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

Smorenburg, B.J.L.

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).

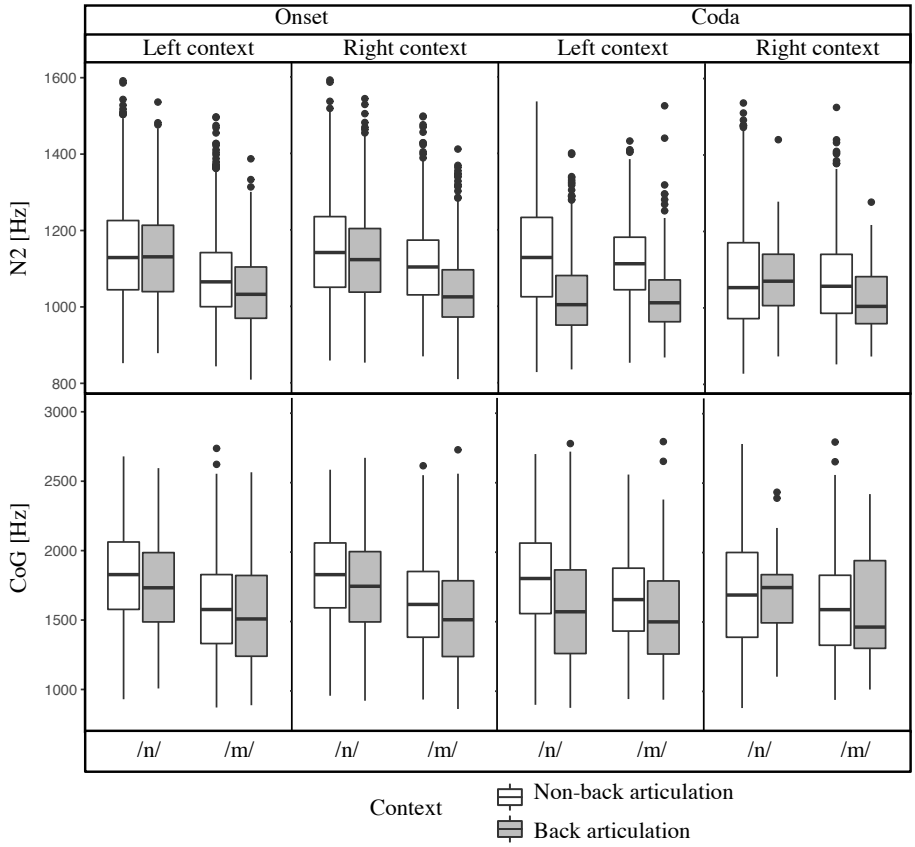

Figure 3.1. *Boxplots for N2 and CoG (Hz) by place of articulation of left and right context, nasal consonant, and syllabic position*

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/.

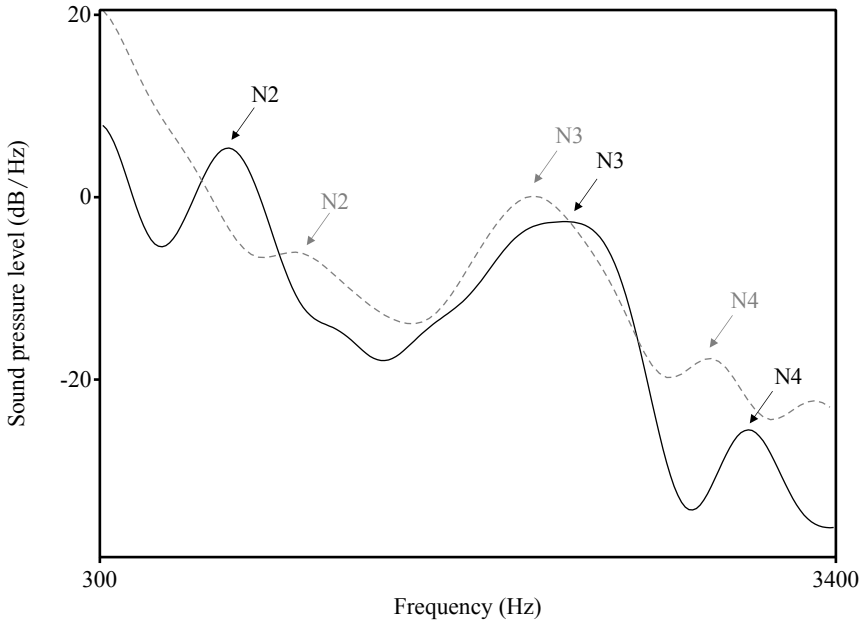

Figure 3.2: *Spectral slices for two /m/ tokens from the same speaker, taken from the mid-50% of each token with cepstral smoothing (500 Hz). Grey dashed line: /m/ in phonetic context with a non-back place of articulation* (was meestal, *'was usually'*, /ʋɑs.mes.tɑl/; *N2 = 1,143 Hz, N3 = 2,126 Hz, N4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

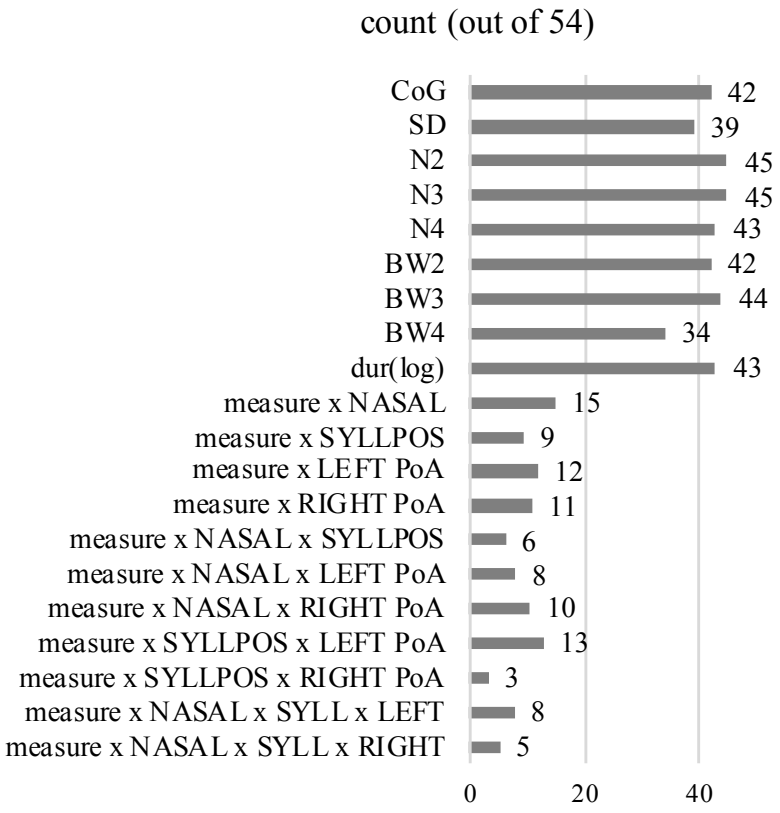

Figure 3.3. *Count of non-zero coefficients for 54 speakers. Counts of interaction predictors were averaged over acoustic measures (N = 9).*

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