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).

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.