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

Smorenburg, B.J.L.

# 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 ($C_{\text{llr}}$: 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 $C_{\text{llr}}$ 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 $C_{\text{llr}}$ 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.

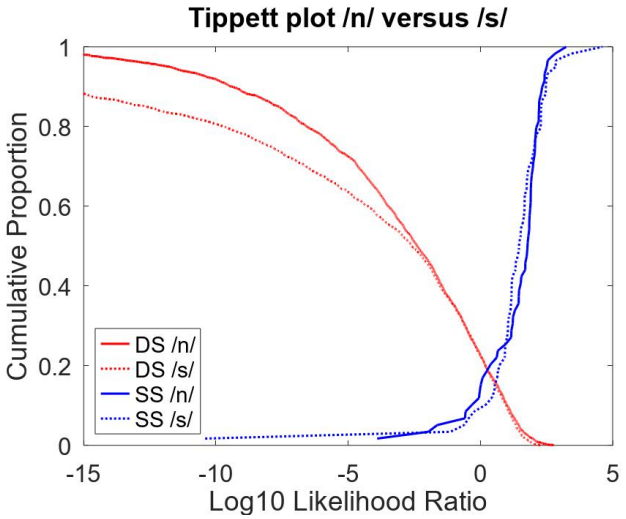


Figure 5.1: *Tippett plot for the LLRs generated using all available /n/ versus /s/ tokens per speaker.*

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.

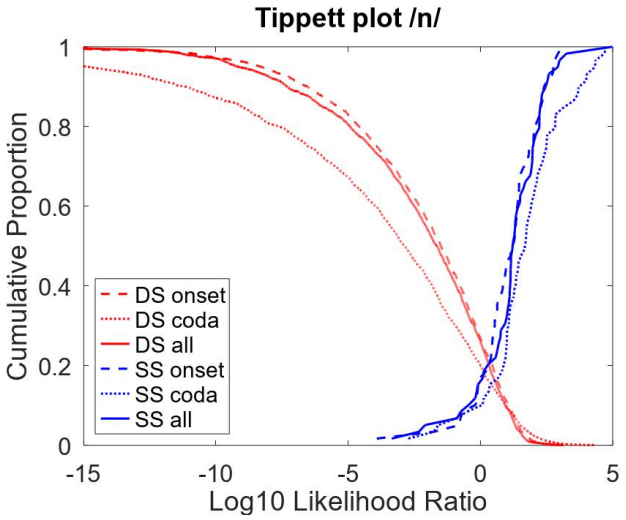


Figure 5.2: *Tippett plot for the LLRs generated for /n/ using tokens sampled across linguistic environments, from onsets, or from codas (sample size per speaker across all conditions $N \leq 16$).*

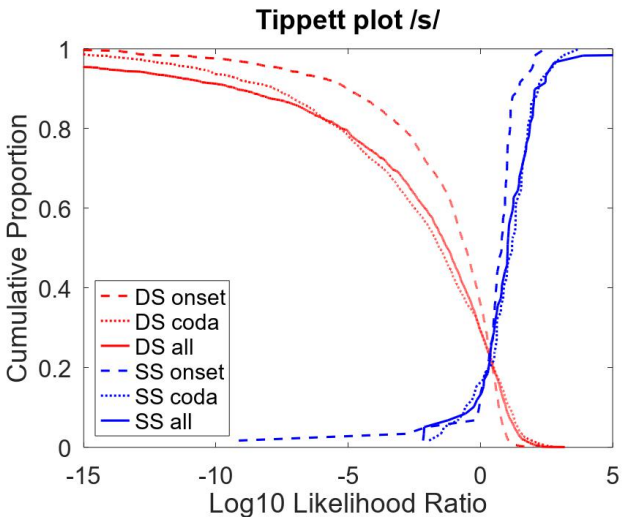


Figure 5.3: *Tippett plot for the LLRs generated for /n/ using tokens sampled across linguistic environments, from onsets, or from codas (sample size per speaker across all conditions N ≤ 16).*

## 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.