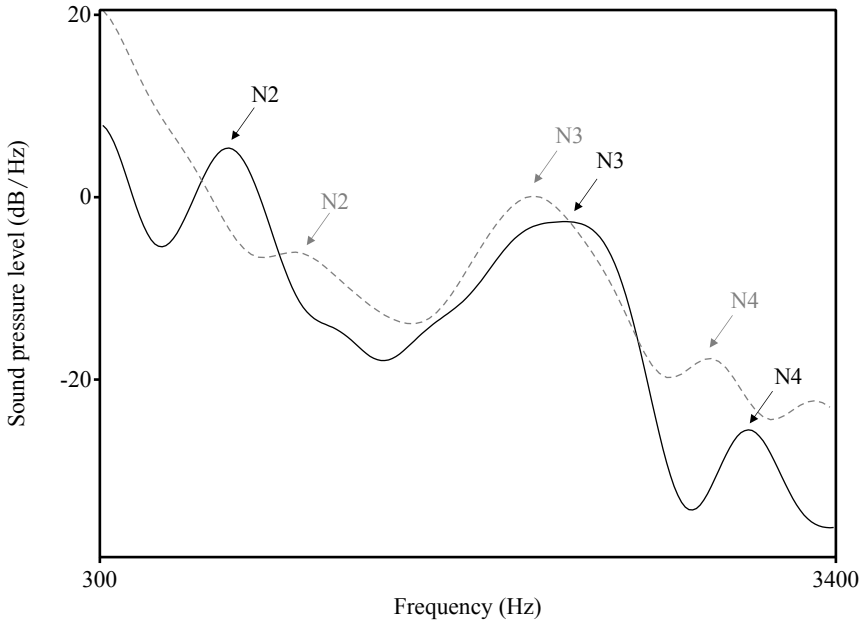

Figure 3.2: *Spectral slices for two /m/ tokens from the same speaker, taken from the mid-50% of each token with cepstral smoothing (500 Hz). Grey dashed line: /m/ in phonetic context with a non-back place of articulation* (was meestal, *'was usually'*, /ʋɑs.mes.tɑl/; *N2 = 1,143 Hz, N3 = 2,126 Hz, N4 = 2,890 Hz). Black solid line: /m/ in phonetic context with a back PoA (*hoe moet, *'how must'*, /ɦu.mut/; *N2 = 842 Hz, N3 = 2,248 Hz, N4 = 3,052 Hz).*

Regarding the between-speaker variation in these linguistic effects, random by-speaker slopes over Left Context were included in the best-fitting model for SD, N2, and BW2. Over Right Context, only the model for SD contained by-speaker slopes. Best-fitting models for CoG, SD, N3, N4, and log-transformed duration contained by-speaker slopes over Syllabic Position. For the factor Nasal Consonant, all measures except for log-transformed duration included random by-speaker slopes.

The random structures of the models indicate that there is significant between-speaker variation in these effects.

3.4 EXPERIMENT II: Speaker classification

3.4.1 Materials

The same materials were used as in experiment I.

3.4.2 Statistical analysis

Speaker-classification systems were built using multinomial logistic regression (MLR) in R version 3.6.3. (R Core Team, 2019). Specifically, function *glmnet()* from R package 'glm-net' (Friedman et al., 2010) was used to perform lasso regression, which uses coefficient shrinkage to simplify models and avoid overfitting, thus improving prediction accuracy and generalizability. Coefficient shrinkage uses a penalty λ, which was determined with cross-validation using function *cv.glmnet()*. By default, this function divides the data into ten folds; one is used for validation (i.e., to generate predictions with) and the remaining nine folds are used to fit the model with a sequence of different λ values. The λ value at which the minimal prediction error was found across folds was selected to shrink the coefficients in the final model, which was built using function *glmnet()*. This shrinkage can be seen as a threshold for contributing predictor coefficients; coefficients that did not improve prediction accuracy across folds in the cross-validation are now shrunk to zero, thus only leaving the coefficients that improved prediction accuracy across folds to be non-zero. The following predictors were entered in the model: nine acoustic measures (CoG, SD, N2, N3, N4, BW2, BW3, BW4, and log-transformed duration) and four binary factors (Nasal, Syllabic Position, Left Context, and Right Context), and all possible interactions between predictions (e.g., CoG × Nasal × SyllPos × Left

Context), excepting those between acoustic measures (e.g., CoG × SD) and between Left Context and Right Context.

Models were built on 70% of the data and predictions were generated from the other 30% of the data, using ten iterations of random sampling. In the first part of this analysis, 70% of the data from /n/ and /m/ was used and non-zero predictor coefficients from the best-fitting model were inspected to see which acoustic measures and linguistic factors significantly improved speaker discrimination. A speaker-classification accuracy was also generated. In speaker-classifications, the model selects the speaker with the highest probability for each token and this decision is then checked to see whether the correct speaker was selected. The classification accuracy of a model equals the number of correctly classified tokens divided by the total number of tokens.

Experiment I showed effects of phonetic context that differed by nasal consonant and syllabic position and further showed significant between-speaker variation (as indicated by the inclusion of random by-speaker slopes) for many acoustic measures. In a second part of this analysis, the data were split on factor Nasal (/n/, /m/), and each nasal on Syllabic Position (ONSET, CODA). Train and test data were then sampled from matching conditions to see whether the speaker discrimination was dependent on these linguistic factors.

3.5 Results II

The speaker-classification model using all /n/ and /m/ data had a mean speaker-classification accuracy of 18.7% over ten iterations of random sampling (range: 18.2% - 20.5%). Inspecting the non-zero predictor coefficients of the model (see Figure 3.3), much speaker variability was present; different sets of predictors are used for each speaker. Despite the variability, some general observations can be made. Firstly, an average of seven (*SD* = 1.2, range = 4 - 9) out of nine acoustic measures were included per speaker, indicating that each speaker needed at least four acoustic measures for optimal predictions. Secondly, there were no large

differences in how many times specific acoustic measures were included across the 54 speakers (M = 41.9, SD = 3.5, range = 34 - 45), which indicates that all the acoustic measures contained useful speaker information. Thirdly, there was a lot of speaker variability in the inclusion of interaction predictors, indicating that the information whether a measurement came from /n/ versus /m/, onset versus coda position, or whether preceding and following context had a non-back versus back place of articulation was not consistently predictive for speakers.

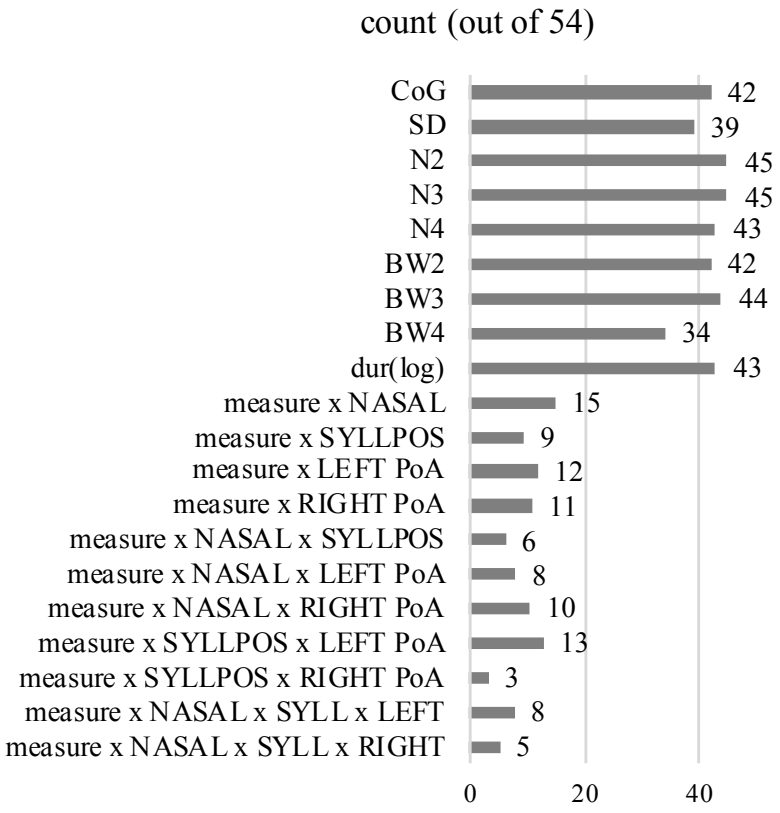

Figure 3.3. *Count of non-zero coefficients for 54 speakers. Counts of interaction predictors were averaged over acoustic measures (N = 9).*

In Table 3.6 we present the speaker-classification accuracies by nasal consonant and syllabic position. Generally, classification accuracies across linguistic conditions are very similar, i.e., all between 17.7% - 22.0%. These classification-accuracy differences between linguistic conditions are about the same size as differences that arise from random sampling iterations within conditions (see classification-accuracy ranges in Table 3.6), indicating that they should be considered minor differences. Nevertheless, some patterns are discernable; /m/ outperforms /n/, /n/ codas outperform /n/ onsets, and /m/ onsets outperform /m/ codas.

Table 3.6. *Speaker-classification accuracies (median and range in percentages over ten iterations of random sampling)*

		Syllabic Position	
	All data	Onset	Coda
/n/	19.4%	18.8%	20.0%
	(17.0 - 20.7%)	(16.8 - 21.0%)	(17.8 - 24.0%)
/m/	21.1%	22.0%	17.7%
	(17.8 - 22.9%)	(20.3 - 23.3%)	(14.5 - 22.5%)

3.6 Discussion

The current work investigated the within and between-speaker variability in nasal consonant acoustics as a function of linguistic factors. Using conversational telephone speech, the first experiment confirmed that there were effects of phonetic context. For the second nasal formant and spectral center of gravity in particular, effects of left and right context differed by nasal consonant and also by syllabic position. For /m/, there were larger effects of following context in onset position and, for /n/, there were larger effects of preceding context in coda position. This is partly in accordance with previous findings that found that /m/ had larger degrees of coarticulation with the following segment than /n/ in onset position (Su et al., 1974) and that articulatory timing mechanisms in nasal consonants differ by syllabic position (Byrd et al., 2009; Krakow, 1993). Su et al. (1974) suggested that /m/ displayed larger degrees of lingual coarticulation than /n/ because there is no articulatory target for the tongue in /m/, whereas in /n/ the tongue is constrained to an alveolar position. It now seems that this finding does not generalize to coda position, perhaps due to the relative weakness of coda /n/ in Dutch. Word-final /n/ in weak syllables is often elided in verb and plurality suffix *-en* such as in the verb *lopen* (/loːpə/ 'walking'). In spontaneous speech, the final /n/ in the plurality suffix is only realized 2.5% of the time and only 35.0% in read speech (Silva et al., 2003). Previous research has also shown that /n/ shows an asynchrony in articulatory timing in codas, with the tongue-tip and velum gestures occurring earlier, i.e., during the articulation of the previous vowel (Byrd et al., 2009). It is possible that this timing asynchrony also affects the nasal murmur.

Current results showed larger effects of lingual coarticulation within the syllable; /m/ showed larger effects of following context in onsets and /n/ showed larger effects of preceding context in codas. Similar syllable-boundary effects on labial coarticulation were found for fricative consonants from the same telephone dialogues (Smorenburg & Heeren, 2020). This seems to indicate that there is more resistance to coarticulation across syllable boundaries, although other studies indicate

that the effect of prosodic boundaries on coarticulation is generally small or absent (e.g., Cho & McQueen, 2005; Hardcastle, 1985).

In the speaker classification in experiment II, we found that /m/ outperformed /n/, /m/ onsets outperformed /m/ codas, and /n/ codas outperformed /n/ onsets (although differences between linguistic conditions were considered minor given they are of the same size as variations due to random sampling of training and test data within conditions). Better speaker classifications indicate that more between-speaker variation was present in those linguistic contexts. Linking the results from experiment II to those found for experiment I, it seems to be the case that conditions showing larger effects of phonetic context, i.e., onset /m/ and coda /n/, had more between-speaker variation and therefore slightly better speaker-classification accuracies. The increased between-speaker variation in these linguistic contexts is assumed to arise from between-speaker variation in the coarticulatory movement. These results are in accordance with earlier work on fricatives which used a subset of the speakers in the current study; speaker classification was only slightly better from fricatives with labial coarticulation than from fricatives without labial coarticulation (Smorenburg and Heeren, 2020). These results provide some further evidence for the hypothesis that articulatory weak parts of speech such as codas and speech sounds in contexts subject to coarticulation, show more between-speaker variation (cf. He et al., 2019) and can therefore be more speaker-specific (Smorenburg & Heeren, 2020).

For forensic speaker comparisons, results indicate that considering the specific linguistic contexts nasals are sampled from only leads to minor differences in speaker-classification accuracy using regularized MLR. In practice, these differences seem too insignificant to be concerned about in forensic case work. Especially since material in forensic casework is usually scarce and only sampling from specific contexts would add a dimension of difficulty. Moreover, the standard in forensic casework has become to use likelihood ratios (LR) in the Bayesian framework, which estimates the likelihood of the evidence assuming that two speech samples come from the same speaker relative to the likelihood of the evidence assuming that two speech samples come from different speakers. This type of analysis was not used in the current

work because of the relatively small number of speakers and because LR models do not allow for the inclusion of interactions with linguistic factors in the modelling of acoustic-phonetic features. It is unclear how the current results would compare to LR speaker classification, but one study reports that small differences in speaker-classification obtained with multinomial logistic regression are not maintained in an LR analysis (Heeren, 2020a). It was suggested that this may be caused by differences in the weighting of between- and within-speaker variation in these two methods. Interestingly, the non-zero coefficients from the regression model indicated that different predictors were included per speaker. This indicates that different combinations of predictors were successful in discriminating different speakers. Moreover, not a single measure was included across all speakers; Figure 3.3 shows that the acoustic measures that were included for most speakers, i.e., N2 and N3, were both included for 45 out of 54 speakers. For forensic speaker comparisons, this may indicate that combining different measures within segments may be crucial for optimizing speaker discrimination in a large set of speakers. Recent studies using forensic methods, that is LR analysis, are also observing speaker variability in speaker predictors (Lo, 2021; Wang et al., 2021).

One limitation of the current work is the possible recording-related variability in the acoustics due to the relatively uncontrolled recording circumstances; speakers conversed on the telephone in their home environment and speech was intercepted via a wiretap. Regarding possible effects of speech channel, previous research has shown that vowel formants that are not in the direct vicinity of the lower and upper limits for the telephone band, i.e., F2 and F3, are generally not affected by the telephone band (Byrne & Foulkes, 2004). However, we cannot claim that there was no influence of background noise or the specific recording device on the speaker-classification accuracies in particular. Recording variability could be controlled by performing by-recording normalization on acoustic measurements, but since the variable 'recording' shows high overlap with 'speaker', we chose not to do this. Recording effects were somewhat controlled by excluding tokens with audible background noise, and all data were wiretapped in the same way. Moreover, the current work was not so much concerned with absolute

speaker-classification accuracies, but rather with relative differences in accuracy between linguistic contexts.

3.7 Conclusion

Nasals have often been cited to be rather speaker specific (e.g., Amino & Arai, 2009; Rose, 2002). In the current exploratory work, we investigated whether nasal consonants /n/ and /m/ show effects of phonetic context and syllabic position in their acoustics and tested whether speaker classifications with acoustic-phonetic features were dependent on the nasals' linguistic environment. Nasal consonants were found to display effects of phonetic context, which differed by nasal consonant and by syllabic position. Speaker-classification results seem to indicate that there might be a positive relation between the degree of coarticulation and speaker-classification accuracy. These results suggest that there are between-speaker differences in the degree and timing of co-articulatory gestures, which may add speaker-specific information from articulatory behavior.

Supplementary materials

The supplementary materials for this article can be found online at: https://asa.scitation.org/doi/suppl/10.1121/10.0005845

CHAPTER 4

Effects of the landline telephone filter

Abstract

Previous work on telephone speech investigating effects of phonetic context and syllabic position on acoustics and speaker variation found different effects for Dutch fricatives /x/ and /s/ (Smorenburg & Heeren, 2020). This was attributed to the narrowband telephone filter cutting of spectral energy from /s/, not /x/. Using English data that was simultaneously recorded as broadband and telephone speech, this work shows that linguistic effects are affected by the telephone filter. Additionally, linguistic context effects on speaker variation again show

that fricatives in labial contexts contain more between-speaker variation than fricatives in non-labial contexts. However, this was only the case for following labial context, not preceding labial context, and no substantial difference was found between /s/ in coda and onset position.

This chapter has been submitted and parts of this chapter have been presented at:

Smorenburg, L., & Heeren, W. (2021). Effects of speech channel on acoustic measurements and speaker discrimination from /s/. In *29th conference of IAFPA*. Marburg, Germany: University of Marburg.

Smorenburg, L., & Heeren, W. (2022). The effects of linguistic contexts on the acoustics and strength-of-evidence of /s/. In *30th conference of IAFPA* (pp. 13–14). Prague, Czech Republic: Charles University.

4.1 Introduction

Social and idiosyncratic information in speech play a large role in everyday communication. Perception studies have for instance shown that sentence interpretation is dependent on (inferred) speaker information (Van Berkum, Van Den Brink, Tesink, Kos, & Hagoort, 2008). Speech acoustics can be used to characterize individual speakers and in forensic speaker comparisons (FSC), the idiosyncratic information in voices is analyzed, and may serve as evidence in court. To improve FSC, researchers have been trying to establish what factors, both linguistic and extra-linguistic, affect the idiosyncratic information in speech.

Different speech segments hold different amounts of idiosyncratic information. Namely, vowels typically contain more speaker information than consonants (e.g., Van den Heuvel, 1996), although see Schindler and Draxler (2013). Amongst the consonants, nasals and fricatives contain more speaker information than other consonants (Kavanagh, 2012; Van den Heuvel, 1996). Moreover, there is some evidence that the same segment might also contain slightly different amounts of speaker information in different linguistic contexts or positions (e.g., see Heeren, 2020a on word class; McDougall, 2004 on lexical stress; Smorenburg & Heeren, 2020 and Su, Li, & Fu, 1974 on phonetic context and idiosyncrasies in coarticulation). On the one hand, some linguistic contexts and positions may result in lower within-speaker variation which may serve to increase speaker-specificity, for example in content words (Heeren, 2020a) and stressed vowels (McDougall, 2004). On the other hand, the degree and timing of coarticulatory movements and reduction may be specific to speakers (cf. Nolan, 1983, Chapter 3), thus increasing between-speaker variation (Smorenburg & Heeren, 2020; Su, Li, & Fu, 1974).

One major concern in FSC is the effects of telephone filters on speaker discrimination. In the Netherlands, wiretapped telephone conversations are common in FSC and it is therefore relevant to know

how telephone filters affect speech acoustics and speaker discrimination. Although the effects of telephone filters on speech acoustics have previously been investigated for some vowels (Byrne & Foulkes, 2004; Künzel, 2001), less is known about their effect on consonants. Given that some consonants, such as sibilant fricatives, have their spectral peak at frequencies outside of the upper limit of most telephone filters, the effect of telephone filters may be high for some consonants. In fact, it has been observed that fricative discrimination in narrowband telephone signals can be difficult (Bessette et al., 2002). Sibilant fricative /s/ in particular has a spectral center of gravity above 7 kHz in some groups of speakers (Munson, McDonald, DeBoe, & White, 2006). Given that /s/ acoustics can convey some information about speaker identity, the telephone filter is expected to have an effect on the idiosyncratic information in /s/.

Previous research on fricatives /s/ and /x/ showed that /s/ still contained significant amounts of idiosyncratic information, even in a landline telephone bandpass of 300 – 3,400 Hz (Smorenburg & Heeren, 2020). Dutch /s/, however, has lower-frequency spectral characteristics than English /s/, which could mean that less idiosyncratic information is available for English /s/ in narrowband signals. Spectral characteristics from fricatives are furthermore strongly affected by labial coarticulation (e.g., Koenig, Shadle, Preston, & Mooshammer, 2013; Munson, 2004), which seemed to affect the speaker-specificity of Dutch fricatives in systematic ways (Smorenburg & Heeren, 2020). The current work investigated effects of linguistic context on the acoustics and speaker variation of British English /s/, also considering effects of and interactions with the landline telephone filter. Although the signal characteristics of landline signals are not entirely representative of the mobile signals that are commonly used in modern communications, the band pass of landline filters is still relevant in the forensic context.

4.1.1 Fricative /s/ acoustics

The alveolar fricative /s/ is articulated by making a narrow constriction at the alveolar ridge. This creates a turbulent airflow which results in an acoustic signal with aperiodic frication noise (Stevens, 2000). This frication noise predominantly reflects the resonance characteristics of the anterior cavity, which, for /s/, is the space between the alveolar constriction and the lips (Stevens, 2000). The smaller that space, the higher the frequency of the frication noise. The alveolar sibilant /s/, for example, has higher-frequency frication noise than post-alveolar /ʃ/ (e.g., Jongman, Wayland, & Wong, 2000). This difference in anterior cavity size is also reflected in effects of sex; male speakers generally have a larger vocal tract and thus lower /s/ frequencies than female speakers (Li et al., 2016; Schwartz, 1968). Cross-linguistic differences have also been attested. Speakers of Dutch, e.g., have laminal articulations of /s/ where the constriction is made with the tongue front/blade. This is different for speakers of English or French where the constriction is apical, i.e., made with a pointed tongue tip. As a result, the anterior cavity in /s/ articulation is larger for speakers of Dutch, resulting in a lower center of gravity in Dutch than in English (Collins & Mees, 1984; Quené, Orr, & Van Leeuwen, 2017). Considering the differences in phoneme inventories and articulatory settings, there are some potentially relevant differences between English and Dutch. For example, it has been observed that Dutch generally has more muscular tension in the lips, whereas in British English the lips are less active, resulting in the stereotype of a 'stiff upper lip' (cf. Collins & Mees, 1984). This goes hand in hand with the vowel inventory: Dutch has more rounded vowels than English, which can be front or back, whereas English round vowels are all back. This is relevant for the effect of phonetic context in this work, as lengthening of the anterior cavity can be achieved by both protruding the lips or having a more posterior tongue constriction in fricative articulation.

Phonetic context also affects the size of the anterior cavity; protruding the lips in anticipatory lip-rounding lengthens the anterior cavity and lowers the frication noise (e.g., Koenig, Shadle, Preston, & Mooshammer, 2013; Munson, 2004; Shadle & Scully, 1995). Another linguistic effect that influences fricative acoustics is syllabic position, although there are contradicting reports, specifically for /s/. Generally

speaking, consonants in coda position are articulated with less effort than consonants in onset position (Ohala & Kawasaki, 1984). For fricatives, coda reduction is observed for fricatives in general but not consistently across temporal and spectral measurements for /s/ in particular (Cunha & Reubold, 2015; Redford & Diehl, 1999; Solé, 2003).

Previous research has shown that there can be cross-linguistic differences in patterns of coarticulation. Most generally, it has been hypothesized that languages can be characterized by the direction of coarticulation. For example, it has been claimed that French shows predominantly anticipatory coarticulation, whereas English shows predominantly carry-over coarticulation (Hoole, Nguyen-Trong, & Hardcastle, 1993). However, acoustic evidence only varyingly corroborates this hypothesis. For example, Magen (1997) found no evidence for more carry-over than anticipatory V-V coarticulation in English. Niebuhr, Clayards, Meunier, and Lancia (2011), on the other hand, found that sibilant sequences in English show exclusively carry-over place articulation, whereas French showed both carry-over and anticipatory place assimilation. Looking at labial coarticulation specifically, many studies show generally large effects of anticipatory labialization in English (e.g., Bell-Berti & Harris, 1982; Koenig et al., 2013; Munson, 2004; Nittrouer & Whalen, 1989; Soli, 1981). Not many studies focus on carry-over labialization in English, although some studies do investigate the combined effect of carry-over and anticipatory labialization in VCV sequences (e.g., Shadle & Scully, 1995). Due to these crosslinguistic differences, it is possible that previous findings on the context-dependency of speaker variation in Dutch /s/ do not generalize to English.

4.1.2 Idiosyncratic information in /s/

Amongst the consonants, nasals and fricatives seem to contain the highest amounts of idiosyncratic information. Nasals are often reported to be robust to many contextual influences and therefore show relatively little within-speaker variation, which makes them relatively speaker specific

(Rose, 2002). Fricatives, particularly /s/, also carry social information about the speaker and therefore have relatively high between-speaker variation, which also makes them relatively speaker-specific. Regarding the between-speaker variation in fricatives, it has been shown that social class and gender significantly affect /s/ productions (Stuart-Smith, 2007) and that even sexual orientation is encoded in and perceived from the acoustics of /s/ (Munson et al., 2006; Tracy, Bainter, & Satariano, 2015). For speakers of Dutch, /s/ acoustics have also been shown to contain information about ethnicity (Ditewig, Smorenburg, Quené, & Heeren, 2021) and region (Ditewig, Pinget, & Heeren, 2019). These social and linguistic variables, along with the acoustic reflection of the speaker's vocal tract size, all contribute to this sound being relatively speaker-specific, which makes it a potentially useful sound in FSC.

There also seem to be differences in the amount of idiosyncratic information within speech sounds that can be related to prosodic structure and phonetic context. Regarding prosodic structure, it seems that speech articulated with more effort is more precise and therefore more consistent within (and also between) speakers. For example, content words seem to contain slightly more speaker information than function words (Heeren, 2020a) and stressed vowels seem to contain slightly more speaker information than unstressed vowels (McDougall, 2004). Conversely, less articulatory effort allows for more freedom in reduced forms. He, Dellwo and colleagues studied between-speaker variation in intensity and formant contours of syllables and found more variation in the second half of syllables, i.e., towards the syllable coda (He & Dellwo, 2017; He, Zhang, & Dellwo, 2019). This was explained by the relative articulatory freedom of codas, whereas realizations of onsets are more constrained. It has also been observed that idiosyncrasies exist in coarticulation (cf. Nolan, 1983, Chapter 3). With regards to /s/, fricative realizations are highly dependent on contextual labialization and /s/ in labial contexts generally showed slightly more between-speaker variation than fricatives in non-labial contexts (Smorenburg & Heeren, 2020). Similarly, this was shown for nasal consonants /n/ and /m/ in contexts with coarticulation (e.g., Su, Li, & Fu, 1974).

4.1.3 Telephone signals and telephone speech

Speech transmitted over telephones loses acoustic information due to the limited band passes used in telephony. Telephone signals can be subdivided into two main types; landline and mobile signals. Landline telephone signals have a narrow band pass of about 300 – 3,400 Hz, meaning that spectral energy below 300 and above 3,400 Hz is strongly attenuated or lacking altogether (Künzel, 2001). Although some mobile signals have a very similar band pass to landline signals, the signal is much less stable. For example, the Adaptive Multi-Rate (AMR) narrowband codec (the compression technology used in 2G and 3G signals) that was standardized for the Global System Mobile Communication (GSM) network has a similar band pass of 200 – 3,400 Hz (Bessette et al., 2002). However, its bit rates can change rapidly, which can lower the upper frequency cut-off from 3,400 Hz to 2,800 Hz (Guillemin & Watson, 2006). More modern cellular technology uses much wider bandwidths, e.g., the Adaptive Multi-Rate Wideband (AMR-WB) codec used in 4G signals covers a 50 – 7,000 Hz band pass and thus provides better fricative differentiation (Bessette et al., 2002). For speech sounds with high-frequency characteristics such as /s/, this upper cut-off captures more information than landline signals and mobile predecessors. However, the AMR-WB still has a varying bit rate depending on channel conditions; the signal changes to half-rate when channel conditions are considered good based on harmonics-to-noise ratios (Bessette et al., 2002).

In the Netherlands, telephony providers are legally required to make wiretapping available for both landline and mobile telephone signals (Van de Pol, 2006). When a call is wiretapped, an authorized third party can listen in on the call and record it. Such recordings may be processed for police investigations. As a result, much of the speech material in forensic casework consists of wiretapped telephone conversations.

Effects of telephone signals on speech can be both signal-related and behavioral in nature. Signal-related effects have mostly been described for vowels in landline signals; vowel formants that are situated

near the lower telephone cut-off are affected in landline signals (Künzel, 2001) and in mobile signals (Byrne & Foulkes, 2004). Specifically, the measurements of F1 values might shift upward. In automatic speaker recognition, which uses more holistic speech features such as Mel-frequency cepstral coefficients, mismatches in speech channel also have significant effects on speaker discrimination when it concerns telephone versus studio recordings (Van der Vloed, Kelly, & Alexander, 2020). For auditory-acoustic analysis however, where linguistic-phonetic speech features are examined and which is more common in forensic casework across the globe (Gold & French, 2011, 2019), it is not yet clear what effects different kind of telephone filters may have on consonants in particular. Some previous research has attempted to replicate telephone filters by using a 500 – 4,000 Hz frequency range for extracting measurements from /s/ in broadband signals (Kavanagh, 2012). Using discriminant analysis, Kavanagh (2012) found similar speaker-classification accuracies for the simulated telephone filtering condition compared to a broadband condition (500-8,000 Hz). When using likelihood-ratio testing, however, better speaker classifications were obtained in the narrowband compared to the broadband filtering condition, which Kavanagh remarked was notable. As will become clear in the current work, telephone signals are not only different from broadband signals in their frequency range, but generally show different spectral shapes due to noise and compression mechanisms in the telephone codec. It is therefore necessary to use actual telephone signals to test the effect of telephone filters on /s/.

Regarding behavioral effects, a speaker's "telephone voice" is often subject to the Lombard effect, i.e., the increase of vocal effort in the presence of noise (Junqua, Fincke, & Field, 1999). When conversing over the telephone speakers cannot be seen by the listener, meaning that hand gestures and facial expressions cannot be used and acoustic means might replace them. In a study on the use of intonation in turn-taking in telephone versus face-to-face conversations, differences were found in speakers' pitch ranges, where a larger pitch range was associated with holding the turn in face-to-face conversation but with changing the turn in telephone conversation (Oliveira & Freitas, 2008). Although perception results subsequently showed that intonation alone did not

seem a sufficient cue for turn transition, this study confirms that speakers display different uses of intonation in their production across speech conditions.

To summarize, telephone speech behavior may differ from other speech behavior and telephone signals are limited in their frequency range, which can affect fricative discrimination. It is not yet clear how the loss of acoustic information may impact the speaker information in /s/, although some research has shown that a limited frequency range does not necessarily lead to decreased speaker classification for /s/.

4.1.4 Research questions

This study investigated the effects of the telephone filter and of phonetic context and syllabic position on the acoustics and speaker characteristics of /s/. Previous research has shown that acoustic-phonetic features from Dutch /s/ still contain significant amounts of idiosyncratic information in landline telephone recordings (Smorenburg & Heeren, 2020). That speech corpus, however, only contained telephone signals and, therefore, did not allow for a direct comparison between telephone and studio channels (i.e., high-quality recordings). Here, we investigate the effect of the landline telephone filter on /s/ in direct comparison with simultaneously recorded studio speech in British English data from the West Yorkshire Regional English Database (WYRED; Gold, Ross, & Earnshaw, 2018). The acoustics of /s/ may be assumed to be highly affected by the telephone wiretapping because its spectral peak falls outside of the telephone band. However, it is possible that the between-speaker variation in spectral peak values is also (partly) reflected in the weaker spectral energy at lower frequencies. Moreover, some acoustic-phonetic measurements might be more robust to telephone filters than others.

Additionally, previous research showed that phonetic context and syllabic position affect the acoustics and speaker information in fricatives from Dutch landline telephone speech (Smorenburg & Heeren, 2020). This work further investigated the possible interactions between

linguistic effects and signal bandwidth and the generalizability of previous Dutch results across languages. It is predicted that English /s/ will show effects of both contextual labialization and syllabic position. Based on the hypothesis that English is a carry-over language, it is expected that carry-over labialization effects will be larger than anticipatory labialization effects. Moreover, given that English /s/ is apical and therefore has higher-frequency spectral characteristics than Dutch /s/ (Quené, Orr, & Van Leeuwen, 2017), it is expected that linguistic effects will only be observed in the broadband studio recordings and not, or to a lesser extent, in the narrowband telephone recordings.

4.2 Method

4.2.1 Materials and segmentation

Materials were taken from the West Yorkshire Regional English Database (WYRED; Gold et al., 2018). This corpus contains four different speech tasks from male speakers from three different regions in Yorkshire, England. For this study, Task 2 was selected, which is a telephone conversation between a suspect (played by the participant) and an accomplice (played by a researcher in another room). Visual speech maps were used to elicit certain speech sounds. These conversations were simultaneously wiretapped from the landline telephone as well as recorded over a microphone placed in front of the participant. Participants performed Task 2 once, meaning that the within-speaker variation in this data is derived from a single 15-min telephone conversation. Since dialect was not of interest to the current study, only speakers from a single region were included, namely all 60 speakers from the Wakefield region (mean age = 21.15, *SD* = 2.85, range = 18–30).

The orthographic transcriptions that are available for each conversation were used in a forced-alignment protocol to generate segmentations at the phonemic level. To achieve the best possible

accuracy, the high-quality studio recordings were used for this. However, given the (semi-)spontaneous nature of the speech, the resulting automatic alignments were often inaccurate and needed manual correction. Target intervals were therefore estimated on four exclusion criteria and boundaries manually corrected until all speakers had at least 100 usable /s/ tokens. Tokens were excluded when they (1) were not auditorily and visually identifiable by the waveform and spectrogram as a sibilant fricative (due to reduction or elision), (2) contained interfering ambient noise or speech by the interlocutor, (3) contained laughter, or (4) contained accent imitations or other vocal imitations such as impersonations. All tokens were manually corrected and labelled on syllabic position and on whether preceding and following speech sounds were labial (consonants: /p, b, m, w/, vowels: /u, ʊ, o, ɔ, ɒ/, and (partially) rounded diphthongs: /əʊ, ɔɪ, aʊ/ were coded as labial, all other sounds were as non-labial). Diphthongs were coded as rounded irrespective of whether the rounding was immediately adjacent to the fricative (cf. temporal patterns of lip-rounding: Bell-Berti & Harris, 1982).

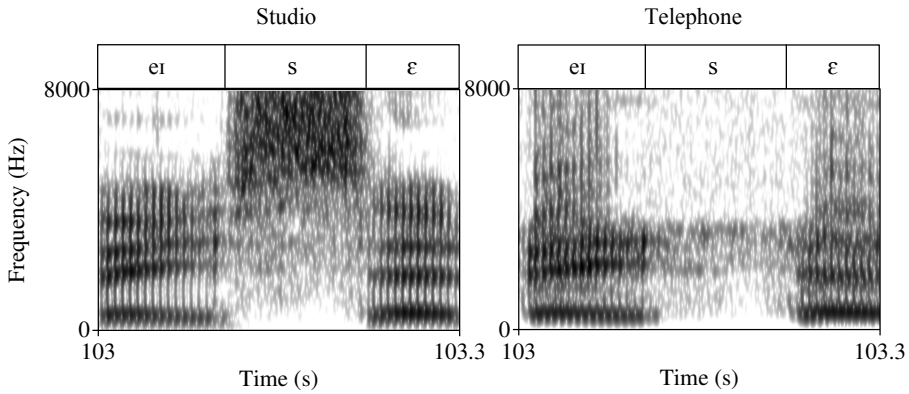

Figure 4.1: *Spectrogram for the same /s/ token in studio versus telephone channel*[6].

[6] Note that, for some tokens (such as this one), there is a very slight misalignment between the studio and telephone recording. This should only have minimal effects

For the analyses focusing on the effects of signal type and bandwidth, exactly 100 tokens per speaker ($N = 60$) were included in the analyses. For the linguistic context analysis, only speakers with at least 10 tokens per factor level were included in the analysis ($N = 55$, see Table 4.1 for the number of tokens per factor level).

Table 4.1: *Number of tokens per factor level with statistics by speaker*.

		Syllabic position		Left context		Right context	
	All	Onset	Coda	Non-labial	Labial	Non-labial	Labial
Total N	6,634	3,865	2,769	5,704	930	5,416	1,218
M	121	70	50	104	17	98	22
(*SD*)	(26)	(16)	(14)	(23)	(5)	(21)	(10)
Range	91-194	42-114	19-87	80-169	10-32	61-146	10-48

4.2.2 Acoustic analysis

Before extracting acoustic measurements for the target intervals, the simultaneously recorded telephone and studio recording for each speaker ($N = 60$) were manually aligned where needed. Given the different signal characteristics per condition, different frequency ranges were used when taking acoustic-phonetic measurements (see Table 4.2). Low frequencies up to 550 Hz were excluded to lessen the effect of ambient background

because spectral moments were measured over the middle 50% of each /s/, which is relatively stable.

noise and intruding voicing (cf. Koenig et al., 2013; Smorenburg & Heeren, 2020). For the studio condition, an upper limit of 8 kHz was chosen, because most phonetic contrasts are captured in this bandwidth in male adult speakers (although higher-frequency information plays a role in sibilants, e.g., see Monson, Lotto, & Story, 2012, the phonetic contrast between sibilants is present in the signal up to 8 kHz in male adult speakers, e.g., see Holliday, Reidy, Beckman, & Edwards, 2015). For /s/, the spectral region of interest is the one that is associated with the anterior cavity peak, found around 5 ~ 7 kHz (Koenig et al., 2013). For the studio recordings, acoustic-phonetic measurements were also taken over the 550 – 3,400 Hz range (similar to Kavanagh, 2012, see Appendix A), to see if measurements from studio and telephone recordings differed when the frequency range of measurement was equal.

Table 4.2: *Signal characteristics per channel*

	Studio	Telephone
Sampling rate [samples/s]	44,100	44,100
Frequency range [Hz]	0 – 22,050	300 – 3,400[a]
Measurement range [Hz]	550 – 8,000/3,400	550 – 3,400

[a] Telephone signal is present from 0 – 4,000 Hz, but is attenuated outside of the telephone filter of 300 – 3,400 Hz.

Four spectral moments, the spectral peak and spectral tilt were measured over the middle 50% of each /s/ token. Spectral moments capture the overall spectral shape and are often used to describe fricatives, particularly sibilants (e.g., Forrest, Weismer, Milenkovic, & Dougall, 1988; Jongman et al., 2000; Shadle & Mair, 1996). The first spectral moment (M1) is the spectral center of gravity and, in Praat (Boersma & Weenink, 2020), is computed as the mean frequency of the

spectrum in Hz. The second moment (M2) is the spectral dispersion and is computed as the variance around M1 in Hz. The third moment (L3, but M3 is also seen in the literature) is the skewness, which is a coefficient that indicates how much the spectral shape below M1 differs from that above M1, i.e., whether it leans to the left (lower frequencies) or right (higher frequencies). Lastly, the fourth moment (L4) is the kurtosis, which indicates how much the shape of the spectrum differs from a Gaussian shape, i.e., how peaked the distribution is.

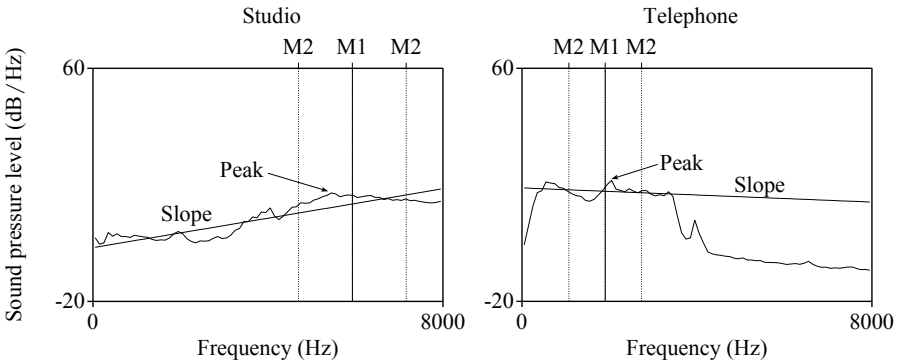

Figure 4.2: *Averaged spectra for one randomly selected speaker (WYRED speaker 041, N = 100) in the studio versus telephone channel. Measurements were taken over the 550 – 3,400 Hz range for the telephone channel and over the 550 – 8,000 Hz range for the studio channel.*

The spectral peak captures the frequency of the amplitudinal maximum in the power spectrum. For sibilant fricatives, the peak associated with the anterior cavity resonance (at 5 ~ 7 kHz: Koenig et al., 2013) falls outside of the telephone band; instead, some other amplitudinal maximum within the telephone band will be selected, which might be rather random. The spectral peak measurement should capture roughly the same type of information as M1, i.e., the size of the vocal

tract and in particular the anterior cavity, and these measurements therefore correlate highly (e.g., Ditewig et al., 2021). However, whereas the spectral peak is tied to a specific spectral event, M1 is not. M1 is highly dependent on the frequency range of measurement. L3 and L4 should also be highly affected by speech channel, as the available spectrum in the narrowband telephone filter will have a different shape than that in the broadband studio recording due to the telephone cut-offs, signal noise, and possibly the telephone codec's compression (see Figure 4.2).

Given the possible relevance of (co)articulatory information in /s/, it has been proposed that acoustic analyses of /s/ should include dynamic acoustic measurements (Koenig et al., 2013). M1 was therefore also measured dynamically, in five non-overlapping windows, each 20% of the total duration of each /s/ token. These five measurements across time were then captured in a polynomial function. Both quadratic (R^2 = 0.81, R^2 adjusted = 0.62) and cubic (R^2 = 0.92, R^2 adjusted = 0.67) functions were estimated; the cubic function was not a significantly better fit to the data than the quadratic one: $\chi^2(1) = 1.15$, $p = 0.28$. For the statistical analysis, the dynamic measures therefore consisted of two coefficients (the linear and quadratic terms). The intercept of the function was excluded because that value is conceptually the same measurement as the M1, only differing slightly in measurement window. Dynamic coefficients might be slightly more robust to speech condition because they capture the relative movement, rather than the absolute values, of M1 across the duration of /s/.

Our last measurement, the spectral tilt, refers to the overall slope of the power spectrum in the specified frequency ranges of measurement and is computed as a logarithmic regression fitted to the power spectrum using least squares. This measurement does not reflect a specific spectral event but rather a trendline of the spectrum. From Figure 4.2 it seems that the averaged spectrum in the telephone condition has a very different shape relative to the same data in the studio condition. Therefore, it is expected that, across all our data, the spectral tilt measurement is also highly affected by the telephone filter.

4.2.3 Statistical analysis

Statistical analysis was performed in R version 4.0.1. (R Core Team, 2019) and consisted of four parts. First, Pearson's correlation coefficients between acoustic-phonetic features within conditions were computed to see which features reflected the same type of information and to see which features could be combined in a follow-up speaker-discrimination test.

Second, linear mixed-effects modelling (LME) was used to firstly assess acoustic effects of the different recording types and bandwidths (Telephone 550 – 3,400 Hz versus Studio: 550 – 8,000 Hz versus Studio: 550 – 3,400 Hz) and secondly to assess the acoustic effects of Phonetic Context (NON-LABIAL, LABIAL) and Syllabic Position (ONSET, CODA) on eight acoustic-phonetic features. In the random structure of each model, a by-speaker intercept and by-speaker slopes over the fixed factors were assessed. Models were built automatically using backward stepwise elimination with BIC (Bayesian Information Criterion) estimation of random and fixed effects using function *buildmer()* from R package 'buildmer' (Voeten, 2020)[7]. The *p*-values for significance were Bonferroni-corrected for the number of acoustic measurements ($N = 8$), as several acoustic measures are extracted from the same recording and the results from these different models can therefore not be assumed to be independent.

Third, to assess speaker-specificity by recording type and bandwidth, as well as by phonetic context and syllabic position, linear discriminant analysis (LDA) was used, utilizing R package 'MASS' (Venables and Ripley, 2002). LDA is commonly used to classify a variable with multiple classes and, in speech science, is often used for

[7] Although one might expect truncated distributions for some acoustic measurements in the telephone recording, visual inspection of histograms did not show truncated distributions for any measurements. Only the spectral peak measurement showed a highly non-normal distribution, with visible peaks in the distribution at 1,500 – 2,000 Hz and 3,000 – 3,400 Hz. This indicates that, since the actual spectral peak of /s/ could not be captured due to the limited telephone band pass, other spectral peaks were found (predominantly in one of the aforementioned frequency regions).

automatic speech recognition in which speech is classified into phonetic classes (e.g., Viszlay, Juhár, & Pleva, 2012). In the current analysis, it is used to classify speakers using the acoustic-phonetic features as predictors. Speaker classifications were first computed over all data (N = 60, n = 6,000), disregarding linguistic contexts, to assess which features and combinations of features performed best at discriminating speakers and to assess the effect of the signal type and bandwidth on speaker classifications. To achieve a direct comparison, the same tokens (in each condition) were selected for the training and test data. Specifically, the first 70% of data by condition and by speaker were used as training data and the last 30% were used as test data. This way, any differences in results may be wholly attributed to signal-related effects, without potential sampling effects or other confounding variables. Before running the LDA, correlations between acoustic-phonetic features were inspected, within each of the two recording conditions. Highly-correlating predictors (r > .60) should not be entered into an LDA model together as multi-collinearity can lead to imprecise model coefficients (Klecka, 1980). The predictor set of the best-performing LDA model was used in subsequent analyses on linguistic contexts.

4.3 Results

4.3.1 Acoustic effects of the landline telephone

As expected, the M1, spectral peak, and spectral tilt measurements were highly correlated in both the studio and telephone conditions (see Table 4.3). High correlations (r > .60) were also found between M1 and L3 and between L3 and L4, although not consistently across conditions. Looking at the same measurement across conditions (see the diagonal in Table 4.3), only weak correlations were found. This suggests that the measurements in the telephone condition reflect different acoustic information than measurements in the studio condition, and also suggests large effects of the telephone filter will be found in LME modelling.

Table 4.3: *Pearson's correlations between acoustic measurements (df = 5,998) within the studio recordings (left of diagonal), within the telephone recordings (right of diagonal) and between the studio and telephone recordings (on diagonal). Significant correlations are indicated in bold.*

	Acoustic measure	Telephone M1	M2	L3	L4	$M1^{lin}$	$M1^{quad}$	Peak	Tilt
Studio	M1	**–0.44**	**–0.11**	**–0.23**	0.05	–0.03	**–0.38**	**0.78**	**0.90**
	M2	–0.05	**–0.12**	–0.07	**–0.29**	**0.13**	**–0.13**	–0.00	**–0.27**
	L3	**–0.71**	**–0.30**	**–0.08**	**0.81**	–0.02	**0.20**	**–0.23**	**–0.11**
	L4	**–0.09**	**–0.51**	**0.42**	**0.10**	–0.03	**0.08**	–0.00	**0.13**
	$M1^{lin}$	–0.02	**0.16**	–0.02	–0.05	**0.31**	0.03	–0.04	–0.06
	$M1^{quad}$	**–0.35**	**–0.23**	**0.45**	**0.14**	**0.16**	**0.13**	**–0.33**	**–0.28**
	Peak	**0.82**	–0.02	**–0.54**	–0.03	0.00	**–0.35**	**–0.26**	**0.63**
	Tilt	**0.72**	**–0.40**	**–0.15**	0.00	**–0.07**	–0.02	**0.47**	–0.01

Best-fitting LME models that assessed the effect of the telephone filter on acoustic measurements from /s/ are presented in Table 4.4. The highly-correlated measures M1, spectral peak, and spectral tilt all show large effects of the telephone filter with much lower values in the telephone than in the studio condition. According to expectations, skewness (L3) was more positive in the telephone than in the studio recording, indicating that the spectral shape is more left-leaning in the telephone condition. This makes sense, given that the telephone band has little spectral energy over 3,400 Hz. Somewhat counterintuitive, kurtosis (L4) was much higher in the telephone than the studio recording, indicating that the spectra in the telephone recording are more peaked than in the studio recording. This might be a result of the sharp cut-off of the spectrum at 3,400 Hz, resulting in a steeper peak even in the absence of the actual spectral peak (see Figure 4.2). The dynamic linear coefficient of M1 is the only measure that does not show a highly significant effect of channel, indicating that some of the dynamics of /s/

are similar across conditions. This might be related to the fact that /s/ is rather stable across time in the linear dimension. The quadratic coefficient, however, shows a large effect of condition, with a much larger dynamic movement in the studio recording. This indicates that the telephone recording does not fully capture the dynamic movement of /s/ across time.

This model was also run including a factor level for the simulation of the telephone signal, i.e., measurements taken in the studio recording using a 550 – 3,400 Hz bandwidth. Even when using this telephone-band frequency range, significant differences for all measurements (except M2, L4, and the linear coefficient of M1) were found between the studio and telephone recordings. This indicates that, although using a landline bandwidth on microphone-recorded materials makes it more similar to the landline signal, there are other differences between the conditions that are not strictly related to bandwidth.

Table 4.4: *Fixed effects in best-fitting linear mixed-effects models (N = 60, n = 6,000, default factor level = Studio: 550 – 8,000 Hz).*

	M1 [Hz]				M2 [Hz]			
Effect	*Est.*	*SE*	*t*	*p*	*Est.*	*SE*	*t*	*p*
(intercept)	5,022	74	67.8	***	1,268	12	104.6	***
Channel: Telephone	–2,943	95	30.8	***	–536	4	–135.6	***
	L3				L4			
(intercept)	0.19	0.04	4.6	***	3.85	0.99	3.9	***
Channel: Telephone	1.34	0.03	40.9	***	32.48	1.03	31.4	***

	dynamic M1^{linear} [Hz]				dynamic M1^{quadratic} [Hz]			
(intercept)	–4	26	–0.2	.8623	–739	26	–28.0	***
Channel: Telephone	81	25	3.2	.0013	545	32	17.0	***
	Peak [Hz]				Tilt [dB/decade]			
(intercept)	4,777	96	49.9	***	15.36	0.82	18.6	***
Channel: Telephone	–2,669	130	20.5	***	–11.86	1.22	9.7	***

Note. Bonferroni-corrected levels for significance: *p < 6.25e–03, **p < 1.25e–03, ***p < 1.25e–04

Regarding the random structure, best-fitting models included by-speaker intercepts for all acoustic measures. M1, spectral peak, spectral slope, and the two dynamic M1 coefficients also included by-speaker slopes over speech condition. There was a negative linear relationship between the by-speaker intercept and slope over speech condition reflecting that speakers who had higher-frequency /s/ productions showed larger acoustic effects of speech condition (see Figure 4.3), which is in line with expectations.

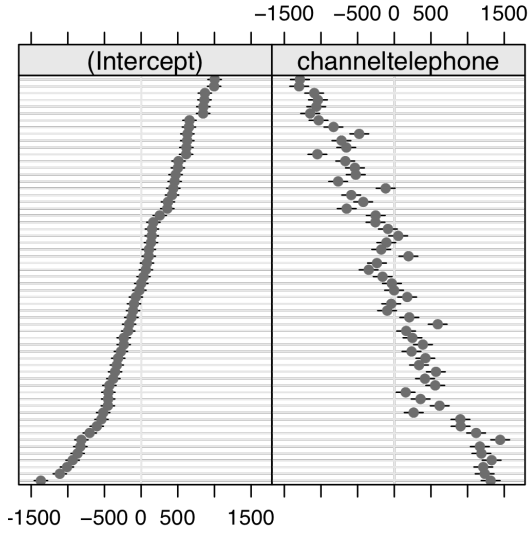

Figure 4.3: *By-speaker intercepts and slopes (N = 60) from the best-fitting LME model for M1 relative to the model intercept (5,022 Hz) and the effect of speech channel (–2,943 Hz).*

4.3.2 Acoustic effects of linguistic contexts (LME)

Starting with the measures related to the anterior resonance frequency, i.e., M1, spectral peak and tilt, these generally shows the expected effects in the studio recording. Namely, when preceding and following labial contexts or when tokens occur in coda position, the resonance frequency is lower (see Table 4.5). The effect of left context, i.e., carry-over coarticulation, is larger than that of right context, i.e., anticipatory coarticulation, which is in line with the hypothesis that English coarticulation patterns are predominantly carry-over (cf. Hoole et al., 1993). There was, however, an interaction between Right Context and Syllabic Position which showed that the effect of right labial context was larger for codas than onsets. Looking at the best-fitting models for the same speech data in the telephone recording, it can be seen that effects

are not maintained. Instead, effects in the telephone recording sometimes, but not as a rule, go in the opposite direction and generally do not resemble the patterns found in the studio recording. This indicates that detailed spectral information reflecting linguistic information is absent in the narrowband signal.

Table 4.5: *Fixed effects from linear mixed-effects models per channel*.

		Studio			Telephone		
	Effects	*Est.*	*SE*	*t*	*Est.*	*SE*	*t*
M1	(intercept)	5,190	77	67.3	2,075	32	64.2
[Hz]	Left context = LABIAL	–365	20	–18.7	112	10	10.6
	Right context = LABIAL	–94	22	–4.3	–31	12	–2.6
	Syll. Position = CODA	–200	15	–13.2	–1	8	–0.1
	Right x Syll. Position	–118	37	–3.2	68	20	3.4
M2	(intercept)	1,249	26	48.7	728	6	123.5
[Hz]	Left context = LABIAL				21	4	4.8
	Right context = LABIAL				25	4	6.2
	Syll. Position = CODA	35	6	6.0			
L3	(intercept)	0.04	0.09	0.5	1.58	0.07	21.9
	Left context = LABIAL	0.48	0.03	17.3	–0.33	0.08	–4.2
	Right context = LABIAL	0.19	0.03	7.4			
	Syll. Position = CODA	0.13	0.02	6.5	–0.19	0.06	–3.4
L4	(intercept)	3.75	0.37	10.2	34.63	1.50	23.2
	Left context = LABIAL	0.75	0.22	3.4			
Tilt	(intercept)	17.0	0.8	21.2	3.9	0.9	4.5
[dB/	Left context = LABIAL	–2.3	0.2	–10.7	2.0	0.3	6.6
decade]	Right context = LABIAL	–0.5			–1.9	0.3	–5.6
	Syll. Position = CODA	–2.2	0.2	–13.3	–0.6	0.2	–2.4
	Right x Syll. Position	–1.5	0.4	–3.6	2.5	0.6	4.3
$M1^{lin}$	(intercept)	–173	29	–5.9	17	17	1.0
[Hz]	Left context = LABIAL	144	33	4.4			
	Right context = LABIAL	–259	37	–7.0	–22	15	–1.5
	Syll. Position = CODA	492	25	19.4	136	10	13.3
	Right x Syll. Position	–276	62	–4.4	119	25	4.8
$M1^{quadr}$	(intercept)	–761	29	–26.1	–194	12	–16.7
[Hz]	Right context = LABIAL	–70	22	–3.2	–31	8	–3.8
	Syll. Position = CODA	116	15	7.7	26	7	3.8
	Left x Syll. Position				–105	19	–5.7
	Right x Syll. Position	185	37	–5.0			

4.3.3 Telephone effects on speaker discrimination (LDA)

In Table 4.6, the speaker-classification accuracies for the different LDA models are presented. With 60 speakers, chance level for classification accuracy was 1/60 (= 1.7%). Amongst the individual measures, the ones associated with the size of the anterior cavity, i.e., M1, spectral peak and tilt, performed best at discriminating speakers. M1 reached the highest accuracy, but spectral tilt was more robust to the telephone filter, possibly because spectral tilt – as a trend line fitted to the spectrum – is less tied to specific spectral events than M1. It seems that M1, spectral peak and tilt contain the most idiosyncratic information. Regarding the effect of condition, as expected, acoustic measures taken in the studio recording generally have more discriminatory power than acoustic measures taken in the telephone recording. The spectral tilt and the linear and quadratic terms of M1 showed only minor differences between speech channels and thus seem to be the most robust to bandwidth effects. Following the acoustic results, this was expected for the linear term; smaller acoustic effects should correspond to smaller effects of condition on the speaker classification. Despite the fact that spectral tilt and the quadratic term showed significant bandwidth effects on the acoustics, they seem relatively robust to these effects on the speaker classification, as we initially predicted based on the nature of these measurements.

Table 4.6: *LDA speaker-classification accuracies (in %) for independent features and combined features across recording type and bandwidth.*

Predictor (set)	Studio (550 - 8,000 Hz)	Telephone (550 - 3,400 Hz)	Studio (550 - 3,400 Hz)
M1	5.7	3.9	5.2
M2	4.1	2.9	3.3
L3	4.4	2.1	2.0
L4	3.1	2.0	2.4
$M1^{linear}$	2.9	3.0	1.9
$M1^{quadratic}$	3.1	2.9	2.3
peak	4.7	3.2	2.9
tilt	4.8	4.4	5.1
M1 + M2 + L4	9.7	5.6	6.2
M2 + L3	7.8	2.8	3.9
$M1^{linear}$ + $M1^{quadratic}$	4.2	2.9	2.5
M1 + M2 + L4 + $M1^{linear}$ + $M1^{quadratic}$	11.4	6.7	6.8
M2 + L3 + $M1^{linear}$ + $M1^{quadratic}$	9.4	4.2	4.0
M2 + L3 + tilt	12.9	6.5	6.8
M2 + L3 + $M1^{linear}$ + $M1^{quadratic}$ + tilt	14.5	7.5	7.6

The best-performing LDA model, with M2, L3, the linear and quadratic M1 coefficients and the spectral tilt as predictors had a 14.5% accuracy (95% CI [12.5, 15.8]) for the studio data and a 7.5% accuracy (95% CI [6.5, 8.8]) for the telephone data. For both the studio and telephone data, two linear discriminant (LD) functions were needed to account for ~80% of the between-speaker variance (studio: LD1 = 48%, LD2 = 31%; telephone: LD1 = 66%, LD2 = 15%). Mirroring the results on the models with individual predictors, the scaling coefficients from this combined model further indicate that LD1 in both the studio and

telephone data was largely explained by the spectral tilt, i.e., this was the most-contributing predictor. For LD2, L3 was the most-contributing predictor. The scaling coefficients also indicated that the dynamic linear and quadratic terms of M1 had the least discriminatory power, which is in accordance with previous research on /s/ in Dutch spontaneous speech (Smorenburg & Heeren, 2020). In spontaneous speech, /s/ dynamics are probably mainly determined by contextual effects, leaving little room for idiosyncrasies in articulatory dynamics (cf. Heeren, 2020b).

In the acoustic analysis, strong positive correlations were found between by-speaker intercepts and slopes over speech channel for M1, peak, tilt, and the dynamic M1 coefficients. This indicates that speakers with high average values for measurements also showed larger effects of speech channel. Despite this, there does not seem to be a strong relationship between the size of speech channel effects on the acoustics and on the speaker-classification accuracy on the speaker level. Pearson's correlations between speakers' channel effects on the acoustics and speaker-classification accuracy from the best-performing LDA model were not significant for any of the measures except for a weak correlation for spectral slope ($r = -.26$, $p < .05$).

As can be seen in Table 4.7, the speaker-classification accuracy is much better in the studio than in the telephone recording across all conditions. Linguistic context does not affect the speaker-classification accuracy save one exception: when the right phonetic context is labial, there is better performance than when the right phonetic context is non-labial. However, in the telephone recording, this difference is neutralized. All other differences between contexts are considered negligible because they are smaller than chance level accuracy (100% / 55 speakers = 1.82%).

Table 4.7: *LDA speaker-classification accuracies (in %) per factor level and recording type using the predictors from the best-performing model from Table 4.6.*

	Context	Studio (550 - 8,000 Hz)	Telephone (550 – 3,400 Hz)
All		14.5	7.5
Syllabic position	ONSET	14.4	7.6
	CODA	14.7	8.2
Left context	LABIAL	14.9	7.1
	NON-LABIAL	14.2	6.7
Right context	LABIAL	20.2	6.2
	NON-LABIAL	14.5	6.7

4.4 Discussion

Previous research has found that vowels' F1 measurements in telephone signals may shift upward by an average of 15% in landline signals and by an average of 29% (with up to 60% rises in F1 values) in mobile signals relative to studio recordings (Künzel, 2001; Byrne & Foulkes, 2004). As expected, because the sibilant fricative's spectral peak falls far outside of the upper limit of the narrowband telephone filter, the effect of landline filters on sibilant fricative /s/ acoustics is much larger than that of F1 for several vowels. This is, of course, mostly a reflection of the decrease in bandwidth in the telephone channel relative to the studio channel. Whereas the telephone filter only shaves off some of the spectral energy for F1, the average spectral peak for /s/ was 4,777 Hz in the studio recording (i.e., 1,377 Hz above the telephone band's upper limit), making it impossible to measure in the telephone signal.

These large effects of speech condition on the acoustics were reflected in the idiosyncratic information in /s/ as shown by the speaker

classification results; in the best-performing model, classification accuracy decreased by about half as a function of bandwidth. From these results we can conclude that the signal from 550 – 3,400 Hz does not capture much of the between-speaker variation that is present in the 550 – 8,000 Hz signal. This is in accordance with the observation that fricative discrimination in broadband signals is better than in narrowband signals (Bessette et al., 2002). However, the current results also show that some idiosyncratic information remains in /s/ from landline telephone speech, as speaker-classification accuracies on telephone speech are (at least slightly) above chance level in the LDA results (see also Smorenburg & Heeren, 2020).

Furthermore, for some acoustic measures, findings from the correlations, the acoustic analysis, and the speaker-classification analysis show interesting patterns. Spectral tilt, e.g., shows no correlation across channels ($r = -.01$, see Table 4.3), which is congruent with the acoustic analysis that shows large effects of speech condition on this measurement. In speaker classification, however, spectral tilt performs relatively well, with only a minor difference (0.4%) in classification accuracy between bandwidths. This implies that, while the measured tilt is significantly altered (lower in the telephone channel), the same amount of speaker information is available in the measurement. Another interesting observation is that the M1 and spectral peak measurements are highly correlated within recording types, even though the spectral peak that is usually targeted (often the spectral maximum around 5 ~ 7 kHz, Koenig et al., 2013) cannot be measured accurately in the telephone band. We expected that the peak measurement in the telephone recordings would therefore be rather random. However, its high correlation with the M1 measurement and the above-chance speaker-classification accuracy in the telephone recording seem to indicate that spectral peak measurements in telephone recordings still systematically capture some information about resonance properties in /s/. This is further corroborated by the distribution of spectral peak measurements, which shows a bimodal distribution (see footnote 2). This indicates that, when the actual spectral peak cannot be measured in the telephone band, another peak is found, not randomly, but predominantly in one of two specific frequency regions.

As for linguistic sampling context, British English /s/ acoustics show effects of contextual labialization and coda reduction in the studio recording, including an interaction between anticipatory labialization and syllabic context which showed more anticipatory labialization for codas. Interestingly, we find evidence for the hypothesis that English has predominantly carry-over coarticulation (Hoole et al., 1993), as effects of left context are larger than effects of right context, i.e., anticipatory coarticulation. In earlier work on Dutch, /x/ showed somewhat larger anticipatory coarticulation, also particularly in coda position (Smorenburg & Heeren, 2020). This might be indicative of other labial coarticulation patterns in British English versus Dutch, with the former being more carry-over and the latter more anticipatory in nature. Comparing linguistic effects on the acoustics across recording type, we see that acoustic effects are not maintained in the telephone recording. In fact, effects in the landline telephone recording are not similar to effects observed in the studio recording and are also not congruent with previous findings on linguistic effects on fricative acoustics; they do not seem to follow any discernable pattern relative to effects in the same speech data recorded over the studio recording. Remember that these results were obtained using landline telephone signals. Although landlines are still in use and therefore relevant, mobile signals are common in daily communications. Mobile signals differ from landline signals in that they can have varying bit rates and therefore varying bandwidths. Future work should consider also examining the effects of mobile signals on different speech sounds across linguistic contexts.

The linguistic effects generally seemed to have no effect on the amount of speaker information in /s/, with the exception of one phonetic context: /s/ tokens perform better when followed by labial segments, which is a context with increased between-speaker variation assumedly due to speaker-specific patterns in anticipatory labialization. This effect was only observable in the studio recording and seemed to be neutralized in the telephone recording.

Previous research has shown that listeners are generally less able to identify speakers over the telephone than over studio-recorded speech (Reynolds, 1995). Knowing that /s/ is a relatively speaker-specific consonant, the results of the current study may contribute to explaining

why speaker identification is lower in telephone speech. When looking at the segmental level, it has also been shown that certain acoustic-phonetic features may be associated with certain social factors. For /s/, some of its acoustic-phonetic features have been linked to social factors regarding gender and sexual orientation (Munson et al., 2006; Tracy et al., 2015). Although other acoustic-phonetic features (from vowels) also encode this type of information, it would be interesting to see whether the telephone effect on individual acoustic-phonetic features affects the perception of social information. For example, can listeners perceive a speaker's sexual orientation equally well from narrowband telephone signals as from broadband signals? Previous research has identified spectral skewness (L3) as an important feature in the perception of sexual orientation of male speakers (Munson et al., 2006). The current results show large effects of speech channel on both the acoustics and speaker classification of L3, which might mean that the perception of sexual orientation from /s/ is more difficult in telephone speech.

4.5 Conclusion

To conclude, for forensic speech science, it seems clear that the idiosyncratic information contained in telephone speech is severely compromised compared to studio-recorded speech. The current analysis on /s/ represents an extreme case of the telephone effect due to the high-frequency spectral characteristics of /s/; the telephone filter causes large changes in acoustic-phonetic measurement values and in speaker-classification accuracies. Despite large acoustic effects of speech channel, some measurements, in particular spectral tilt, showed relatively small effects of speech channel on speaker classification and can therefore still be useful for speaker discrimination in telephone speech. As for linguistic sampling context, although the landline telephone filter greatly affects the presence of expected linguistic effects on the acoustics, these are generally not, or only slightly, reflected in LDA speaker-classification accuracy.

CHAPTER 5

Effects of linguistic context on the LR strength-of-evidence

Abstract

Findings from previous work show that the linguistic environment that tokens are sampled from affect the acoustic realization and the within- and between-speaker variation of fricatives and nasal consonants. Specifically, more between-speaker variation and better speaker-classification accuracy using multinomial logistic regression were found for codas versus onsets and for tokens in highly coarticulated phonetic contexts versus in other contexts. The question remains whether these linguistic differences are relevant for forensic speaker comparisons. In

the current work, the effects of syllabic position on the strength of evidence from nasal /n/ and fricative /s/ were analyzed. Using a multivariate kernel density (MVKD) implementation of the Bayesian likelihood-ratio framework, results were in line with previous findings using other statistical methods. Namely, consonants in coda position perform slightly better at discriminating speakers than consonants in onset position. These results are discussed in terms of practicality in forensic speaker comparisons.

This chapter has been submitted.

5.1 Introduction

Reports on practices in forensic phonetic research show that auditory-acoustic analyses in forensic casework often make use of consonantal information (Gold & French, 2011, 2019). Although state-of-the-art methods in this field are evolving towards using automatic speaker recognition (ASR), this type of analysis is not always possible due to different legal contexts per country. For this reason, it is helpful to know what features from which segments are effective in auditory-acoustic analysis. Recent studies have shown that the same segment can carry different amounts of speaker-dependent information depending on the linguistic environment it was sampled from for both fricative and nasal consonants (Smorenburg & Heeren, 2020, 2021a). The current work aims to investigate the strength-of-evidence expressed by likelihood ratios (LRs) from Dutch nasal and fricative consonants, which have previously been shown to outperform other consonants in terms of their speaker discriminability (Amino & Arai, 2009; Kavanagh, 2012; Van den Heuvel, 1996). Syllabic position effects will be investigated, to see if linguistic contexts affect the strength-of-evidence from these consonants.

5.1.1 Articulation and acoustics of fricatives and nasals

In this work, we focus on Dutch fricative /s/ and Dutch nasal /n/. Firstly, because, amongst the consonantal sounds, nasals and fricatives are often shown to be the most speaker-specific, although there is some variation in the literature when it comes to the comparison between nasals and fricatives (Amino & Arai, 2009; Kavanagh, 2012; Van den Heuvel, 1996). Secondly, they are highly frequent speech sounds in Dutch (Luyckx et al., 2007) and therefore likely to be available in forensic case material in this language. Lastly, previous work (Smorenburg & Heeren, 2020; 2021a) has also shown that these segments retain useful speaker information in wiretapped recordings from landline telephones, despite

the compromised acoustics. For the fricatives specifically, alveolar /s/ was selected over other fricatives, even though its acoustics are compromised by the landline telephone filter. The main reason for this is that it outperformed dorsal fricative /x/ – the acoustics of which are not compromised by the landline filter – in an LDA speaker-classification test using spectral moments (cf. Smorenburg & Heeren, 2020). For the nasals, the selection of /n/ over the other two nasals in Dutch /m/ and /ŋ/ was two-fold; firstly, /n/ is more frequent than the other two segments (Luyckx et al., 2007). Secondly, previous work on Dutch showed /n/ to be more speaker-dependent than /m/ (Smorenburg & Heeren, 2021c; Van den Heuvel, 1996).

5.1.1.1 Fricatives

Articulatorily and acoustically, nasal and fricative consonants are very different. Fricatives are articulated by making a narrow constriction through which air is pressed with high velocity, resulting in aperiodic fricative noise. Looking at the acoustics, the resonance frequencies of fricatives are mainly dependent on the length of the anterior cavity, i.e., the space from the constriction to the lips. This is because, in voiceless fricatives, the noise source is not at the vocal cords but at the fricative constriction, which is then only filtered by the cavity anterior to that constriction before it passes the lips. Dorsal fricatives have larger anterior cavities and thus lower-frequency spectral energy and coronal fricatives have smaller anterior cavities and thus higher-frequency spectral energy. For example, Dutch alveolar /s/ has a spectral center of gravity of around 5.4 kHz (Ditewig et al., 2019), whereas Dutch velar/uvular /x/ has its spectral peak around 1.6 kHz (Van der Harst, Van de Velde & Schouten, 2007). Given that the spectral peaks for anterior fricatives such as /s/ are very high, their spectral peaks fall outside of the upper limit of narrowband (300 - 3,400 Hz) telephone filters (e.g., Smorenburg & Heeren, 2021b). Large effects of the narrowband filter would thus be expected for anterior fricatives but not for dorsal fricatives such as Dutch /x/. Any factors that significantly affect the length of the anterior cavity have a direct effect on fricative acoustics. Most obviously, speakers with larger vocal tracts will also have larger anterior cavities.

For example, male speakers have lower resonance frequencies for fricatives than female speakers (e.g., Jongman et al., 2000). The teeth have also been found to influence fricative acoustics; the teeth form an obstacle to the air that is pushed through the narrow constriction (i.e., the frication noise) and therefore the presence or absence of teeth (or dentures) can alter fricative spectra (Shadle, 1986).

Some fricatives have been associated with various social variables. Both Dutch /x/ and /s/ productions show regional variation in the Dutch language area. Fricative /x/ in particular is a very clear marker for region perceptually, with the 'soft' velar variant in Southern parts of the Dutch language area, and a 'harsher' uvular variant, which can sound very guttural due to the uvular trill, in the North and urban Randstad area (Van der Harst & Van de Velde, 2006). Fricative /s/ has been shown to be more retracted and [ʃ]-like in the Netherlands and more fronted and sharp-sounding in Flemish regions (Ditewig et al., 2019). For /s/, it has also been shown that social class and gender significantly affect /s/ productions, as working-class women were found to have /s/ acoustics similar to men (Stuart-Smith, 2007). Sexual orientation is also encoded in and perceived from the acoustics of /s/ (Munson et al., 2006; Tracy et al., 2015). For speakers of Dutch, /s/ acoustics have also been shown to contain information about ethnicity, with endogenous Dutch speakers producing more retracted /s/ articulations than Moroccan Dutch speakers (Ditewig et al., 2021). Fricative acoustics thus seem to convey social information about the speaker, which could contribute to the high between-speaker variation found in these sounds.

5.1.1.2 Nasals

Nasal consonants are articulated with a lowered velum, which opens the nasal cavity, allowing sound produced at the vocal cords to resonate there (Stevens, 2000, pp. 187-194 and 487-513). The vocal tract in nasal consonants runs from the glottis to the nostrils, with the oral cavity as a side branch that is closed at the mouth (for /m/), at the alveolar constriction (for /n/), or at the velar constriction (for /ŋ/). The resonance frequencies in nasals, i.e., the nasal formants, are associated with the larynx and the nasal cavity and are more or less a direct reflection of a

speaker's anatomy (ref). In most models for nasal consonants (cf. Stevens, 2000; Johnson, 2003; Fant, 1970), the oral cavity is modelled to produce antiresonances because it is a closed off side branch of the main vocal tract. These antiresonances, or antiformants, dampen sound at specific frequencies, which can shift or attenuate the nasal formants. The location of antiformants is dependent on the size of the oral cavity and thus varies by place of articulation. Additionally, the coupling of the nasal cavity with all its crevices adds surface area to the vocal tract, which further dampens the sound, i.e., lowers the amplitude and resonance frequencies, in nasals (Stevens, 2000, pp. 187-194 and 487-513). The low amplitude of nasals means that they are relatively weak sounds acoustically, which is especially noticeable in low quality recordings.

However, nasals are often reported to be robust to many contextual influences and therefore show relatively little within-speaker variation, which makes them relatively speaker specific (Rose, 2002). Nasal consonants are also affected by the telephone filter; their most prominent spectral characteristic, the first nasal formant, can be as low as 250 Hz (N1 for /m/: Fant, 1970), which is below the lower boundary of some narrowband telephone filters. In sum, nasal consonant acoustics better reflect information about a speaker's unique anatomy and physiology than oral consonants, resulting in relatively low within-speaker and high between-speaker variation. Articulatory-acoustic differences between nasal consonants cross-linguistically have not received a lot of attention (although see Tabain et al., 2016 on three Australian languages). Besides a study showing only minor differences between bilabial /m/ in Dutch versus English – with a slightly higher (31 Hz) second nasal formant in English than in Dutch (De Boer & Heeren, 2021) – not much is known about how Dutch nasals differ from nasals in other languages articulatorily and acoustically.

5.1.2 Linguistic context effects

It is well-known that there is variation in consonantal realizations due to linguistic variables such as prosodic structure and phonetic context. These effects might be relevant when selecting tokens to analyze in forensic speaker comparisons. In this section, prosodic effects on fricatives and nasals are described, both in terms of the linguistic effects on their acoustic realizations and their idiosyncratic information.

5.1.2.1 Prosodic effects

Prosodic structure can affect a segment's acoustics, which mainly seems to be related to the articulatory effort being higher in some linguistic positions relative to others. Some positions in speech are more constrained and are therefore articulated with more effort and precision. One clear example of this is syllabic position; compared to codas, onsets play a larger role in lexical perception (e.g., Gow et al., 1996; Marslen-Wilson & Zwitserlood, 1989) and are therefore articulated more clearly than codas, which are generally reduced in amplitude and duration, are more centralized in place of articulation and have lower signal-to-noise (SNR) ratios (Ohala & Kawasaki, 1984). Perhaps more generally, there seem to be boundary effects of prosodic constituents such as syllables, prosodic words, and intonational phrases (e.g., Cho & McQueen, 2005; Fougeron, 2001; Fougeron & Keating, 1998). For example, vowels in prosodically strong locations such as vowels with a nuclear pitch accent or vowels in initial versus final position within the prosodic constituent undergo less coarticulatory influence by neighboring segments (Cho & McQueen, 2005).

Prosodic structure and speech effort and precision have been linked to the amount of within- and between-speaker variation. The effects of articulatory effort generally go in two directions. On the one hand, parts of speech that are articulated with more effort and precision can be expected to have lower within-speaker variation (and lower between-speaker variation) because speakers make more effort to produce speech close to the model which conveys their desired linguistic

effects. For example, in perceptual speaker identification, listeners showed better accuracy for syllables containing onsets than syllables not containing onsets (Amino et al., 2007). On the other hand, parts of speech that are articulated with less effort and precision can be expected to have higher between-speaker variation (and within speaker variation). From the phonetic and phonological literature, it has often been mentioned that segment classification systems (such as automatic speech recognition systems) perform better on onset tokens than on coda tokens due to more speaker variation in coda position. For example, measures of spectral change between the nasal murmur and the following vowel show a clearer difference for place of articulation (here between alveolar /n/ and bilabial /m/) in onset than in coda position (Seitz et al.,1990). For formant and intensity contours of syllables, it was found that more between-speaker variation is present in the second half of syllables, i.e., the mouth closing gesture towards the coda of the syllable (cf. He & Dellwo, 2017; He et al., 2019). The authors hypothesized that less articulatorily constrained positions in speech, such as codas but more generally the second half of syllables, have more between-speaker variation, which could result in them being more speaker-specific.

Some studies have looked at effects of prosodic structure on speaker classifications and forensic strength-of-evidence. For example, McDougall (2004) has looked at effects of lexical stress and Heeren (2020) at effects of word class. The former found that nuclear-stressed vowels outperformed non-nuclear unstressed vowels in speaker-discrimination tests, which can be attributed to the increased speech effort, precision, and length in stressed positions (cf. McDougall, 2004). Regarding word class, function and content words have different acoustic realizations. For example, lexical frequency was found to have a shortening effect on the duration of content but not function words, with function words being shorter than content words in general (Bell et al., 2009). Dutch vowels from function words are not only shorter but also more centralized compared to vowels from content words (Van Bergem, 1993, pp. 38-39). This is likely related to the different phonological status of content versus function words, with the former always containing a strong syllable that can receive lexical stress and pitch accents and the latter only doing so in special circumstances such as when spoken in

isolation (cf. Selkirk, 1996). Heeren (2020a) found slightly better speaker-classification for content over function words using multinomial logistic regression, but similar performance using likelihood-ratio (LR) strength-of-evidence.

5.1.2.2 Phonetic context and coarticulation

For some speech sounds, coarticulation can provide idiosyncratic information (Nolan, 1983, Chapter 3). Fricative acoustics are highly dependent on contextual labialization. When fricatives are preceded or followed by rounded vowels or labial consonants, the lip-rounding movement can extend into the fricative, which lengthens the anterior cavity and lowers the resonance frequency (e.g., Koenig et al., 2013; Munson, 2004; Shadle & Scully, 1995). There seems to be between-speaker variation in the timing and degree of this coarticulatory lip-rounding, because /x/ and /s/ productions in labial contexts were found to contain more between-speaker variation than in other phonetic contexts (Smorenburg & Heeren, 2020).

Nasals are generally thought be rather unaffected by linguistic contexts due to the higher involvement of the nasal cavity instead of the oral cavity. However, models for nasal acoustics do indicate that the oral cavity has some effect on the nasal spectra through the nasal antiformants which are produced there. In production, it has indeed been shown that phonetic context affects nasal acoustics (Kurowski & Blumstein, 1987; Smorenburg & Heeren, 2021a; Tabain et al., 2016). In fact, it has been shown that the coarticulation between a nasal and the following vowel provides speaker-specific information (Smorenburg & Heeren, 2021a; Su et al., 1974). The claim that nasals have low within-speaker variation and high between-speaker variation due to the involvement of the rigid nasal cavity thus seems to lack some nuance.

5.1.3 Research questions

This work investigates whether selecting tokens from specific linguistic environments (which benefits the homogeneity of a set of segment realizations) can improve forensic speaker comparisons. For both fricative and nasal consonants, it has been shown that linguistic factors can affect the acoustics and speaker information available in those sounds. Specifically, tokens that occur in relatively less articulatorily constrained positions, such as codas compared to onsets and tokens in phonetic contexts that are highly coarticulated phonetic compared to other phonetic contexts, generally seem to contain more between-speaker variation and perform better in speaker classifications using multinomial logistic regression (Smorenburg & Heeren, 2020; 2021a). Given that tokens in these different linguistic environments have different acoustic realizations, it might therefore be preferable to select tokens from specific contexts to maximize the speaker discriminability and to have a set of homogenous tokens. However, being selective about the linguistic environment of tokens could result in insufficient datasets regarding the number of tokens, which can be problematic in often already short and/or low-quality forensic case material. In this work, we investigate the effect of syllabic position on the strength of evidence from two frequently-occurring Dutch consonants that have previously been shown to be relatively speaker-specific, namely fricative /s/ and nasal /n/.

5.2 Method

5.2.1 Materials

The main data analyzed in this work comes from the Spoken Dutch Corpus (Oostdijk, 2000). Specifically, component ‘c’ of the corpus, where speakers have spontaneous telephone conversations with other speakers that are previously known to them. This corpus was chosen because of the informal speaking style and because the wiretapped landline telephone recordings (300 - 3,400 Hz bandwidth) resemble

speech found in forensic case work. Speakers were wiretapped from their own home environments in the year 2002 using a digital switchboard, assumedly using their personal telephones, which means that recording conditions (ambient noise and telephone model) were not identical across speakers. Fricative /s/ and nasal /n/ tokens from 62 male adult speakers were segmented and analyzed. Each speaker had one to four 10-minute conversations available (M = 1.8, SD = 1.1). For speakers who had more than one conversation available, it is not clear to what degree these recordings were non-contemporaneous because only the recording year is available in the meta data. From the content of the conversations, the author thinks it likely recordings were made (successively) on the same day for any given speaker. Given that the sub-setting of data according to syllabic position would sometimes result in insufficient sets of tokens, all available data per speaker was used and treated as contemporaneous.

5.2.2 Segmentation

The orthographic transcriptions that are available for both corpora were used to produce automatic segmentations using Praat's forced-alignment (Boersma & Weenink, 2020). Because of the spontaneous nature of the conversations, these segmentations were often inaccurate. Therefore, the automatic segmentations were used to query tokens in the signal, which were manually estimated and corrected if necessary. Tokens were estimated using several exclusion criteria; they were excluded when there was overlapping speech between interlocuters, when there was laughter, when there were accent or person imitations, or when the token was not auditorily identifiable as the target token by the first author, who is a native speaker of Dutch.

Each token was then labelled on syllabic position and phonetic context. Syllabic position was defined lexically. Although syllabic position is sometimes defined phonetically – i.e., excluding ambisyllabic codas, which are codas followed by vowels – this resulted in low token numbers (N < 10) per condition per speaker for many speakers in this corpus. Wanting to use the same set of speakers across segments and

syllabic position, the lexical definition of syllabic position yielded sufficient tokens ($N > 10$) per condition per speaker to have a set of 59 speakers. Using the phonetic definition yielded a set of only 36 speakers with at least 10 tokens per syllabic position for both segments. Only speakers with at least 10 tokens per factor level across factors were included in the analysis. The resulting token numbers per segment and syllabic position are presented in Table 5.1.

As can be seen in Table 5.1, tokens are not equally numerous across syllabic positions; fewer tokens were available in coda than in onset position. For some speakers, fewer than 16 tokens were available per segment and syllabic position. Given that at least one 10-minute telephone recording was available for each speaker (note that these were conversations and that some speakers spoke less than others, instead listening to the interlocuter) and that not even 16 tokens were available across syllabic positions, it seems clear that selecting tokens from specific linguistic environments is challenging.

Table 5.1: *Token numbers per segment and sampling context*

Segment	Speakers		All	Onset	Coda
/s/	59	*N*	3,485	2,223	1,228
		M (*SD*)	58 (24)	38 (16)	21 (10)
		Range	26-150	15-85	10-66
		Speakers with N < 16		1	6
/n/	59	*N*	3,761	2,988	1,473
		M (*SD*)	63 (32)	50 (21)	25 (10)
		Range	20-137	14-116	10-75
		Speakers with N < 16		1	17

5.2.3 Acoustic analysis

For both fricatives and nasals, traditional acoustic-phonetic features from the literature that are easy to measure and interpret were selected to be estimated as speaker predictors. For fricatives, spectral moments are often used to describe the overall shape of fricative spectra, particularly sibilant fricatives (e.g., Forrest et al., 1988; Jongman et al., 2000; Shadle & Mair, 1996). More generally, these four dimensions can be used to describe Gaussian-like distributions. Importantly, spectral moments are not associated with specific events in the spectrum and can therefore be measured even in compromised signals. For Dutch in particular, fricative /s/ is clearly identifiable both auditorily and visually in the spectrum due to its lower spectral characteristics than in other languages such as English (cf. Smorenburg & Heeren, 2020). Spectral moments are sometimes also used to describe nasal consonants (e.g., Tabain et al., 2016), however, nasals have a formant structure, which makes the spectral moments less precise compared to nasal formants and bandwidths for nasal consonants (cf. Smorenburg & Heeren, 2021c).

For fricative /s/, the four spectral moments and duration were measured. The first spectral moment (M1) is the spectral centre of gravity and, in Praat (Boersma & Weenink, 2020), is computed as the mean frequency of the spectrum in Hz. The second moment (M2) is the spectral standard deviation and is computed as the dispersion of energy, i.e., variance, around M1 in Hz. Skewness (L3), the third spectral moment, is a coefficient that indicates how much the spectrum below the spectral mean differs from the shape of the spectrum above the spectral mean, i.e., whether the spectral shape leans to the left (lower frequencies) or right (higher frequencies). The kurtosis (L4), or fourth spectral moment, is a coefficient that indicates how much the shape of the spectrum differs from a Gaussian shape, i.e., how peaked the distribution is. The spectral moments were measured over the middle 50%1 of each fricative consonant over a 500 - 3400 Hz measurement range. Frequencies below 500 Hz were excluded to decrease effects of ambient noise and intruding voicing into the fricative.

For nasal /n/, the second (N2) and third nasal formants (N3) along with their bandwidths (BW2, BW3) were measured. The first nasal formant (N1), although it is the strongest component of the nasal spectrum, falls below or very close to the 300-Hz cut off of the narrowband telephone filter (also see Tabain et al., 2016) and could therefore not be measured reliably. Formants and their bandwidths were measured over the middle 50%[8] of each nasal consonant over the 800 - 3,400 Hz band using the Burg method, querying three formants in that range.

With regards to dynamic measurements across the consonant, a previous analysis showed that dynamic M1, N2 and N3 measurements did not contain much discriminatory power for Dutch /s, x, n, m/ (Smorenburg & Heeren, 2021c), so these were not considered in the current work.

5.2.4 Statistical analysis

The statistical analysis consisted of likelihood-ratio (LR) testing to obtain the strength-of-evidence for different linguistic contexts, specifically onsets versus codas. Speaker discriminability was tested with likelihood ratios (LRs). LRs reflect the ratio of the probability of the evidence under the hypothesis that two speech samples come from the same speaker (SS) to the probability of the evidence under the hypothesis that two speech samples come from different speakers (DS). The leave-one-out implementation with calibration (Morrison, 2007) based on the multivariate kernel density (MVKD) algorithm proposed by Aitken and

[8] Both fricative and nasal consonants show effects of phonetic context in acoustic measurements (spectral moments and nasal formants), even when measured at the middle 50% of these segments (Smorenburg & Heeren, 2020; 2021a). Nasal /n/ showed larger effects of phonetic context (coded as back versus non-back articulations) in coda position than in onset position (Smorenburg & Heeren, 2021a). Fricatives acoustics show effects of labialization of the context, but these did not show up in Dutch /s/ from landline recordings, assumedly due to the narrowband filter (Smorenburg & Heeren, 2020).

Lucy (2004) was used in software programme Octave (Eaton et al., 2019). In this implementation, within-speaker variation is modelled as a normal distribution and between-speaker variation is modelled with a multivariate kernel density.

For each LR system, same-speaker and different-speaker LRs were first computed in a development phase. Since not all speakers had multiple recordings, the tokens per speaker were divided in half to generate SS comparisons. This resulted in 59 same-speaker and 1711 different-speaker comparisons and accompanying LR scores. For the same-speaker comparisons, the leave-one-out MVKD implementation loops through all speakers, using the remaining 58 speakers as background data (Morrison, 2007). For the different-speaker comparisons, it loops through speaker pairs, using the remaining 57 speakers as background data. In a subsequent round of calibration, the LR scores from the previous step were used to obtain calibration parameters (shift, slope) to generate calibrated 59 same-speaker and 1711 different-speaker calibrated LLRs (log base = 10). System performance was then assessed through same-speaker and different-speaker LLRs, the equal error rate (EER) and the log-likelihood-ratio costs (C_{llr}: Brümmer & Du Preez, 2006), as well as the minimum log-likelihood-ratio costs ($C_{\text{llr}}^{\text{min}}$). For the LLR, a value of 1 means that the evidence is 10 times more likely under the same-speaker hypothesis and a value of –1 means that the evidence is 10 times more likely under the different-speaker hypothesis. The EER metric is based on the percentages of the system's false misses (i.e., same-speaker as different-speaker) and false hits (i.e., different-speaker as same-speaker). The C_{llr} also expresses false LR misses and hits, but as a gradient, therefore taking into account the magnitude of errors. The $C_{\text{llr}}^{\text{min}}$ shows the system's discrimination potential when optimally calibrated. Subtracting the $C_{\text{llr}}^{\text{min}}$ from the C_{llr} thus gives the calibration loss ($C_{\text{llr}}^{\text{cal}}$). For all three performance measures, closer to 0 is better. Median LLRs and performance measures were obtained using R package 'sretools' (Van Leeuwen, 2011).

The LR systems built for nasal /n/ contained duration and the second and third nasal formants and bandwidths (N2, BW2, N3 and BW3) as predictors and the systems built for fricative /s/ contained duration and the four spectral moments (M1, M2, L3, L4) as predictors. Correlations

between predictors within a single system were all weak to medium ($r <$ .60). For both the nasal and fricative segment, the first system was built using all available tokens for that segment. Then, systems were built using either onset or coda data. Because the available numbers of tokens differ across speaker and syllabic position, up to 16 tokens were randomly sampled per speaker per syllabic position (in some cases, some speakers had fewer than 16 but at least 10 tokens available per syllabic position). The first system was then also run again using ≤16 tokens per speaker, to make for a fair comparison.

5.3 Results

As can be seen in Table 5.2 and Figure 5.1, the nasal consonants /n/ and fricative /s/ perform rather similarly when all available tokens per speaker are used.

Table 5.2: *Same-speaker (SS) and different-speaker (DS) LLRs, C_{llr}, C_{llr}^{min}, and EER per segment and syllabic position.*

		SS LLR	DS LLR	C_{llr}	C_{llr}^{min}	EER
/n/	All tokens	1.79	–2.39	0.55	0.48	16.74
	N ≤ 16	1.23	–1.64	0.62	0.55	18.30
	Onset N ≤ 16	1.26	–1.49	0.64	0.59	20.58
	Coda N ≤ 16	1.70	–2.86	0.50	0.45	13.89
/s/	All tokens	1.46	–2.60	0.59	0.46	14.16
	N ≤ 16	1.03	–1.25	0.66	0.60	21.58
	Onset N ≤ 16	0.80	–0.56	0.81	0.64	22.48
	Coda N ≤ 16	1.20	–1.55	0.64	0.59	20.02

In line with expectations from reported low within-speaker variation for nasals, the nasal /n/ shows slightly better same-speaker comparisons (as shown by the higher same-speaker LLRs for /n/ than for /s/ in Figure 5.1). The fricative /s/, on the other hand, shows slightly better different-speaker comparisons (as shown by the lower different-speaker LLRs for /s/ than for /n/ in Figure 5.1). This is also in line with expectations given the reported high between-speaker variation for fricatives (e.g., Smorenburg & Heeren, 2020). When only up to 16 tokens per speaker are considered, which were randomly sampled across syllabic positions, i.e., from the full set of available tokens with no consideration to linguistic context, performance decreases significantly. This suggests that 16 tokens per speaker (with some speakers having fewer tokens, see Table 5.1) did not provide a representative sample for these speakers.

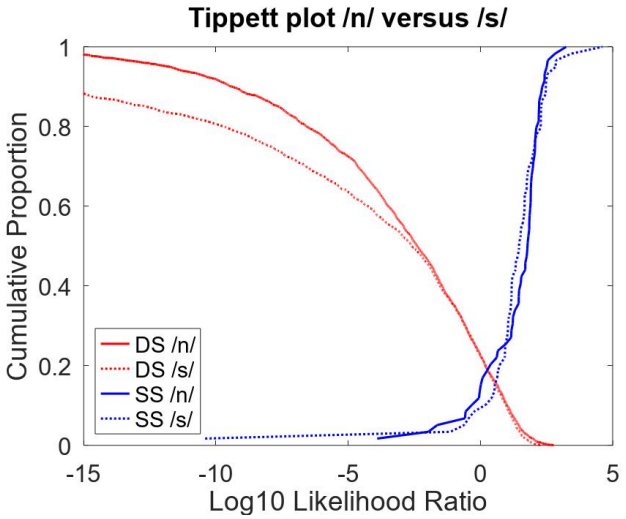

Figure 5.1: *Tippett plot for the LLRs generated using all available /n/ versus /s/ tokens per speaker.*

Regarding the linguistic effects, from figures 5.2 and 5.3 (as well as from the performance statistics in Table 5.2), it can be seen that the

strength of evidence for both /n/ and /s/ differ by syllabic position, which is in line with the multinomial regression analysis from previous work (Smorenburg & Heeren, 2020; 2021a). For onsets, there is no advantage in strength of evidence from creating a homogenous set of onsets compared to not taking syllabic position into account. The LLRs for codas (the dotted lines in figures 5.2 and 5.3) show better speaker discrimination as shown by the larger separation between different-speaker LLRs and same-speaker LLRs. Particularly for /n/, the coda position, even though the number or tokens are relatively low ($N \leq 16$), performs similarly to when all available tokens per speaker ($M = 63$) are used (see Table 5.2). Given that only segmenting and analyzing 16 tokens is less laborious than selecting many more tokens from all available contexts, the former might be preferable. One caveat being that there is enough speech available to find sufficient tokens that occur in coda position. For /s/, having more tokens results in better performance. These differences between segments are discussed further in the next section.

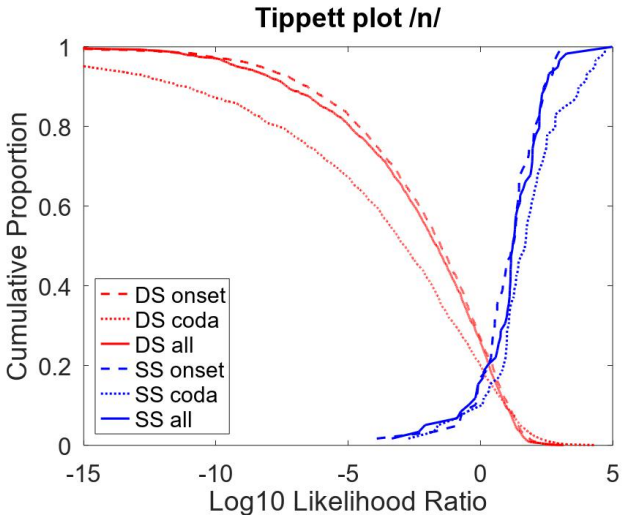

Figure 5.2: *Tippett plot for the LLRs generated for /n/ using tokens sampled across linguistic environments, from onsets, or from codas (sample size per speaker across all conditions $N \leq 16$).*

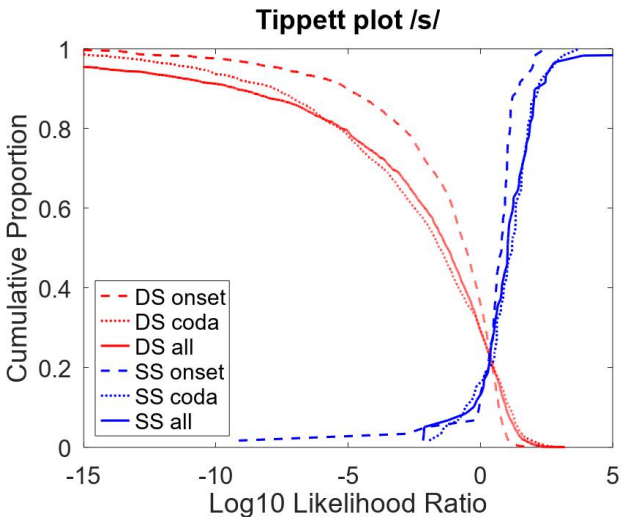

Figure 5.3: *Tippett plot for the LLRs generated for /n/ using tokens sampled across linguistic environments, from onsets, or from codas (sample size per speaker across all conditions N ≤ 16).*

5.4 Discussion and conclusion

Previous research has shown that linguistic factors can have large effects on a segment's acoustics. For nasals and fricatives, it has previously been shown that both nasal and fricative consonants show effects of syllabic position and phonetic context on the acoustics and speaker-dependent information (Smorenburg & Heeren, 2020, 2021a). Specifically, codas were reduced compared to onsets, and nasals and fricatives were highly coarticulated in back-articulated and labial contexts respectively. Given these acoustic differences, better speaker-discrimination might be achieved when segments are more homogenous within a speaker, which could be achieved by selecting tokens from a specific linguistic environment. Additionally, it is possible that some linguistic

environments contain more speaker information than others. Specifically, it has been suggested that less articulatory constrained positions in speech show more between-speaker variation (cf. He & Dellwo, 2017; He et al., 2019). Codas can be described as less articulatorily constrained as onsets, which is reflected in numerous observations of coda reduction. Previous work showed that codas and segments in highly coarticulated contexts had more between-speaker variation and performed better in speaker classifications with multinomial logistic regression (Smorenburg & Heeren, 2020; 2021a).

The current work shows that differences for syllabic position persist in likelihood ratio analysis, with greater strength of evidence for tokens in coda position compared to onset position. However, for /s/, despite the fact that selecting tokens from specific linguistic environments has a small effect on the strength of evidence, similar results were obtained when all available tokens across linguistic environments were used, even when the sample size was capped at 16 tokens per speaker. This means that, for /s/, selecting tokens in coda position specifically does not benefit the strength of evidence in speaker comparisons. For /n/, selecting tokens in coda position specifically resulted in similar performance compared to when all available tokens per speaker were used. This suggests that /n/ is more robust to sample size (at least compared to /s/). This might be explained by the low within-speaker variation in /n/, resulting in little difference in performance when the sample size per speaker is large ($M = 63$) or smaller ($N \leq 16$), because even a small sample per speaker seems to give a good estimation for the within- and between-speaker variation for /n/.

One major consideration is the availability of tokens per segment and syllabic position. Many decisions in this work, such as which segment to select and how to define syllabic position (lexically versus phonetically), were influenced by the number of available tokens per speaker. Compared to what is sometimes available in forensic casework, one to four 10-minute conversations per speaker seems like sufficient material, but even for highly-frequent consonants, the availability of tokens per condition was low for many speakers, particularly for segments in coda position. Not only do the segments studied here simply seem more frequent in onset position, due to coda reduction (and the

common coda /n/ deletion in weak syllables in Dutch: Silva et al., 2003), some segments in coda position could not segmented. Not unrelated, the landline telephone recordings used in this work have compromised acoustics due to the narrowband filter (300 – 3,400 Hz). Nasals have relatively low amplitudes, especially above 500 Hz, and can therefore be hard to measure in low-quality recordings such as the narrowband telephone speech used here. Measuring the first nasal formant, which can be as low as 250 Hz (Fant, 1970), is therefore highly unreliable. Measurements from fricative /s/ are highly affected because the spectral centre of gravity is generally higher than the 3,400 Hz limit of the narrowband telephone filter (Smorenburg & Heeren, 2021b). It is therefore possible that the comparison between /n/ and /s/ yields different results when looking at high-quality microphone recordings. Thus, selecting tokens from either onset or coda position does not seem feasible or particularly beneficial for forensic casework, as the numbers of tokens can be insufficient even in 10-minute conversations (partly due to reduction in coda position) and there is no strong advantage in terms of the strength of evidence.

This comparison between consonants in the current results is interesting in terms of the sources of within- and between-speaker variation for these segments. Given that various social variables have been shown to affect fricative acoustics (particularly /s/), it has to be assumed that the source of the between-speaker variation is perhaps not mainly the speaker's unique anatomy and physiology, but rather the speaker's expression of their social identity. Nasal consonants, on the other hand, are claimed to mostly reflect a speaker's unique anatomy and physiology due to the coupling of the relatively rigid nasal cavity which has different shapes and sizes between speakers. Because the oral cavity is less involved is nasal sounds (acting not as a main resonator but as a closed-off side branch which produces antiformants), the within-speaker variation is also relatively low. From a forensic perspective, the latter source of between-speaker variation is preferable because it is relatively unchangeable. Earlier work on the speaker-specificity of Dutch consonants from read nonsense words found that /n/ had higher speaker-specificity than /s/ (here defined as the ratio of between- to within-speaker variation in acoustic measurements: Van den Heuvel, 1996). This

is in line with current results using consonants from spontaneous telephone conversations when the numbers of tokens per speaker was capped at 16 tokens, but not necessarily when all available tokens per speaker were used, as /n/ and /s/ then perform similarly.

To conclude, likelihood ratio analysis showed results congruent with previous work using multinomial logistic regression analysis, namely that linguistic factors can have small effects on the speaker discrimination. However, these effects seem too small to benefit forensic speaker comparisons, especially in the light of the scarcity of material in case work. Rather, prioritizing the quantity of tokens seem to result in stronger strength of evidence.

CHAPTER 6

Summary and conclusions

6.1 Summary

This dissertation aimed to investigate how the speaker-specificity of consonants is dependent on linguistic factors, specifically segments' immediate phonetic context and syllabic position. Focus was placed on nasal and fricative consonants, which have previously been found to be relatively speaker-specific. In the following sections, the chapters reported above are briefly summarized, after which they are discussed in terms of the theoretical and practical implications. Lastly, some

suggestions for future work are made based on the findings and the limitations of the current work.

Chapter 2

In this chapter, two linguistic effects on the acoustics and speaker-specificity of Dutch fricatives were examined. Fricatives /s/ and /x/ were selected for their frequency of occurrence in Standard Dutch and, in the case of /s/, because previous research found this sound to be relatively speaker-specific (e.g., Kavanagh, 2012; Van den Heuvel, 1996). These fricatives were sampled from spontaneous telephone conversations in the Spoken Dutch Corpus (Oostdijk, 2000) and were investigated on their within- and between-speaker variation as a function of two linguistic factors: phonetic context and syllabic position. Significant effects of these factors were found on the acoustics, predominantly for /x/. For syllabic position, the acoustics showed coda reduction. For phonetic context, the acoustics showed effects of coarticulatory labialization, which is in line with previous literature showing that labialization lowers the spectral mean in fricative spectra (e.g., Bell-Berti & Harris, 1979; Koenig et al., 2013). Using multinomial logistic regression analysis in a following speaker-classification test, codas showed slightly better speaker-classification accuracy than onsets and fricatives with labial neighbors showed slightly better speaker-classification accuracy than fricatives in other phonetic contexts. This was attributed to between-speaker variation in the degree of reduction and coarticulation. It seems that speakers have individual ways in which codas are reduced and in which fricatives in labial contexts are coarticulated with regards to the specific timing and degree of articulatory gestures.

Acoustic effects were mostly observed for dorsal fricative /x/ and not for coronal /s/. Given the previous literature showing coarticulatory labialization for /s/ and the current findings for /x/, it was assumed that the lack of linguistic effects for /s/ were due to the narrowband telephone

filter of 300 – 3,400 Hz, which does not capture all the relevant acoustic information for /s/, while it does seem to do so for /x/.

The results in this chapter point to the need to consider linguistic factors when sampling segments in the forensic setting, as some specific linguistic contexts seem to yield more speaker information than others. However, the speaker-classification gain in these contexts were relatively small, possibly too small to need to be considered in forensic speaker comparisons (as was discussed in chapter 5).

Chapter 3

The line of research described in chapter 2 was extended to include two Dutch nasal consonants in chapter 3. The nasals /n/ and /m/ were sampled from the same spontaneous telephone conversations from the Spoken Dutch Corpus (Oostdijk, 2000) used in chapter 2. Again, the effects of syllabic position and phonetic context on the acoustics and within- and between-speaker variation were examined. Whereas fricatives are often found to be affected by contextual labialization, nasals can show effects of front-to-mid versus back-articulated context, with lower (second) nasal formant values when the nasal has a back-articulated neighbor. For phonetic context, a distinction was therefore made between back and non-back neighbors (opposed to the labial versus non-labial distinction for fricatives).

Results showed interactions between syllabic position and phonetic context in both the acoustics and speaker-classifications. For bilabial /m/, high degrees of place coarticulation mostly occur anticipatorily in onset position, while for alveolar /n/, there is mostly carry-over place coarticulation in coda position. Coarticulation thus seems to occur mostly within the syllable domain, but in opposite directions for the two nasal consonants. This could possibly be related to frequency of occurrence of these segments in onset versus coda position, as in these Dutch data /n/ was more frequent in coda position than /m/. The relative markedness of /m/ in coda position could thus have led to

resistance to coarticulation (see section 6.2.1. for more discussion on this topic).

Subsequent speaker classifications using multinomial logistic regression showed that /m/ onsets, which showed larger degrees of coarticulation, show better speaker-classification accuracy than /m/ codas. In line with the acoustics, for alveolar /n/ the pattern was the reverse; /n/ codas, which showed larger degrees of coarticulation, showed better speaker-classification accuracy than /n/ onsets. We concluded that highly coarticulated tokens contain more speaker information because of the between-speaker variation in the timing and degree of coarticulation.

Chapter 4

In chapter 4, a remaining question from chapter 2 was addressed. In chapter 2, it was assumed that the lack of acoustic effects of linguistic factors for /s/ was due to the narrowband telephone filter, which cuts off spectral energy for this fricative. This assumption was tested using an English speech corpus that includes wiretapped narrowband telephone conversations that were simultaneously recorded with a high-quality microphone placed in front of the speaker. Using an additional language would show whether previous results extend to another, albeit similar, language.

Results showed that English fricative /s/ showed the expected effects of coda reduction and coarticulatory labialization on the acoustics when measured in the high-quality microphone recording. Although the literature so far had mostly focused on anticipatory labialization, the degree of carry-over labialization was larger than anticipatory coarticulation. This finding is in line with the idea that patterns of English coarticulation are predominantly carry-over (Hoole et al., 1993). This contrasts with results on Dutch in chapter 2, which showed larger anticipatory labialization for Dutch /x/, indicating that Dutch and English might have different patterns for labialization. More importantly, results

showed that linguistic effects could not be observed in the acoustics of the narrowband telephone recording (300 – 3,400 Hz landline filter). Although some significant linguistic effects were found, they were not similar to the effects found in the studio recording in terms of magnitude and direction and no clear pattern could be discerned. This suggests that the telephone filter can have unpredictable effects on the acoustics. The speaker classifications showed some sampling effects in the broadband studio recordings, but not in the narrowband telephone recordings. This means that linguistic effects can potentially be relevant in broadband signals, but less so when dealing with narrowband signals, at least for segments with high-frequency spectral energy such as /s/.

Chapter 5

In chapter 5, some findings from previous chapters were tested in the Bayesian likelihood-ratio framework, to see whether sampling tokens from specific linguistic contexts affected the strength of evidence using likelihood ratios as it affected the speaker classifications using multinomial logistic regression in chapters 2 to 4. Given that these linguistic factors have been shown to affect the acoustics in chapters 2 to 4, sampling from specific contexts should result in more homogeneous sets of tokens. However, speech material can be scarce in forensic case work, meaning that sampling from specific linguistic contexts can lead to insufficient tokens per speaker. Results in this chapter showed that sampling from codas leads to stronger evidence than sampling from onsets for both /n/ and /s/. However, differences between speaker-classification accuracy across linguistic contexts were minor, and results also showed that prioritizing token numbers yielded the best speaker discrimination results. Given the minor differences across linguistic contexts and the often-scarce materials, it was therefore concluded that sampling from specific contexts in forensic contexts is not practical.

6.2 Conclusions

6.2.1 Theoretical implications

This section will discuss some of the theoretical implications with regards to the findings described in this dissertation.

6.2.1.1 Phonetic context effects

A large body of previous phonetic research has shown that phonetic context can affect the acoustics of speech segments. The current work, however, has not made a distinction between phonetic and phonological variation in speech sounds in its examination of phonetic context. Coarticulation refers to the acoustic and articulatory overlap between articulatory gestures in speech sounds in connected speech. In other words, there is coarticulation because the articulators have to move from an articulatory target for one sound to another articulatory target for another sound in quick succession, assimilating features to facilitate articulation. Coarticulation is thus a phonetic, gradient process. Assimilation, on the other hand, is often used to refer to a phonological[9] and categorical process in speech that does not stem solely from the physiological properties of the vocal tract, but from the acquired phonological rules in a certain language. These rules operate in specific phonological environments and result in allophones, i.e., different realizations of the same phoneme. Whereas coarticulation is obligatory (you cannot tell your articulators to time-jump into a new position, they have to travel there), assimilation is optional in the sense that it is language-specific[10]. For example, in the Received Pronunciation (RP)

---

[9] Note that some have argued that there is no clear distinction between phonetic and phonological variation and that gradient and categorical changes can overlap (see e.g., Scobbie, 2012).

[10] Although not further discussed here, phonological rules can furthermore be obligatory and therefore predictable or optional and free within languages. For example, in English, voiceless stops /p t k/ are always aspirated in the onset of stressed syllables [ph th kh] unless they follow an /s/ as in [spi:k]. Additionally, these sounds also show free variation, i.e., overlapping but not contrastive distribution, with their

accent of English, lateral consonant /l/ is produced as dark [ɬ] at the end of words or before consonants, but as clear [l] anywhere else (compare the clear [l] in *letter* to dark [ɬ] in *feel* or *milk*). In both the English and Welsh in southern Wales, on the other hand, clear [l] is found is all positions (Penhallurick, 2008). The former language variety thus has two allophones for /l/, whereas the latter does not have the dark [ɬ] allophone.

Although the current work has not made a distinction between phonetic and phonological aspects in the observed effects of phonetic context, based on the findings on coarticulatory labialization in Dutch (chapter 2) versus in English (chapter 4), some tentative conclusions can still be drawn. Namely, in both languages there is a phonetic aspect of coarticulatory labialization that seems unavoidable, resulting in at least some degree of coarticulatory labialization across syllabic positions (onset and coda position), directionality (anticipatory and carry-over), and languages (Dutch and English). However, clear differences were also observed. Specifically, coarticulatory labialization in English seemed to occur predominantly in a perseverative manner, i.e., effects of left context were larger than effects of right context. This provides some evidence for the hypothesis that English has predominantly perseverative, or carry-over, coarticulation (Hoole et al., 1993). In Dutch (chapter 2), the dorsal fricative /x/ showed somewhat larger anticipatory coarticulation. This might be indicative of other labial coarticulation patterns in English versus Dutch, with the former being more carry-over and the latter more anticipatory in nature. This difference is possibly due to different timing mechanisms in motor control planning between Dutch and English, specifically in the onset and/or length of the labial gestures. Hence, these seem to be language-specific, and thus acquired, patterns of labialization.

The results in chapters 2 and 3 show that previously observed effects of phonetic context are also observable in spontaneous speech, which makes them more robust. However, more research is still needed to describe the differences in phonetically- and phonologically-restrained variation across languages. For example, previous research on

---

unreleased variants [p¬ t¬ k¬] in word-final position such as in [stɒp¬] (e.g., Rowe & Levine, 2018, pp. 68-69).

coarticulation between vowels and nasal consonants /n/ and /m/ found more coarticulation for /m/, presumably because /m/ has no particular articulatory tongue target, whereas for /n/ the tongue target is alveolar and therefore more resistant to anticipatory coarticulation (cf. Su et al., 1974). This is in line with what was found for nasal consonants in onset position in chapter 3, but not for nasal consonants in coda position, where /n/ showed higher degrees of coarticulation than /m/. This might be specific to Dutch, where word-final /n/ is highly frequent and often elided (Silva et al., 2003), and word-final /m/ is more marked due to its low frequency of occurrence. Low frequency of occurrence could result in more resistance to coarticulation. For example, it has been shown that, in English, high frequency words show higher degrees of coarticulation, whereas lower frequency words show more resistance to coarticulation (e.g., Yun, 2006). Similar findings exist for syllables, where it has been suggested that highly frequent syllables are stored in a mental syllabary that includes articulatory routines (cf. Cholin et al., 2006; Levelt & Wheeldon, 1994). Experimental work indeed shows that there are syllable-frequency effects on the degree of coarticulation, with larger gestural overlap in highly frequent syllables (e.g., Herrmann, Whiteside & Cunningham, 2008). However, this explanation does not extend to onset position, where there is no such clear difference in frequency of occurrence between /n/ and /m/, but where the bilabial nasal showed higher degrees of coarticulation than the alveolar nasal.

In read speech, the articulation of word-final /n/ in Dutch seems to be affected both by social variables such as region and the interaction between sex and age, as well as by linguistic variables such as the word type (e.g. mono- versus polymorphemic) and the following phonetic context (vowel, consonant, pause, schwa or clitic: Van de Velde & Van Hout, 2001). Although the social variables were mostly excluded in this speaker set, i.e. they were all males aged 18-50 who spoke Standard Dutch, our factors did not distinguish between these specific phonetic contexts. Rather, pauses and non-back vowels and consonants were grouped together and back vowels and consonants were grouped together. In future work, the reduction of /n/ in the spontaneous Dutch data worked with here could be re-evaluated using the contexts described in Van de Velde & Van Hout (2001). Given the acoustic nature of the present work,

/n/ could only be measured when not deleted (or reduced to an extend that segmentation was no longer possible) and given the interest in added speaker information from coarticulation specifically, the current work chose to focus on non-back versus back-articulated phonetic context for nasals.

Interactions between phonetic context and syllabic position effects in the current results showed that effects of phonetic contexts were larger within the syllable domain than across a syllable boundary. Namely, for the nasal consonants in chapter 3, /m/ showed larger effects of following context in onsets and /n/ showed larger effects of preceding context in codas. Similar syllable-boundary effects on labial coarticulation were found for fricative consonants from the same telephone dialogues in chapter 2, where these syllable boundaries additionally coincided with word boundaries in all cases. This seems to indicate that there is more resistance to coarticulation across syllable boundaries, although other studies indicate that the effect of prosodic boundaries on coarticulation is generally small or absent (e.g., Cho & McQueen, 2005; Hardcastle, 1985).

6.2.1.2 Sources of speaker information

In this dissertation, both fricative and nasal consonants were examined on their speaker information. Previous phonetic theory and observations have indicated that fricatives and nasals seem to contain qualitatively different types of speaker information. The results in this dissertation corroborate this.

Fricative acoustics are partly dependent on the size of the vocal tract, resulting in lower spectral averages in males than in females (for /s/: Jongman et al., 2000). Additionally, fricatives seem to convey social information about the speaker such as social class (Stuart-Smith, 2007), sexual orientation (Munson et al., 2006), ethnicity (Ditewig et al., 2021), and region (see Ditewig et al., 2019 for /s/ and Van der Harst & Van de Velde, 2006 for /x/). In chapters 2 and 3, a set of adult male speakers of Standard Dutch was selected from the Spoken Dutch Corpus (Oostdijk, 2000). Although this makes for a relatively homogeneous group of speakers, differences between social factors, ethnicity, and region will

still exist in this group to a certain extent. In other words, the observed speaker variation may partly be due to social differences between speakers, which is group behavior rather than speaker-specific behavior. As a consequence, although /s/ was quite successful in distinguishing speakers in this group of adult male speakers of Standard Dutch, it is possible that /s/ is less speaker-specific in even more homogenous groups of speakers, who have been matched on several social variables.

Nasal consonants, on the other hand, seem to be a better reflection of a speaker's vocal tract, with less influence from (socio)linguistic factors. In other words, these sounds are more dependent on the metaphorical hardware (i.e., the vocal tract) and less on the software (acquired language behavior). This is thought to be the case because of the involvement of the nasal cavity, which is relatively rigid and therefore relatively invariable, but have highly different shapes and sizes between speakers (cf. Rose, 2002). In chapter 5, results showed that /n/ was more robust to smaller sample sizes than /s/, presumably due to the low within-speaker variation in /n/ compared to /s/. At the same time, in chapter 2 it was shown that nasals display more variation than is generally assumed, in this case from coarticulation with the phonetic context. Although nasal acoustics are strongly affected by the coupling of the nasal cavity, the oral cavity still serves as a side chamber to the vocal tract that, in nasal consonants, runs from the glottis to the nostrils. That is how place of articulations are distinguished in nasal consonants; by variations in tongue position in the oral cavity which acts as a side chamber and produces anti-formants at different frequencies.

The type of speaker variation found in nasals, which is predominantly associated with the shape and size of the vocal tract, might be more stable across populations that differ in their level of homogeneity and might therefore be preferable in a forensic context. However, one disadvantage of nasal consonants is their relative acoustic weakness. Due to the involvement of the nasal cavity, which adds a lot of surface to the vocal tract, nasal sounds are more dampened and lower in frequency than oral sounds (Stevens, 2000). On top of that, nasal consonants, like vowels, have complex formant structures. This makes them more difficult to measure, particularly when using semi-automatic measuring methods and especially in lower-quality recordings such as wiretapped telephone

conversations. Fricatives, on the other hand, contain high-velocity airflow resulting from the narrow fricative constriction (Stevens, 2000). They are therefore relatively easily identifiable in spectrograms, even in lower-quality recordings. They also have the advantage that they can be adequately captured by relatively simple measurements, namely spectral moments, which are often used to represent the general spectral shape in fricative sounds (cf. Koenig et al., 2013) and are also easy to explain (opposed to more highly-dimensional acoustic features such as MFCCs). When comparing the strength of evidence from nasal consonant /n/ to fricative consonant /s/, both perform very similarly when all available tokens per speaker were included, but /n/ seems to be more suitable when fewer tokens are available because it is slightly more robust to sample size per speaker, which seems due to the lower within-speaker variation for nasals compared to fricatives.

With regards to the type of acoustic features, it seems that spectral measurements contain more speaker information than temporal and amplitudinal measures. This is probably related to the fact that these measures reflect the size and shape of the relevant resonance chambers of the vocal tract, which are dependent on not only acquired speech behavior, but also on a speaker's hardware, i.e., the vocal tract. This is not the case for temporal and amplitudinal measures (or at least to a lesser extent, e.g., see some recent discussion on the stability and variation in patterns of respiration: Fuchs, 2022). Dynamic spectral measurements did not contain a lot of speaker information either, which was surprising given the general findings in this dissertation that some contexts contain more speaker information that seemed to be due to idiosyncrasies in (co)articulation. Possibly, the consonants under study here are too short and the contexts too variable to get much useful information from dynamic measurements from consonant onset to offset (cf. Heeren, 2020b on the lack of information in dynamic measurements for vowels in spontaneous speech). The observations on the relative contributions of acoustic-phonetic features to the speaker classification tests were consistent across the two different fricatives that were investigated in chapter 2 and extended to nasal consonants in chapter 3.

6.2.1.3 Distribution of speaker information

In the introduction of this dissertation, two competing hypotheses were put forward with regards to the dependency of a sound's speaker information on its linguistic environment. One predicted that speech sounds in articulatorily strong positions and contexts would show less within-speaker variation and therefore be speaker-specific. This hypothesis was mostly based on work on speaker information in stressed versus unstressed vowels (McDougall, 2004) and speaker information in content versus function words (Heeren, 2020a). The second hypothesis predicted that speech sounds in articulatorily free positions (with less linguistic constraint) would show more between-speaker variation and therefore be more speaker-specific. This hypothesis was based on findings on there being more between-speaker variation in the second half of syllables – i.e., the mouth closing gesture towards the coda – in both formant and intensity contours (He & Dellwo, 2017; He, Zhang, & Dellwo, 2019). Relatedly, consonants that are in highly coarticulated environments were expected to contain additional articulatory speaker information (cf. Nolan, 1983, Ch. 3).

In the current dissertation, it was shown that there is a tendency for speech segments in contexts or positions that are less articulatorily constrained to display relatively more between-speaker variation than within-speaker variation. Generally, this concerns codas (compared to onsets) and tokens in highly coarticulated phonetic contexts such as fricatives in labialized contexts (compared to other phonetic contexts). However, from the findings in chapter 3 it can be concluded that the hypothesis that codas are less articulatorily constrained than onsets and therefore have more between-speaker variation required some nuance. Namely, the specific linguistic environments that are more speaker-specific are not entirely consistent across speech segments and languages. Regarding segments, variation was observed even within sound classes. Specifically, whereas Dutch alveolar /n/ conformed to the previously observed pattern of more speaker-specific codas than onsets, Dutch bilabial /m/ did not show this pattern in the conversational telephone data from the Spoken Dutch Corpus. For /m/, onsets were more coarticulated than codas and – presumably as a result – also contained more speaker-dependent information. Regarding cross-linguistic

variation, the findings in chapter 4 implied that Dutch and English have different patterns of labial coarticulation, with English being more regressive in nature than Dutch. The earlier hypothesis that the second half of syllables display more speaker-variation may thus be too general. Rather, the current findings should be regarded as specific to the articulatory-acoustic dependencies that exist in Dutch fricatives and nasal consonants (chapters 2 and 3) and English fricative /s/ (in chapter 4).

In other words, findings in this dissertation do not seem to be directly generalizable to other languages because which parts of the speech signal are more reduced and coarticulated is language-specific. For example, in languages like French, "labial constriction is much more crucial for vocalic rounding contrast than in English" (Noiray et al., 2010, p. 166). In a previous articulatory study, differences were found between the rounding mechanisms in young speakers of Canadian French and American English when modelling the anticipatory labial motor control for rounded vowel /u/ on preceding sounds. Noiray et al. (2010) "found very regular anticipatory behaviors for six of the seven French children tested" (p. 166), which the authors thought was related to the relative importance of labial constriction in French compared to English. Interestingly, although there were differences between the languages, it was also reported that all speakers showed idiosyncrasies in rounding gestures (here defined as labial protrusion and constriction). Anticipatory motor control provides the glue, or overlap, by which sequential speech sounds and syllables are held together. At its core, this is a motor control issue that seems to be language-dependent to some degree (e.g., Noiray et al., 2010).

Within languages, motor control also shows variation dependent on prosodic boundaries. For example, at the phrase level, articulatory gestures slow when a phrase boundary is approached and speed up again after the phrase boundary has passed (Byrd & Saltzman, 2003). In this dissertation, the examination of prosodic structure was mostly restricted to syllabic structure, focusing on coda reduction (Ohala & Kawasaki, 1984). In the introduction of chapter 2, the seeming cross-linguistic variation in coda reduction for /s/ as found in previous research is described: In English, coda /s/ displayed lower intensity (Solé, 2003) and

duration (Redford & Diehl, 1999), but in German, no reduction on either spectral or temporal measures was reported (Cunha & Reubold, 2015), although in both languages, codas did show more variability and were generally less identifiable. This latter observation was also found for both Dutch (chapters 2 and 3) and English (chapter 4) in the current dissertation. Codas generally seem to place less constraint on motor control and articulatory targets than onsets, although, again note that the bilabial nasal seems to be a clear exception to this pattern in the current data.

Regarding the amount of speaker information found in different linguistic environments, it is tentatively concluded that those parts of speech that are less linguistically constrained and therefore have more articulatory freedom contain relatively more between-speaker variation than within-speaker variation. For the consonantal segments examined in this dissertation, the coda would be such a position (except in the case of /m/). Segments in contexts that show high degrees of coarticulation with neighboring segments also seem to contain additional speaker information. These findings are in line with the idea that there are speaker-specificities in reduction and coarticulatory gestures (cf. Nolan, 1983, Ch. 3) and that speech segments in contexts with more reduction and coarticulation can therefore be (slightly) more speaker-specific.

6.2.2 Practical implications

For forensic speaker comparisons, the findings in this dissertation may perhaps lay some concerns to rest. Although significant effects of linguistic context were found on the acoustic realizations, the magnitude of these effects on speaker discrimination using multinomial logistic regression, linear discriminant analysis, and likelihood ratio analysis were relatively small. In some cases, it might be beneficial to sample tokens from specific linguistic environments. For example, when sufficient speech data is available, one might decide to sample only from consonants in coda position. However, the reality in forensic speech comparisons is that speech evidence can often be short and taking

acoustic measurements for a segment in specific linguistic environments might simply not be possible due to scarcity of material. For forensic speech science we can thus conclude that sampling from specific linguistic contexts may yield some small benefits with regards to the strength of evidence, although differences are generally too small to make a difference for the conclusions of forensic speaker comparisons, which will often be expressed in verbal terms for interpretation in court. The Netherlands Forensic Institute's guidelines for interpreting the strength of evidence as derived with likelihood ratios in the Bayesian method includes a six-point scale of LR ranges and corresponding verbal conclusions. The evidence can be 'about equally probable' under either hypothesis, up to 'extremely more probable' under one of them. Using these labels, the probability of the evidence under the same-speaker hypothesis and under the different-speaker hypothesis can be related to one another, allowing for conclusions in both directions (Nederlands Forensisch Instituut, 2017). The likelihood ratios obtained in chapter 5 generally do not change the strength of evidence according to the six-point scale, or at least not more than one scale, which mostly occurred in cases where there was also a discrepancy in how many tokens were included per speaker. To conclude, not considering linguistic environment when sampling tokens (in this dissertation restricted to syllabic position and phonetic context effects on fricative and nasal consonants) does not seem to have overly large consequences on forensic speaker comparisons.

Rather, including more tokens might be more beneficial than sampling from specific contexts. In chapter 5, it was shown that, for /s/, there is better performance when all available tokens are included, maximizing the number of tokens. For /n/, on the other hand, sampling only from coda position yields higher performance than when all available tokens are included. This seems to be related to the different types of speaker information available in these sounds. Fricative /s/ is associated with several social variables and displays more between-speaker variation, whereas nasal /n/ shows relatively little influence from social (or even linguistic) variables and displays less within-speaker variation, i.e., is more stable within speakers even using smaller samples. Although both perform similarly when all available tokens are included

(even showed a small advantage for /s/), /n/ is clearly preferable when materials are scarce.

In chapter 4, the effect of the telephone filter on the amount of speaker information was examined, also including the different linguistic contexts. Both fricative /s/ and nasal /n/ were expected to show effects of the landline telephone filter. The former because its spectral peak falls outside of the upper limit of the filter and the latter because its main spectral characteristic – the first nasal formant – falls (partly) below the lower limit of the filter, leaving only the second to third (or fourth) nasal formants to be measured reliably. In chapter 4, the effect of the landline telephone filter on fricative /s/ was tested, which arguably constitutes a worst-case scenario due to both the high-frequency nature of /s/ and the small range of the landline filter compared to more modern mobile filters. Acoustic results showed that, even when taking the same measurement range (550 – 3400 Hz) from parallel studio and telephone recordings, significant acoustic differences were found in acoustic-phonetic features. This means that simulating a telephone filter by simply narrowing the frequency range in the studio recording does not approach the acoustics of the telephone filter. Landline telephone recordings have a 300 – 3400 Hz bandpass but usually show signal from 0 – 4000 Hz[11]. This is because bandpass filters are not rectangular, but rounded at the edges, resulting in attenuated signal outside the 300 – 3400 Hz band pass. That there are significant differences between recording types even when measuring within that band pass indicates that the signal within the bandpass displays additional effects. Most obviously, the telephone hardware and the different positioning of speaking into telephones compared to microphones could affect the acoustics. However, it could also be signal-related as captured in specific telephone filter algorithms.

For English /s/, the acoustic differences between linguistic contexts were neutralized by the landline filter. On the one hand, this can be regarded as positive, as linguistic contexts therefore do not need to be taken into account. On the other hand, it clearly indicates that this speech sound is acoustically compromised by the telephone filter, neutralizing

[11] Note that there is some variation in landline filters across countries; this is the band pass in the Netherlands.

both linguistic and speaker information. Previous research has already looked at vowel formants, for which telephone filter effects were predictably smaller than for /s/. Future research should include more consonantal speech sounds, to get a more complete view of telephone effects on forensic speaker comparisons using auditory-acoustic methods. From a sociolinguistic perspective, it would be interesting to see how different telephone filters affect the production and perception of social variables such as gender identity and sexual orientation. The current results on English /s/ would imply that perceiving such information from telephone acoustics might be more difficult.

6.2.3 Limitations

In this section, some of the limitations of the current work will be discussed.

Firstly, the data analyzed in this dissertation comes from pre-existing speech corpora. The Dutch data in particular, from the Spoken Dutch Corpus (Oostdijk, 2000), was recorded around two decades ago, which potentially makes it somewhat dated with regards to ongoing sound changes in Dutch such as fricative devoicing beyond the coda position (Pinget, Van de Velde, & Kager, 2014). With regards to the devoicing trend in particular, when two sounds are in the process of merging, speakers often display more variation, resulting in more or less variation for the sounds in question – here /s/-/z/ – in a set of speakers, which may affect speaker discrimination.

Another limitation in the Dutch data is the uncontrolled recording circumstances. The telephone conversations in the Spoken Dutch Corpus were recorded by wiretapping the landline telephones in speakers' own homes, presumably using their own telephones. One advantage of this corpus is that speakers converse with speakers that are known to them on any topic of their choosing (participants were asked to converse about anything they wanted for about ten minutes). As a result, the conversations contain natural speech in informal register that reflects everyday communications between speakers. One major disadvantage is

that it has to be assumed that speakers used different telephones, namely the landline in their own home, although the documentation of the corpus is not entirely clear on this. This means that it is possible that the acoustics possibly contain some idiosyncrasies that are not necessarily dependent on the speaker, but on the specific telephone that was used, the quality of the wiretapped signal, and the specific background noises in the speaker's home. Note that this does not include different phone-holding behaviors, which can also affect the acoustics but are more speaker-dependent in nature. Examples of background noises include a crying baby in the background and a pet bird. In the data annotation, tokens with audible background noise were excluded from analysis, but it is still possible that the general acoustics of the space of each speaker exerted some influence on the recordings and the speech sound acoustics that were analyzed in this dissertation. This was deemed somewhat acceptable because the research questions in this dissertation focused on the effects of linguistic factors *within* speakers and not so much on building the best-performing speaker discrimination system possible.

The English data from the West Yorkshire Regional English Database (WYRED, Gold et al., 2018) does not have these specific limitations, as recording conditions were much more controlled. Each speaker was recorded in the lab using the same recording equipment. Although this corpus is more contemporary, it only includes speakers from a particular dialect area in England (in this dissertation, only the speakers from Wakefield in Yorkshire were included, as region was not of particular interest). It is therefore potentially only representative for Yorkshire English (as spoken in Wakefield).

For both the Spoken Dutch Corpus and WYRED, only contemporaneous data was used, which, for any speaker, should be assumed to underestimate the within-speaker variability. Although one to four telephone conversations from the Spoken Dutch Corpus were used in the analyses presented in this dissertation, it is not clear to what degree these recordings are non-contemporaneous as only the recording year is available in the meta data. From the content of the conversations, some seem to have taken place consecutively, making them contemporaneous. Another possible disadvantage for both corpora regards the use of the landline telephone. Mobile phones have risen in popularity the past two

decades and are probably more representative for telephone communications currently. Mobile signals have a larger bandpass and varying bit rates, which gives the signal better quality, but only variably so. However, as mentioned in the introduction of this dissertation, the use of burner phones by criminals, which are likely not compatible with newer generation mobile networks, result in many wiretapped signals that are highly similar to landline signals. Nevertheless, future work should consider the effects of mobile telephone filters on different consonantal speech sounds, also examining the interactions with linguistic factors that were found in chapter 4.

Lastly, it should be mentioned that the use of rather simplistic acoustic-phonetic features in the current dissertation is a possible limitation. Measurements such as spectral moments for fricatives and nasal formants for nasals were used to be able to compare current findings to previous phonetic research. Importantly, these rather simple measurements are relatively easy to measure and easy to interpret, as they have clear associations with vocal tract configurations. This is desirable in auditory-acoustic forensic speaker comparisons, where practitioners may have to be able to explain the speech evidence in court, for which permissible evidence depends on the specific legal context of that country. Importantly, these measurements seemed to adequately capture the linguistic effects that were of interest in this work. Having stated that, it is possible that some between-speaker variation in the sounds examined here is captured in more detail using acoustic measures with higher dimensionality, such as discrete cosine transforms (DCT: Jannedy & Weirich, 2017) or mel-frequency cepstral coefficients (MFCC), which are often used in automatic speaker recognition. To conclude, future work should consider extending the current line of research to using more advanced acoustic-phonetic measurements on contemporary speech data that include more contemporary (modern) telephone signals.

Appendix

Appendix A. *Fixed effects in best-fitting linear mixed-effects models for English /s/ including an additional factor for measurements taken in the studio recording over the 550 – 3,400 Hz bandwidth (N = 60, n = 6,000, default factor level = Telephone: 550 – 3,400 Hz).*

	M1 [Hz]				M2 [Hz]			
Effect	*Est.*	*SE*	*t*	*p*	*Est.*	*SE*	*t*	*p*
(intercept)	2,078	31	66.9	***	732	6	126.2	***
Channel: Studio (550-8000 Hz)	2,943	95	30.8	***	536	26	20.9	***
Channel: Studio (550-3400 Hz)	225	16	13.9	***	19	8	2.4	0.017
	L3				L4			
(intercept)	1.53	0.07	20.9	***	36.33	1.62	22.5	***
Channel: Studio (550-8000 Hz)	–1.34	0.13	–9.9	***	– 32.48	1.30	– 25.03	***
Channel: Studio (550-3400 Hz)	–0.43	0.07	–6.6	***	–1.45	1.30	–1.12	0.263
	dynamic $M1^{linear}$ [Hz]				dynamic $M1^{quadratic}$ [Hz]			
(intercept)	76	15	5.0	***	–194	10	–19.9	***
Channel: Studio (550-8000 Hz)	–81	25	–3.2	0.001	–545	7	–75.0	***
Channel: Studio (550-3400 Hz)	–38	16	–2.4	0.018	–31	7	–4.2	**
	Peak [Hz]				Tilt [dB/decade]			
(intercept)	2,108	50	42.2	***	3.50	0.85	4.1	**
Channel: Studio (550-8000 Hz)	2,669	130	20.5	***	11.86	1.22	9.7	***
Channel: Studio (550-3400 Hz)	335	31	10.8	***	5.73	0.55	10.5	***

Bibliography

Aitken, C. G. G., & Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *Journal of the Royal Statistical Society. Series C: Applied Statistics, 53*(1), 109–122. https://doi.org/10.1046/j.0035-9254.2003.05271.x

Amino, K., & Arai, T. (2009). Speaker-dependent characteristics of the nasals. *Forensic Science International, 185*(1–3), 21–28. https://doi.org/10.1016/J.FORSCIINT.2008.11.018

Amino, K., Arai, T., & Sugawara, T. (2007). Effects of the phonological contents on perceptual speaker identification. In C. Müller (Ed.) *Speaker Classification II: Selected papers. Lecture Notes in Computer Science* (Vol. 4441 LNAI, pp. 83–92). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-74122-0_8

Audibert, N., Fougeron, C., & Chardenon, E. (2021, February 4-5). Do you remain the same speaker over 21 recordings? [Conference presentation]. *XVII National congress of the Italian Association of Speech Science (AISV)*, Zürich, Switzerland. Retrieved from https://www.cl.uzh.ch/dam/jcr:e6a8ac08-5b5b-4bba-9146-ac6d1256d72c/AISV2021_book%20of%20abstracts_v2.pdf.

Baayen, R. H., Piepenbrock, R., & Van Rijn, H. (1993). *The CELEX Lexical Database*. Philadelphia Linguistics Data Consortium, University of Pennsylvania. https://doi.org/10.1016/j.bbrc.2013.10.120

Bang, H.-Y., Clayards, M., & Goad, H. (2017). Compensatory Strategies in the Developmental Patterns of English /s/: Gender and Vowel Context Effects. *Journal of Speech Language and Hearing Research, 60*(3), 571–591. https://doi.org/10.1044/2016_JSLHR-L-15-0381

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3). https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint* arXiv:1506.04967.

Bell, A. (1984). Language style as audience design. *Language in society, 13*(2), 145-204. https://doi.org/10.1017/S004740450001037X

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language, 60*(1), 92-111. https://doi.org/10.1016/j.jml.2008.06.003

Bell-Berti, F., & Harris, K. S. (1979). Anticipatory coarticulation: Some implications from a study of lip rounding. *Journal of the Acoustical Society of America, 65*(5), 1268–1270. https://doi.org/10.1121/1.382794

Bell-Berti, F., & Harris, K. S. (1982). Temporal patterns of coarticulation: Lip rounding. *Journal of the Acoustical Society of America, 71*(2), 449–454. https://doi.org/10.1121/1.387466

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., & Wellekens, C. (2007). Automatic speech recognition and speech variability: A review. *Speech communication*, *49*(10-11), 763-786. https://doi.org/10.1016/j.specom.2007.02.006

Bessette, B., Salami, R., Lefebvre, R., Jelínek, M., Rotola-Pukkila, J., Vainio, J., Mikkola, H., Järvinen, K. (2002). The Adaptive Multirate Wideband speech codec (AMR-WB). *IEEE Transactions on Speech and Audio Processing, 10*(8), 620–636. https://doi.org/10.1109/TSA.2002.804299

Biber, D., & Conrad, S. (2005). Register variation: A corpus approach. In D. Schiffrin, D. Tannen, & H E. Hamilton (Eds.) *The handbook of discourse analysis* (pp. 175-196). Blackwell Publishers Ltd. https://doi.org/10.1002/9780470753460.ch10

Boersma, P. & Weenink, D. (2020). Praat: doing phonetics by computer [Computer program]. Version 6.0.40, retrieved from http://www.praat.org/

Bosma, W., Dalm, S., van Eijk, E., El Harchaoui, R., Rijgersberg, E., Tops, H. T., Veenstra, A. & Ypma, R. (2020). Establishing phone-pair co-usage by comparing mobility patterns. *Science & Justice, 60*(2), 180-190. https://doi.org/10.1016/j.scijus.2019.10.005

Bradlow, A. R., Nygaard, L. C., & Pisoni, D. B. (1999). Effects of talker, rate, and amplitude variation on recognition memory for spoken words. *Perception & Psychophysics, 61*(2), 206–219. https://doi.org/10.3758/BF03206883

Brümmer, N., & Du Preez, J. (2006). Application-independent evaluation of speaker detection. *Computer Speech and Language, 20*(2-3), 230–275. https://doi.org/10.1016/j.csl.2005.08.001

Bürki, A. (2018). Variation in the speech signal as a window into the cognitive architecture of language production. *Psychonomic Bulletin & Review, 25*(6), 1973–2004. https://doi.org/10.3758/s13423-017-1423-4

Byrd, D., Tobin, S., Bresch, E., & Narayanan, S. (2009). Timing effects of syllable structure and stress on nasals: A real-time MRI examination. *Journal of Phonetics, 37*(1), 97–110. https://doi.org/10.1016/J.WOCN.2008.10.002

Byrd, D., & Saltzman, E. (2003). The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics, 31*(2), 149-180. https://doi.org/10.1016/S0095-4470(02)00085-2

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

Cho, T., & McQueen, J. M. (2005). Prosodic influences on consonant production in Dutch: Effects of prosodic boundaries, phrasal accent and lexical stress. *Journal of Phonetics*, *33*(2), 121–157. https://doi.org/10.1016/j.wocn.2005.01.001

Cho, T., Lee, Y., & Kim, S. (2014). Prosodic strengthening on the /s/-stop cluster and the phonetic implementation of an allophonic rule in English. *Journal of Phonetics, 46*, 128-146. https://doi.org/10.1016/j.wocn.2014.06.003

Cholin, J., Levelt, W. J., & Schiller, N. O. (2006). Effects of syllable frequency in speech production. *Cognition, 99*(2), 205-235. https://doi.org/10.1016/j.cognition.2005.01.009

Collins, B., & Mees, I. M. (1984). *The sounds of English and Dutch* (5th ed.). Leiden, Netherlands: Brill Archive.

Cunha, C., & Reubold, U. (2015). The contribution of vowel coarticulation and prosodic weakening in initial and final fricatives to sound change. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, UK: the University of Glasgow. ISBN 978-0-85261-941-4. Paper number 0979 retrieved from http://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0979.pdf

De Boer, M. M., & Heeren, W. F. L. (2021, February 4-5). Language-dependency of /m/ in L1 Dutch and L2 English [Conference presentation]. *XVII National congress of the Italian Association of Speech Science (AISV)*, Zürich, Switzerland. Retrieved from https://www.cl.uzh.ch/dam/jcr:e6a8ac08-5b5b-4bba-9146-ac6d1256d72c/AISV2021_book%20of%20abstracts_v2.pdf.

De Boer M. M., Quené H. & Heeren W. F. L. (2022), Long-term within-speaker consistency of filled pauses in native and non-native speech, *JASA Express Letters 2*(3): 035201.

Ditewig, S., Pinget, A. C. H., & Heeren, W. F. L. (2019). Regional variation in the pronunciation of /s/ in the Dutch language area. *Nederlandse Taalkunde, 24*(2), 195–212.

Ditewig, S., Smorenburg, L., Quené, H., & Heeren, W. (2021). An acoustic-phonetic study of retraction of /s/ in Moroccan Dutch and endogenous Dutch. *Nederlandse Taalkunde, 26*(3), 315–338. https://doi.org/10.5117/NEDTAA2021.3.001.DITE

Dumay, N., Content, A., & Frauenfelder, U. (1999). Acoustic-phonetic cues to word boundary location: Evidence from word spotting. In J. O. Ohala, Y. Hasegawa, M. Ohala, D. Granville, & A. C. Bailey (Eds.) *Proceedings of the 14th international congress of phonetic sciences* (pp. 281-284). San Fransisco, USA: University of California. Retrieved

from https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS1999/p14_0281.html

Eaton, J. W., Bateman, D., Hauberg, S., & Wehbring, R. (2019). GNU Octave version 5.2.0 manual: a high-level interactive language for numerical computations [Computer program]. Retrieved from https://www.gnu.org/software/octave/doc/v5.2.0/

Fant, G. (1970). *Acoustic theory of speech production (2nd ed.)*. The Hague: Mouton.

Forrest, K., Weismer, G., Milenkovic, P., & Dougall, R. N. (1988). Statistical analysis of word-initial voiceless obstruents: Preliminary data. *Journal of the Acoustical Society of America, 84*(1), 115–123. https://doi.org/10.1121/1.396977

Fougeron, C. (2001). Articulatory properties of initial segments in several prosodic constituents in French. *Journal of Phonetics, 29*(2), 109–135. https://doi.org/10.1006/JPHO.2000.0114

Fougeron, C., & Keating, P. A. (1998). Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America, 101*(6), 3728. https://doi.org/10.1121/1.418332

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*(1), 1–22. https://doi.org/10.18637/jss.v033.i01

Fuchs, S., & Toda, M. (2010). Do differences in male versus female /s/ reflect biological or sociophonetic factors? In S. Fuchs, M. Toda, & M. Żygis (Eds.), *Turbulent sounds: an interdisciplinary guide* (pp. 281–302). Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783110226584.281

Fuchs, S. (2022, July 10-13). Flexibility and stability of respiration in human actions [Conference presentation]. *The 30th annual IAFPA conference*, Prague, Czech republic.

Fujimura, O. (1962). Analysis of Nasal Consonants. *Journal of the Acoustical Society of America, 34*(12), 1865–1875. https://doi.org/10.1121/1.1909142

Glenn, J. W., & Kleiner, N. (1968). Speaker Identification Based on Nasal Phonation. *Journal of the Acoustical Society of America, 43*(2), 368–372. https://doi.org/10.1121/1.1910788

Gold, E., & French, P. (2011). International practices in forensic speaker comparison. *International Journal of Speech, Language and the Law, 18*(2), 293–307. https://doi.org/10.1558/ijsll.v18i2.293

Gold, E., & French, P. (2019). International practices in forensic speaker comparisons: Second survey. *International Journal of Speech, Language and the Law, 26*(1), 1–20. https://doi.org/10.1558/ijsll.38028

Gold, E., Ross, S., & Earnshaw, K. (2018). The "West Yorkshire Regional English Database": Investigations into the generalizability of reference populations for forensic speaker comparison casework. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* (Vol. 2018, pp. 2748–2752). International Speech Communication Association. https://doi.org/10.21437/Interspeech.2018-65

Goldinger, S. D. (1998). Echoes of echoes? an episodic theory of lexical access. *Psychological Review, 105*, 251-279. https://doi.org/10.1037/0033-295X.105.2.251

Gow, D. W., Melvold, J., & Manuel, S. (1996). How word onsets drive lexical access and segmentation: Evidence from acoustics, phonology and processing. In *Proceedings of International Conference on Spoken Language Processing ICSLP'96* (Vol. 1, pp. 66-69), Philadelphia, PA, USA: IEEE. https://doi.org/10.1109/icslp.1996.607031

Guillemin, B. J., & Watson, C. I. (2006). Impact of the GSM AMR Speech Codec on Formant Information Important to Forensic Speaker Identification. In P. Warren & C. I. Watson (Eds.) *Proceedings of the 11th Australian International Conference on Speech Science & Technology* (pp. 483-488). Australian Speech Science & Technology Association Inc.

Gussenhoven, C. (1999). Dutch. In *Handbook of the international phonetic association* (pp. 74–77). Cambridge: Cambridge University Press.

Hardcastle, W. J. (1985). Some phonetic and syntactic constraints on lingual coarticulation during /kl/ sequences. *Speech Communication, 4*(1–3), 247–263. https://doi.org/10.1016/0167-6393(85)90051-2

Harst, S. Van der, Velde, H. Van de, & Schouten, B. (2007). Acoustic characteristics of Standard Dutch /x/. In J. Trouvain & Barry, W. J. (Eds.) *Proceedings of the 16th International Congress of Phonetic Sciences* (pp. 1469–1472). Saarbrücken, Germany. Retrieved from http://www.icphs2007.de/conference/Papers/1479/1479.pdf

Hawthorne, K., Mazuka, R., & Gerken, L. (2015). The acoustic salience of prosody trumps infants' acquired knowledge of language-specific prosodic patterns. *Journal of Memory and Language*, *82*, 105-117.

He, L., & Dellwo, V. (2017). Between-speaker variability in temporal organizations of intensity contours. *Journal of the Acoustical Society of America, 141*(5), EL488-EL494. https://doi.org/10.1121/1.4983398

He, L., Zhang, Y., & Dellwo, V. (2019). Between-speaker variability and temporal organization of the first formant. *Journal of the Acoustical Society of America, 145*(3), EL209-EL214. https://doi.org/10.1121/1.5093450

Heeren, W. F. L. (2018). Does a token's linguistic context affect its speaker-dependent information? [Conference presentation]. *Proceedings of the 26th IAFPA*, (pp 31–32), Huddersfield, UK. Retrieved from https://www.iafpa.net/tag/2018/

Heeren, W. F. L. (2020a). The effect of word class on speaker-dependent information in the Standard Dutch vowel /a:/. *Journal of the Acoustical Society of America, 148*(4), 2028–2039. https://doi.org/10.1121/10.0002173

Heeren, W. F. L. (2020b). The contribution of dynamic versus static formant information in conversational speech. *International Journal of Speech, Language & the Law, 27*(1), 75—98. https://doi.org/10.1558/ijsll.41058

Herrmann, F., Whiteside, S. P., & Cunningham, S. (2008, June). An acoustic investigation into coarticulation and speech motor control: high vs. low frequency syllables. *Proc. Mtgs. Acoust., 4*(1), 060007. https://doi.org/10.1121/1.3085742

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

Holliday, J. J., Reidy, P. F., Beckman, M. E., & Edwards, J. (2015). Quantifying the Robustness of the English Sibilant Fricative Contrast in Children. *Journal of Speech, Language, and Hearing Research, 58*(3), 622–637. https://doi.org/10.1044/2015_JSLHR-S-14-0090

Hoole, P., Nguyen-Trong, N., & Hardcastle, W. (1993). A comparative investigation of coarticulation in fricatives: Electropalatographic,

electromagnetic, and acoustic data. *Language and Speech, 36*(2–3), 235–260. https://doi.org/10.1177/002383099303600307

Hughes, V., & Foulkes, P. (2015). The relevant population in forensic voice comparison: Effects of varying delimitations of social class and age. *Speech Communication, 66*, 218-230.

Ingemann, F. (1968). Identification of the Speaker's Sex from Voiceless Fricatives. *Journal of the Acoustical Society of America, 44*(4), 1142–1144. https://doi.org/10.1121/1.1911208

Jang, J., Kim, S., & Cho, T. (2018). Focus and boundary effects on coarticulatory vowel nasalization in Korean with implications for cross-linguistic similarities and differences. *Journal of the Acoustical Society of America, 144*(1), EL33-EL39. https://doi.org/10.1121/1.5044641

Jannedy, S., & Weirich, M. (2017). Spectral moments vs discrete cosine transformation coefficients: Evaluation of acoustic measures distinguishing two merging German fricatives. *Journal of the Acoustical Society of America, 142*(1), 395–405. https://doi.org/10.1121/1.4991347

Johnson, K. (2003). *Acoustic and Auditory Phonetics (2nd ed.)*. Oxford: Blackwell. https://doi.org/10.2307/417784

Johnson, K. (2005). Speaker normalization in speech perception. In D. B. Pisoni & R. E. Remez (Eds.), *The handbook of speech perception* (p. 363-389). Oxford: Blackwell.

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *Journal of the Acoustical Society of America, 108*(3), 1252. https://doi.org/10.1121/1.1288413

Jovičić, S. T., Jovanović, N., Subotić, M., & Grozdić, Đ. (2015). Impact of mobile phone usage on speech spectral features: Some preliminary findings. *International Journal of Speech, Language and the Law, 22*(1). https://doi.org/10.1558/ijsll.v22i1.17880

Junqua, J. (1993). The Lombard reflex and its role on human listeners and automatic speech recognizers. *Journal of the Acoustical Society of America, 93*(1), 510–524. https://doi.org/10.1121/1.405631

Junqua, J.-C., Fincke, S., & Field, K. (1999). The Lombard effect: a reflex to better communicate with others in noise. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 2083–2086). IEEE. https://doi.org/10.1109/icassp.1999.758343

Kachkovskaia, T., Menshikova, A., Kocharov, D., Kholiavin, P., Mamushina, A. (2022, May 23-26). Social and situational factors of speaker variability in collaborative dialogues [Conference presentation]. *Proceedings of Speech Prosody 2022* (pp. 455-459) https://doi.org/10.21437/SpeechProsody.2022-93

Kavanagh, C. M. (2012). New consonantal acoustic parameters for forensic speaker comparison [Doctoral dissertation]. York, UK: University of York.

Klecka, W. R. (1980). Discriminant Analysis. In *Quantitative Applications in the Social Sciences* (Vol. 19). London: Sage Publications.

Kleinschmidt, D. F. (2019). Structure in talker variability: How much is there and how much can it help? *Language, cognition and neuroscience, 34*(1), 43-68. https://doi.org/10.1080/23273798.2018.1500698

Koenig, L. L., Shadle, C. H., Preston, J. L., & Mooshammer, C. R. (2013). Toward improved spectral measures of /s/: Results from adolescents. *Journal of Speech Language and Hearing Research, 56*(4), 1175. https://doi.org/10.1044/1092-4388(2012/12-0038)

Krakow, R. A. (1993). Nonsegmental influences on velum movement patterns: Syllables, sentences, stress, and speaking rate. In M. K. Huffman & R. A. Krakow (Eds.) *Phonetics and Phonology: Nasals, Nasalization, and the Velum, Vol. 5* (pp. 87–116). Academic Press. https://doi.org/10.1016/b978-0-12-360380-7.50008-9

Künzel, H. J. (2001). Beware of the "telephone effect": the influence of telephone transmission on the measurement of formant frequencies. *Forensic Linguistics 8*(1), 80–99.

Kurowski, K., & Blumstein, S. E. (1984). Perceptual integration of the murmur and formant transitions for place of articulation in nasal consonants. *Journal of the Acoustical Society of America 76*(2), 383–390. https://doi.org/10.1121/1.391139

Kurowski, K., & Blumstein, S. E. (1987). Acoustic properties for place of articulation in nasal consonants. *Journal of the Acoustical Society of America, 81*(6), 1917–1927. https://doi.org/10.1121/1.394756

Levelt, W. J., & Wheeldon, L. (1994). Do speakers have access to a mental syllabary? *Cognition, 50*(1-3), 239-269.

Li, F., Rendall, D., Vasey, P. L., Kinsman, M., Ward-Sutherland, A., & Diano, G. (2016). The development of sex/gender-specific /s/ and its relationship to gender identity in children and adolescents. *Journal of Phonetics, 57*, 59–70. https://doi.org/10.1016/J.WOCN.2016.05.004

Lo, J. J. H. (2021). Seeing the trees in the forest: Diagnosing individual performance in likelihood ratio based forensic voice comparison. In C. Bernardasci, D. Dipino, D. Garassino, S. Negrinelli, E. Pellegrino, & S. Schmid (Eds.) *L'individualità del parlante nelle scienze fonetiche: applicazioni tecnologiche e forensi [Speaker individuality in phonetics and speech sciences: speech technology and forensic applications] (Studi AISV8*, pp. 77–96. Italy, Officinaventuno. https://doi.org/10.17469/O2108AISV000004

Luyckx, K., Kloots, H., Coussé, E., & Gillis, S. (2007). Klankfrequenties in het Nederlands. In *Tussen taal, spelling en onderwijs. Essays bij het emeritaat van Frans Daems* (pp. 141–154). Academia Press.

Magen, H. S. (1997). The extent of vowel-to-vowel coarticulation in English. *Journal of Phonetics, 25*(2), 187—205. https://doi.org/10.1006/jpho.1996.0041

Maniwa, K., Jongman, A., & Wade, T. (2009). Acoustic characteristics of clearly spoken English fricatives. *Journal of the Acoustical Society of America, 125*(6), 3962–3973. https://doi.org/10.1121/1.2990715

Marslen-Wilson, W., & Zwitserlood, P. (1989). Accessing Spoken Words: The Importance of Word Onsets. *Journal of Experimental Psychology: Human Perception and Performance, 15*(3), 576. https://doi.org/10.1037/0096-1523.15.3.576

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.

McDougall, K. (2006). Dynamic features of speech and the characterization of speakers: Towards a new approach using formant frequencies. International *Journal of Speech, Language and the Law, 13*(1), 89–125. https://doi.org/10.1558/sll.2006.13.1.89

Mermelstein, P. (1977). On detecting nasals in continuous speech. *Journal of the Acoustical Society of America, 61*(2), 581–587. https://doi.org/10.1121/1.381301

Monson, B. B., Lotto, A. J., & Story, B. H. (2012). Analysis of high-frequency energy in long-term average spectra of singing, speech, and voiceless fricatives. *Journal of the Acoustical Society of America, 132*(3), 1754–1764. https://doi.org/10.1121/1.4742724

Morrison, G. S. (2007). Matlab implementation of Aitken & Lucy's (2004) forensic likelihood ratio software using multivariate-kernel-density estimation. [Computer program].

Morrison, G. S., Sahito, F. H., Jardine, G., Djokic, D., Clavet, S., Berghs, S., & Dorny, C. G. (2016). INTERPOL survey of the use of speaker identification by law enforcement agencies. *Forensic science international*, *263*, 92-100.

Mourigh, K. (2017). Stance-taking through sibilant palatalisation in Gouda Moroccan Dutch. *Nederlandse Taalkunde, 22*(3), 421–446. https://doi.org/10.5117/nedtaa2017.3.mour

Munson, B. (2001). A method for studying variability in fricatives using dynamic measures of spectral mean. *Journal of the Acoustical Society of America, 110*(2), 1203–1206. https://doi.org/10.1121/1.1387093

Munson, B. (2004). Variability in /s/ Production in Children and Adults. *Journal of Speech Language and Hearing Research, 47*(1), 58–69. https://doi.org/10.1044/1092-4388(2004/006)

Munson, B., McDonald, E. C., DeBoe, N. L., & White, A. R. (2006). The acoustic and perceptual bases of judgments of women and men's sexual orientation from read speech. *Journal of Phonetics, 34*(2), 202–240. https://doi.org/10.1016/J.WOCN.2005.05.003

Nederlands Forensisch Instituut (2020, February). Vakbijlage – Vergelijkend spraakonderzoek (version October 2016). Ministerie van Veiligheid en Justitie. Retrieved from https://www.forensischinstituut.nl/publicaties/publicaties/2020/02/03/vakbijlage-vergelijkend-spraakonderzoek

Nederlands Forensisch Instituut (2017, May). Vakbijlage – De reeks waarschijnlijkheidstermen van het NFI en het Bayesiaanse model voor interpretatie van bewijs (version 2.2). Ministerie van Veiligheid en Justitie. Retrieved from https://www.forensischinstituut.nl/publicaties/publicaties/2017/10/18/vakbijlage-waarschijnlijkheidstermen

Niebuhr, O., Clayards, M., Meunier, C., & Lancia, L. (2011). On place assimilation in sibilant sequences—Comparing French and English. *Journal of Phonetics, 39*(3), 429–451. https://doi.org/10.1016/J.WOCN.2011.04.003

Nittrouer, S., & Whalen, D. H. (1989). The perceptual effects of child–adult differences in fricative-vowel coarticulation. *Journal of the Acoustical Society of America, 86*(4), 1266–1276. https://doi.org/10.1121/1.398741

Noiray, A., Cathiard, M. A., Abry, C., & Ménard, L. (2010). Lip rounding anticipatory control: Crosslinguistically lawful and ontogenetically attuned. In B. Maassen & P. van Lieshout (eds.) *Speech motor control: New developments in basic and applied research* (pp. 153-171). Oxford: Oxford University Press.

Nolan, F. (1983). The Phonetic Bases of Speaker Recognition. *Cambridge Studies in Speech Science and Communication*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1121/1.392415

Nolan, F. (2001). Speaker identification evidence: its forms, limitations, and roles. In *Law and Language: Prospect and Retrospect* (pp. 1–19). Levi, Lappland. Retrieved from http://www.ling.cam.ac.uk/francis/lawlang.doc

Nortier, J., & Dorleijn, M. (2008). A Moroccan accent in Dutch: A sociocultural style restricted to the Moroccan community?. *International Journal of Bilingualism, 12*(1-2), 125-142.

Odinot, G., Jong, D. de, Leij, J. B. J. van der, Poot, C. J. de, & Straalen, E. K. van. (2010). *Het gebruik van de telefoon- en internettap in de opsporing (Onderzoek)*. Meppel: Boom Lemma uitgevers. Retrieved from https://repository.tudelft.nl/view/wodc/uuid:a4b1041c-0af4-4b30-bca2-ecc28dd79c8d

Ohala, J. J., & Kawasaki, H. (1984). *Prosodic phonology and phonetics*. *Phonology, 1*, 113–127. https://doi.org/10.1017/S0952675700000312

Oliveira, M., & Freitas, T. (2008). Intonation as a cue to turn management in telephone and face-to-face interactions. In *Proceedings of Speech Prosody* (pp. 485–488). Campinas, Brazil: ISCA. Retrieved from https://www.isca-speech.org/archive/speechprosody_2008/oliveirajr08_speechprosody.html

Oostdijk, N. H. J. (2000). Corpus Gesproken Nederlands. *Nederlandse Taalkunde, 5*(3), 280–284. Retrieved from http://repository.ubn.ru.nl/handle/2066/76339

Penhallurick, R. (2008). Welsh English: Phonology. In B. Kortmann & C. Upton (eds.) *Varieties of English: The British Isles* (pp. 105-121). Berlin: Mouton de Gruyter.

Perkell, J. S., & Matthies, M. L. (1992). Temporal measures of anticipatory labial coarticulation for the vowel /u/: Within- and cross-subject variability. *Journal of the Acoustical Society of America, 91*(5), 2911–2925. https://doi.org/10.1121/1.403778

Pinget, A.-F., Van de Velde, H., & Kager, R. (2014). Cross-regional differences in the perception of fricative devoicing. In J. Caspers, J. Pacilly, W. Heeren & Y. Chen (Eds.) *Above and Beyond the Segments* (pp. 230–245). Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/z.189.19pin

Pisanski, K., Nowak, J., & Sorokowski, P. (2016). Individual differences in cortisol stress response predict increases in voice pitch during exam stress. *Physiology & behavior*, *163*, 234-238.

Quené, H., Orr, R., & Van Leeuwen, D. (2017). Phonetic similarity of /s/ in native and second language: Individual differences in learning curves. *Journal of the Acoustical Society of America, 142*, 525. https://doi.org/10.1121/1.5013149

R Core Team. (2019). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Recasens, D., & Dolorspallarè, M. (2001). Coarticulation, assimilation and blending in Catalan consonant clusters. *Journal of Phonetics, 29*, 273–301. https://doi.org/10.006/jpho.2001.0139

Recasens, D. (2004). The effect of syllable position on consonant reduction (evidence from Catalan consonant clusters). *Journal of Phonetics, 32*(3), 435–453. https://doi.org/10.1016/J.WOCN.2004.02.001

Redford, M. A., & Diehl, R. L. (1999). The relative perceptual distinctiveness of initial and final consonants in CVC syllables. *Journal of the Acoustical Society of America, 106*(3), 1555. https://doi.org/10.1121/1.427152

Reynolds, D. A. (1995). Large Population Speaker Identification Using Clean and Telephone Speech. *IEEE Signal Processing Letters, 2*(3), 46–48. https://doi.org/10.1109/97.372913

Rose, P. (2002). *Forensic Speaker Identification*. London, UK: CRC Press. https://doi.org/10.1201/9780203166369

Rowe, B. M., & Levine, D. P. (2018). *A concise introduction to linguistics* (5th ed.). London: Routledge. https://doi.org/10.4324/9781315227283.

Rudzicz, F. (2007). Comparing speaker-dependent and speaker-adaptive acoustic models for recognizing dysarthric speech. In *Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility* (pp. 255-256). https://doi.org/10.1145/1296843.1296899

Schilling, N. (2004). Investigating stylistic variation. In J. K. Chambers, P. Trudgill, & N. Schilling (Eds.), *The handbook of language variation and change* (pp. 375-401). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118335598.ch15

Scobbie, J.M. (2012). Interface and Overlap in Phonetics and Phonology. In G. Ramchand & C. Reiss (eds.) *The Oxford Handbook of Linguistic Interfaces* (pp. 17-52). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199247455.013.0002

Saon, G., Soltau, H., Nahamoo, D., & Picheny, M. (2013, December). Speaker adaptation of neural network acoustic models using i-vectors. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding* (pp. 55-59). IEEE.

Schindler, C., & Draxler, C. (2013). Using spectral moments as a speaker specific feature in nasals and fricatives. In *Proceedings of INTERSPEECH* (pp. 2793–2796). Lyon, France: ISCA. https://doi.org/10.21437/interspeech.2013-639

Schwartz, M. F. (1968). Identification of Speaker Sex from Isolated, Voiceless Fricatives. *Journal of the Acoustical Society of America, 43*(5), 1178–1179. https://doi.org/10.1121/1.1910954

Seitz, P. F., McCormick, M. M., Watson, I. M. C., & Bladon, R. A. (1990). Relational spectral features for place of articulation in nasal consonants. *Journal of the Acoustical Society of America, 87*(1), 351–358. https://doi.org/10.1121/1.399256

Selkirk, E. (1996). The Prosodic Structure of Function Words. In J. L. Morgan & K. Demuth (Eds.), *Signal to Syntax: Bootstrapping From Speech To Grammar in Early Acquisition* (pp. 187–214). Mahwah, NJ: Erlbaum.

Shadle, C. H. (1986). The acoustics of fricative consonants. *Journal of the Acoustical Society of America, 79*(2), 574. https://doi.org/10.1121/1.393552

Shadle, C. H., & Mair, S. J. (1996). Quantifying spectral characteristics of fricatives. In *Proceeding of Fourth International Conference on Spoken Language Processing* (Vol. 3, pp. 1521–1524). Philadelphia, USA: IEEE. https://doi.org/10.1109/ICSLP.1996.607906

Shadle, C. H., & Scully, C. (1995). An articulatory-acoustic-aerodynamic analysisof [s] in VCV sequences. *Journal of Phonetics, 23*(1–2), 53–66. https://doi.org/10.1016/S0095-4470(95)80032-8

Shahamiri, S. R. (2021). Speech vision: An end-to-end deep learning-based dysarthric automatic speech recognition system. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 29*, 852-861. https://doi.org/10.1109/TNSRE.2021.3076778

Shi, R., Gick, B., Kanwischer, D., & Wilson, I. (2005). Frequency and Category Factors in the Reduction and Assimilation of Function Words: EPG and Acoustic Measures. *Journal of Psycholinguistic Research, 34*(4), 341–364. https://doi.org/10.1007/s10936-005-6138-4

Silva, V. De, Iivonen, A., Bondarko, L. V., & Pols, L. C. W. (2003). Common and Language Dependent Phonetic Differences between Read and Spontaneous Speech in Russian, Finnish and Dutch. In M. J. Solé, D. Recasens, and J. Romero (Eds.), *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 2761-2764). Barcelona, Spain. ISBN 1-876346-48-5.

Simpson, A. P. (2009). Phonetic differences between male and female speech. *Language and linguistics compass, 3*(2), 621-640.

Singmann, H. (2019). afex: Analysis of factorial experiments [Computer program]. Retrieved from https://github.com/singmann/afex/

Smorenburg, L., & Heeren, W. (2020). The distribution of speaker information in Dutch fricatives /s/ and /x/ from telephone dialogues. *Journal of the Acoustical Society of America, 147*(2), 949–960. https://doi.org/10.1121/10.0000674

Smorenburg, L., & Heeren, W. (2021a). Acoustic and speaker variation in Dutch /n/ and /m/ as a function of phonetic context and syllabic position. *Journal of the Acoustical Society of America, 150*(2), 979–989. https://doi.org/10.1121/10.0005845

Smorenburg, L., & Heeren, W. (2021b, August 22-25). Effects of speech channel on acoustic measurements and speaker discrimination from /s/ [Conference presentation]. T*he 29th conference of IAFPA*, Marburg, Germany. Retrieved from https://scholarlypublications.universiteitleiden.nl/handle/1887/3210581

Smorenburg, L., & Heeren, W. (2021c, February 4-5). Forensic value of acoustic-phonetic features from Standard Dutch nasals and fricatives [Conference presentation]. *XVII National congress of the Italian Association of Speech Science (AISV)*, Zürich, Switzerland. Retrieved from https://www.cl.uzh.ch/dam/jcr:e6a8ac08-5b5b-4bba-9146-ac6d1256d72c/AISV2021_book%20of%20abstracts_v2.pdf.

Solé, M.-J. (2003). Aerodynamic characteristics of onset and coda fricatives. In M. J. Solé, D. Recasens, and J. Romero (Eds.), *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 2761-2764). Barcelona, Spain. ISBN 1-876346-48-5 retrieved from https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2003/p15_2761.html

Soli, S. D. (1981). Second formants in fricatives: Acoustic consequences of fricative-vowel coarticulation. *Journal of the Acoustical Society of America, 70*(4), 976–984. https://doi.org/10.1121/1.387032

Stevens, K. N. (2000). *Acoustic phonetics* (Vol. 30). London, UK: MIT press.

Stuart-Smith, J. (2007). Empirical evidence for gendered speech production: /s/ in Glaswegian. In J. Cole & J. I. Hualde (eds.) *Laboratory Phonology 9* (pp. 65–86). New York, USA: Mouton de Gruyter.

Su, L., Li, K. -P., & Fu, K. S. (1974). Identification of speakers by use of nasal coarticulation. *Journal of the Acoustical Society of America, 56*(6), 1876–1883. https://doi.org/10.1121/1.1903526

Sumner, M., Kim, S. K., King, E., & McGowan, K. B. (2014). The socially weighted encoding of spoken words: A dual-route approach to speech perception. *Frontiers in Psychology, 4*, 1015. https://doi.org/10.3389/fpsyg.2013.01015

Tabain, M., Butcher, A., Breen, G., & Beare, R. (2016). An acoustic study of nasal consonants in three Central Australian languages. *Journal of the Acoustical Society of America, 139*(2), 890–903. https://doi.org/10.1121/1.4941659

Toda, M., Maeda, S., & Honda, K. (2010). Formant-cavity affiliation in sibilant fricatives. In S. Fuchs, M. Toda, & M. Zygis (Eds.) *Turbulent Sounds - an Interdisciplinary Guide* (pp. 343–374). Mouton de Gruyter. https://doi.org/10.1515/9783110226584

Tracy, E. C., Bainter, S. A., & Satariano, N. P. (2015). Judgments of self-identified gay and heterosexual male speakers: Which phonemes are most salient in determining sexual orientation? *Journal of Phonetics, 52*, 13–25. https://doi.org/10.1016/J.WOCN.2015.04.001

Van Bael, C., Van den Heuvel, H., & Strik, H. (2004). Investigating Speech Style Specific Pronunciation Variation in Large Spoken Language Corpora Large Spoken Language Corpora. In Proceedings of Interspeech (ICSLP), Jeju, Korea. Retrieved from http://www.isca-speech.org/archive/interspeech_2004/i04_2793.html

Van Bergem, D. R. (1993). Acoustic vowel reduction as a function of sentence accent, word stress, and word class. Speech Communication, 12(1), 1–23. https://doi.org/10.1016/0167-6393(93)90015-D

Van Bergem, D. (1995). Acoustic and lexical vowel reduction [Doctoral dissertation]. Universiteit van Amsterdam.

Van Berkum, J. J. A., Van Den Brink, D., Tesink, C. M. J. Y., Kos, M., & Hagoort, P. (2008). The neural integration of speaker and message. Journal of Cognitive Neuroscience, 20(4), 580–591. https://doi.org/10.1162/jocn.2008.20054

Van der Harst, S., & Van de Velde, H. (2006). 17 G's in het Standaardnederlands? *Taal en Tongval, 59*, 172–195.

Van den Heuvel, H. (1996). *Speaker variability in acoustic properties of Dutch phoneme realisations* [Doctoral dissertation]. Nijmegen, Netherlands: Radboud Universiteit.

Van Leeuwen, D. (2011). SREtools: Compute performance measures for speaker recognition [Software program]. Retrieved from https://github.com/davidavdav/sretools.R/

Van Oostendorp, M. (2001). Nasal consonants in variants of Dutch and some related systems. *Neerlandistiek*. Retrieved from https://dspace.library.uu.nl/handle/1874/28504

Van de Pol, W. (2006). *Onder de tap, Afluisteren in Nederland*. Amsterdam: Balans.

Van Son, R. J. J. H., & Van Santen, J. P. H. (2005). Duration and spectral balance of intervocalic consonants: A case for efficient communication. *Speech Communication, 47*(1–2), 100–123. https://doi.org/10.1016/j.specom.2005.06.005

Van de Velde, H., & Van Hout, R. (2000). N-deletion in reading style. *Linguistics in the Netherlands, 17*(1), 209-219.

Van de Velde, H., & van Hout, R. (2001). Sprekertypologie met betrekking tot de realisering van de slot-n in het Standaard-Nederlands. *Taal en tongval*, *14*, 89-112.

Van der Vloed, D., Kelly, F., & Alexander, A. (2020). Exploring the Effects of Device Variability on Forensic Speaker Comparison Using VOCALISE and NFI-FRIDA, A Forensically Realistic Database. In *Proceedings of the Odyssey Speaker and Language Recognition Workshop* (pp. 402–407). Tokyo, Japan: ISCA. https://doi.org/10.21437/odyssey.2020-57

Venables, W. N. & Ripley, B. D. (2002). *Modern Applied Statistics with S* (4th ed.). New York, USA: Springer

Viszlay, P., Juhár, J., & Pleva, M. (2012). Alternative phonetic class definition in linear discriminant analysis of speech. In *19th International Conference on Systems, Signals and Image Processing* (pp. 637–640). Vienna, Austria: IEEE.

Voeten, C. (2020). buildmer: Stepwise Elimination and Term Reordering for Mixed-Effects Regression. R package version 1.5 [Computer program]. https://CRAN.R-project.org/package=buildmer

Voeten, C. C., Heeringa, W., & Van de Velde, H. (2022). Normalization of nonlinearly time-dynamic vowels. *Journal of the Acoustical Society of America, 152*(5), 2692-2710. https://doi.org/10.1121/10.0015025

Wang, B. X., Hughes, V., & Foulkes, P. (2021, February 4-5). System performance and speaker individuality in LR-based forensic voice comparison [Conference presentation]. *XVII National congress of the Italian Association of Speech Science (AISV)*, Zürich, Switzerland.

Retrieved from https://www.cl.uzh.ch/dam/jcr:e1037cb2-8839-4b8a-89f3-a2cc428cf2ff/AISV_2021_Wang%20Hughes%20Foulkes.pdf

Weatherholtz, K., & Jaeger, T. F. (2016). Speech Perception and Generalization Across Talkers and Accents. *Oxford Research Encyclopedia of Linguistics*. https://doi.org/10.1093/acrefore/9780199384655.013.95.

Weirich, M. (2015). Organic sources of inter-speaker variability in articulation: Insights from twin studies and male and female speech. In S. Fuchs, D. Pape, C. Petrone, & P. Perrier (Eds.) *Individual differences in speech production and perception* (Vol. 3, pp. 189—222). Peter Lang International Academic Publishers.

Yun, G. (2006). The effects of lexical frequency and stress on coarticulation. In *Annual Meeting of the Berkeley Linguistics Society* (Vol. 32, No. 1, pp. 441—452).

Samenvatting in het Nederlands

Dit proefschrift onderzocht hoe de spreker-specificiteit van medeklinkers afhankelijk is van linguïstische factoren, met name de directe fonetische context van segmenten en de syllabische positie. De nadruk werd gelegd op nasale en fricatieven medeklinkers, waarvan eerder is vastgesteld dat ze relatief sprekerspecifiek zijn. In de volgende paragrafen worden de hoofdstukken kort samengevat, waarna de theoretische en praktische implicaties worden besproken. Ten slotte worden enkele suggesties voor toekomstig werk gedaan op basis van de bevindingen en de beperkingen van het huidige werk.

Hoofdstuk 2

In dit hoofdstuk zijn twee taalkundige effecten op de akoestiek en sprekerspecificiteit van Nederlandse fricatieven onderzocht. De fricatieven /s/ en /x/ zijn geselecteerd vanwege hun frequentie van voorkomen in het Standaard Nederlands en, in het geval van /s/, omdat uit eerder onderzoek is gebleken dat deze klank relatief sprekerspecifiek is (bv. Kavanagh, 2012; Van den Heuvel, 1996). Deze fricatieven werden

gesampled uit spontane telefoongesprekken in het Corpus Spoken Dutch (Oostdijk, 2000) en werden onderzocht op hun variatie binnen en tussen de spreker als functie van twee taalkundige factoren: fonetische context en syllabische positie. Significante effecten van deze factoren werden gevonden op de akoestiek, voornamelijk voor /x/. Voor syllabische positie vertoonde de akoestiek coda-reductie. Voor fonetische context vertoonde de akoestiek effecten van labialisatie, wat overeenkomt met eerdere literatuur die aantoont dat labialisatie het spectrale gemiddelde in fricatieven spectra verlaagt (bv. Bell-Berti & Harris, 1979; Koenig et al., 2013). Met behulp van multinomiale logistische regressieanalyse in een volgende sprekerclassificatie-test, vertoonden coda's een iets betere sprekerclassificatie dan onsets en fricatieven met labiale buren vertoonden een iets betere sprekerclassificatie dan fricatieven in andere fonetische contexten. Dit werd toegeschreven aan variatie tussen de sprekers in de mate van reductie en co-articulatie; het lijkt erop dat sprekers individuele manieren hebben waarop coda's worden gereduceerd en waarin fricatieven in labiale contexten worden gecoördineerd met betrekking tot de specifieke timing en mate van articulatiegebaren.

Hoofdstuk 3

Het onderzoek gepresenteerd in hoofdstuk 2 is in hoofdstuk 3 uitgebreid met twee Nederlandse nasale medeklinkers. De nasalen /n/ en /m/ zijn gesampled uit dezelfde spontane telefoongesprekken uit het Corpus Gesproken Nederlands (Oostdijk, 2000) dat in hoofdstuk 2 is gebruikt. Wederom werden de effecten van syllabische positie en fonetische context op de akoestiek en variatie binnen en tussen de spreker onderzocht. Terwijl fricatieven vaak worden beïnvloed door contextuele labialisatie, kunnen nasalen effecten vertonen van voor-tot-midden versus achterin-gearticuleerde context, met lagere (tweede) nasale formantwaarden wanneer de nasale een achterin-gearticuleerde buur heeft. Voor fonetische context werd daarom een onderscheid gemaakt tussen achter- en niet-achterburen (in tegenstelling tot het labiale versus niet-labiale onderscheid voor fricatieven).

De resultaten toonden interacties tussen syllabische positie en fonetische context in zowel de akoestiek als de sprekerclassificatie. Voor bilabiale /m/ treedt hoge mate van plaatsco-articulatie meestal anticiperend op in de beginpositie, terwijl voor alveolaire /n/ er meestal sprake is van overdracht van plaatsco-articulatie in codapositie. Co-articulatie lijkt dus vooral plaats te vinden binnen het lettergreepdomein, maar in tegengestelde richtingen voor de twee nasale medeklinkers. Dit zou mogelijk verband kunnen houden met de frequentie van voorkomen van deze segmenten in onset versus coda-positie, aangezien in deze Nederlandse data /n/ vaker voorkwam in coda-positie dan /m/. De relatieve gemarkeerdheid van /m/ in codapositie zou dus kunnen hebben geleid tot weerstand tegen co-articulatie (zie paragraaf 6.2.1. voor meer discussie over dit onderwerp).

Daaropvolgende sprekerclassificaties met behulp van multinomiale logistische regressie toonden aan dat /m/ onsets, die een grotere mate van co-articulatie vertoonden, een betere sprekerclassificatie vertoonden dan /m/ coda's. In overeenstemming met de akoestiek was het voor alveolaire /n/ het omgekeerde; /n/ coda's, die een grotere mate van co-articulatie vertoonden, vertoonden een betere sprekerclassificatie dan /n/ onsets. Er werd geconcludeerd dat tokens met een hoge co-articulatie meer sprekerinformatie bevatten vanwege de variatie tussen de spreker in de timing en mate van co-articulatie.

Hoofdstuk 4

In hoofdstuk 4 is een resterende vraag uit hoofdstuk 2 behandeld. In hoofdstuk 2 werd aangenomen dat het ontbreken van akoestische effecten van linguïstische factoren voor /s/ te wijten was aan de telefoonfilter, dat de spectrale energie voor deze fricatief afsnijdt. Deze aanname is getoetst met behulp van een Engels spraakcorpus met afgetapte telefoongesprekken die gelijktijdig zijn opgenomen met een microfoon die voor de spreker is geplaatst. Het gebruik van een extra taal zou bovendien laten zien of eerdere resultaten zich generaliseren tot een andere, zij het vergelijkbare, taal.

De resultaten toonden aan dat de Engelse fricatief /s/ de verwachte effecten van coda-reductie en co-articulatorische labialisatie op de akoestiek vertoonde, gemeten in de microfoonopname. Hoewel de literatuur tot nu toe vooral gericht was op anticiperende labialisatie, was de mate van overgedragen labialisatie (van de linkerbuur) groter dan anticiperende co-articulatie (voor de rechterbuur). Deze bevinding komt overeen met de hypothese dat Engelse co-articulatie grotere effecten van een vorig segment op een volgend segment vertoont (Hoole et al., 1993). Dit staat in contrast met de resultaten over Nederlandse fricatieven in hoofdstuk 2, die een grotere anticiperende labialisatie lieten zien voor Nederlands /x/, wat aangeeft dat Nederlands en Engels mogelijk verschillende patronen voor labialisatie hebben. De resultaten toonden bovendien aan dat taalkundige effecten niet konden worden waargenomen in de akoestiek van de telefoonopname (300 - 3.400 Hz voor de vaste lijn). Hoewel er enkele significante taalkundige effecten werden gevonden, waren ze niet vergelijkbaar met de effecten die werden gevonden in de studio-opname in termen van grootte en richting en kon er geen duidelijk patroon worden onderscheiden. Dit suggereert dat de telefoonfilter onvoorspelbare effecten kan hebben op de akoestiek. De sprekerclassificaties vertoonden enkele talige context-effecten in de breedband studio-opnamen, maar niet in de telefoonopnamen. Dit betekent dat taalkundige effecten potentieel relevant kunnen zijn bij breedbandsignalen, maar minder bij telefoonsignalen, althans voor segmenten met hoogfrequente spectrale energie zoals /s/.

Hoofdstuk 5

In hoofdstuk 5 werden enkele bevindingen uit voorgaande hoofdstukken getest met Bayesiaanse waarschijnlijkheidstermen die doorgaans worden gebruikt in forensische analyses om te zien of het selecteren van tokens uit specifieke taalcontexten van invloed was op de bewijskracht. Aangezien is aangetoond dat deze taalkundige factoren de akoestiek en de sprekervariatie beïnvloeden in hoofdstukken 2 tot en met 4, zou het selecteren van tokens uit specifieke talige contexten moeten resulteren in meer homogene sets tokens. Spraakmateriaal kan echter schaars zijn in

forensisch casuswerk, wat betekent dat steekproeven uit specifieke taalcontexten kunnen leiden tot onvoldoende tokens per spreker. De resultaten in dit hoofdstuk laten zien dat het analyseren van coda's leidt tot iets sterkere bewijskracht dan het analyseren van onsets voor zowel /n/ als /s/. De verschillen tussen de sprekerclassificatie in verschillende taalcontexten waren echter klein, en de resultaten toonden ook aan dat het prioriteren van het aantal tokens de beste resultaten opleverde voor sprekerdiscriminatie. Gezien de kleine verschillen tussen taalcontexten en de vaak schaarse materialen, werd daarom geconcludeerd dat het selecteren van tokens uit specifieke contexten in forensische contexten niet praktisch of raadzaam is.

Curriculum Vitae

Laura Smorenburg was born in 1992 in Tiel, The Netherlands. She finished her secondary education at O.R.S. Lek en Linge in 2010. The following year, she spent travelling in Australia and New Zealand, after which she decided to study English. She obtained her BA in English language and culture from Utrecht University in 2015 with a thesis on the acquisition of sarcastic prosody in Dutch learners of English. Laura continued to develop her interest in linguistics with a research MA in linguistics at Utrecht University. During her MA, she was a research intern at the Joint Centre of Language, Mind and Brain at the Chinese University of Hong Kong where she researched the role of language experience in the perception of lexical tones in infants. In 2017, Laura graduated *cum laude* from her MA, with a dissertation on the role of female pitch and formants in verbal processing. The following year, she was a lecturer in English language and linguistics and Utrecht University. In 2018, Laura started working as a PhD candidate at Leiden University Centre for Linguistics in dr. Willemijn Heeren's NWO-funded VIDI project *The speaker in speech*. Currently, Laura is a lecturer in English language and linguistics and researcher at Utrecht University.