Stop! Hey, what's that sound? the representation and realization of Danish stops

Puggaard-Rode, R.

CHAPTER 5

Time-varying spectral characteristics of stop releases

5.1 Introduction

The aspirated alveolar stop /t/ in Standard Danish is usually strongly affricated. This was already pointed out by Otto Jespersen (1897–1899: 355). He maintained that /t/ was best described as an aspirated stop, but assumed that Danish was undergoing a sound change whereby all aspirated stops would eventually become affricates, as had happened in some varieties of German a millennium earlier with the Second Consonant Shift. Jespersen assumed that /t/ was most advanced in this sound change, followed by /k/, and finally /p/. Today, more than a century after Jespersen's observations, the affrication of /t/ is taken for granted in the literature; it has been established several times over, and has been shown to be exceptionless (see Section 2.3.4). While it is cross-linguistically common for the initial burst noise of stops to have a similar frequency range to fricatives at the same place of articulation,

A revised paper corresponding to this chapter has been published (Puggaard-Rode 2022b). Audio data are available online in password-protected form (Grønnum 2016); replication data and code are freely available (Puggaard-Rode 2022a).

this usually makes up a comparatively small portion of stop releases in other languages. Brink and Lund (1975) tracked the development of /t/-affrication across more than a century of recordings of Copenhagen Danish, and showed that it went from a widespread phenomenon in the mid-19th century to an exceptionless phenomenon in the mid-20th century.

As discussed in Section 2.3.4, the prominent affrication in /t/ has led to a variety of different phonetic transcription strategies. In very narrow transcription, it is often assumed that /t/ in simple onset is best represented as /d/ with some 'garnish': [d̥ˢ] (e.g. Basbøll 1968, 2005; Grønnum 1998), [d̥ˢʰ] (e.g. Petersen 1983), and [dˢh] (Brink and Lund 1975) are all used in the literature, under the assumption that the only meaningful difference between /d t/ is the release. More broad transcriptions include [tsʰ] (Basbøll and Wagner 1985), and [tˢ] (e.g. Grønnum 1998), the latter of which has emerged as the standard. More recently, Schachtenhaufen (2022) has proposed that the sound is a true affricate and should be transcribed as [ts].

Fischer-Jørgensen (1972d) shows that having the right noise profile during the release is a crucial cue to the perception of the laryngeal contrast in stops at all places of articulation, which suggests that /t/ is not so special after all. While there is consensus about the affrication in /t/, possible affrication patterns in /p k/ have never been investigated. On the one hand, since /p t k/ show class behavior in other matters (e.g. phonotactics; see Section 2.4.2), we might also expect them to show class behavior in phonetic implementation; on the other hand, Chodroff and Wilson (2018) recently found only moderate signs of class behavior in the realization of place cues in American English /p t k/. The most straightforward explanation for the lack of interest in affrication patterns in Danish /p k/ is that it is not particularly salient (if it is there at all); perhaps this is simply because coronal frication is more salient than labial and dorsal frication. This is a reasonable assumption, which can help account for why most affricates cross-linguistically are coronal (Ladefoged and Maddieson 1996).

One goal of this chapter is to investigate Jespersen's prediction a century later: are Danish aspirated stops changing into affricates? This is not straightforward: the boundary between an aspirated stop and an affricated one is fuzzy, as is boundary between an affricated stop and

a proper affricate. I approach the question by looking holistically and dynamically at time-varying spectral characteristics throughout stop releases, and how they vary, using the DanPASS corpus (see Section 4.5.1). I focus on the following questions, which are more readily answerable than the question of whether or not the sounds in question are affricates:

(1) How do the spectral characteristics of Danish stop releases vary across time?

(2) How are the time-varying characteristics of Danish stop releases affected by different phonetic contexts? An example could be coarticulation effects following from features of the following vowel, like backness, height, and rounding, all of which affect the size and shape of the vocal tract.

When analyzing the dynamics of spectral characteristics, researchers usually resort to using a small number of discrete measurements aimed at capturing as much of the relevant spectral information as possible. For vowels and sonorant consonants, an example is formants; for obstruent consonants, examples are spectral moments or coefficients of discrete cosine transformations of the spectrum. A second goal of this chapter is to demonstrate function-on-scalar regression (FOSR; Reiss et al. 2010; Greven and Scheipl 2017a; Bauer et al. 2018) as a method for taking the entire spectrum into account when analyzing sources of phonetic variance. Rather than relying on discrete measurements, FOSR allows for the use of complete spectra as response variables. FOSR gives a clear and easily interpretable overview of the influence of various factors on time-varying spectral characteristics, and does so with minimal reduction of the information in the acoustic signal. Other recent studies have compared full (temporally static) spectra in order to illuminate differences between palatalized and non-palatalized consonants using smoothing spline ANOVA (Iskarous and Kavitskaya 2018) and generalized additive models (Nance and Kirkham 2020); in Section 6.7, I use functional principal component analysis to analyze the main sources of variance in spectra of stop releases. Functional regression models have been used in the analysis of phonetic data previously (e.g. Pouplier et al. 2014, 2017; Cederbaum et al. 2016; Carignan et al. 2020;

Volkmann et al. 2021). However, to the extent of my knowledge, this is the first study to use FOSR to analyze speech spectra.[1]

Section 5.2 of this chapter discusses the acoustic characteristics of aspirated stops and affricates, and the available heuristics (or lack thereof) for determining whether a sound is phonetically one or the other. Section 5.3 discusses available methods for measuring frication and some of the problems associated with these, and Section 5.4 presents FOSR and other smoothing-based approaches to dynamic data analysis as possible solutions to these problems. Section 5.5 presents the methods used in this study in detail, and Section 5.6 shows the results. In Section 5.7, I discuss the hypotheses presented above on the basis of the results, and discuss opportunities and limitations of FOSR as used here. Section 5.8 briefly concludes the chapter.

5.2 Aspirated stops, affricates, and the middle ground

The production of both stop consonants and affricates has been modeled thoroughly in the work of Fant (1960) and Stevens (e.g. 1993a, 1993b, 1998: chs. 7–8). A shared component of both types of sound is a complete occlusion somewhere in the oral cavity, which allows intraoral air pressure to build up. Another shared component is a release phase, in which this pressure is released, resulting in a rapid sequence of acoustic events, including an initial brief transient followed by frication. The transient shows a fairly even distribution of noise throughout the spectrum. Frication noise is subsequently generated at or near the point of occlusion; due to the high pressure behind the constriction and the narrow gap in the oral cavity, the escaping air becomes turbulent and excites the area around the constriction. The nature of this noise gradually changes as the approximation gradually widens. In aspirated stops, air will continue to escape through the open glottis for some time after the release, and turbulence

[1] Wood (2017a: 390ff.) proposes similar models for the analysis of other types of spectra (infrared spectra and protein mass spectra), but in both cases, the spectra are independent variables. In the studies reported here, spectra are the dependent variables.

noise generated at the area around the vocal folds continually excites the vocal tract.

The energy distribution of the turbulent frication noise depends on the nature of the obstruction (Shadle 1991). In labials, since there is no cavity in front of the obstruction, the frication noise is generated directly at the lips, causing a fairly even distribution of noise throughout the spectrum, with a slight linear drop in amplitude at increasing frequencies. In alveolars, the turbulent air stream impinges on the teeth immediately in front of the constriction, meaning there is only a very small cavity anterior to the constriction, causing high resonance frequencies around 5 kHz to be excited. In velars, the turbulent air stream impinges on the hard palate at an oblique angle, before being filtered through a sizeable front cavity, causing relatively low resonance frequencies somewhat below 2 kHz; note, however, that the exact point of occlusion in velars is variable and depends on surrounding vowel(s), since the tongue body is less precisely controlled than the tip and blade (Ouni 2014), and the tongue body is itself more directly involved in the production of vowels than the tip and blade. A more fronted obstruction will cause the air stream to more directly impinge on the hard palate, causing higher resonance frequencies.

During aspiration, low-frequency noise is generated as the airstream passing through the glottis impinges on the vocal folds, epiglottis, and surfaces directly above the glottis; this turbulence noise further excites the natural resonances of the oral cavity, which of course largely depend on e.g. the position of the tongue. The aspiration noise is present throughout the release, but is initially dominated by frication. As the obstruction above the glottis opens, aspiration noise will gradually overtake frication noise in prominence (Hanson and Stevens 2003).

In voiceless unaspirated stops, the frication phase is very brief, but it is an important cue to place of articulation. There are two primary place cues in stops: the spectral characteristics of the initial frication phase (e.g. Stevens 1971; Stevens and Blumstein 1978; Blumstein and Stevens 1979, 1980), and the transitions of formants as the articulators move from occlusion to vowel (Kewley-Port 1982, 1983; Kewley-Port et al. 1983; Stevens et al. 1999). In aspirated stops, formant transitions are relatively weak, because movement of the articulators typically

happens before the onset of voicing. This makes frication as a place cue all the more important in aspirated stops. Frication is also usually a stronger cue in aspirated stops: since the glottis is spread during at least part of the closure, there is a greater build-up of supraglottal air pressure, causing quicker releases and greater burst intensities than in unaspirated stops (see e.g. Löfqvist 1975a, 1980; Jaeger 1983). Long voicing lag can in itself lead to affrication in certain environments: when devoiced, high front vowels can be acoustically similar to fricatives (Mortensen 2012). This can lead to the common sound change whereby /k/ → /tʃ/ before /i/ (Hock 1991; Ohala 1992), as observed in e.g. Slavic, Indo-Iranian, and Middle Chinese (Guion 1998 and references therein), and the common phonological process where /t/ is realized as an affricate or fricative before /i/, as observed in e.g. Finnish and Korean (Kim 2001; Hall and Hamann 2006; Hall et al. 2006).

The timing of gestures in Danish aspirated stops is different from comparable Germanic languages, as discussed throughout Section 2.3. In Icelandic and Swedish, peak glottal opening is achieved relatively early during the closure of aspirated stops (Pétursson 1976; Löfqvist 1980); in English and German as well, the glottis is typically fully spread sometime before the stop release (Sawashima 1970; Hoole et al. 1984). Furthermore, closures in aspirated stops are typically longer than in unaspirated stops (Lisker 1957; Löfqvist 1976; Stathopoulos and Weismer 1983; Braunschweiler 1997). This ensures that supraglottal air pressure is high at the time of the release. In Danish, however, peak glottal opening is typically just after the stop is released (Frøkjær-Jensen et al. 1971), and closure duration is shortest in aspirated stops (Fischer-Jørgensen 1969, 1972b). Taken together, these two facts about Danish aspirated stops – late peak glottal opening, and relatively short closure duration – mean that there are fewer mechanisms in place to ensure high supraglottal air pressure at the time of release, and accordingly, less guarantee of a prominent burst.[2] This can motivate why a constriction would be retained for relatively long in Danish

[2]This exposition suggests that Danish stops are outliers in Germanic, but in fact, all languages which have been examined in detail have idiosyncrasies in their stop articulation. If anything, it should indicate that oral and glottal gestures are largely independently controlled, and that individual languages have a lot of freedom in how phonetic categories are implemented.

stop releases. Functionally, it can also explain the 'need' for affricated releases in Danish: if the place cues of the burst are not otherwise so prominent, they can be strengthened by retaining a constriction after the release.

There are no clear heuristics to decide whether a particular speech sound is an affricated aspirated stop or an affricate – at least not from the acoustic signal alone. In phonology, a decision may be reached on the basis of behavior. Affricates are often assumed to contain a feature like [stop] as well as one usually used in the representation of fricatives, such as [strident] (e.g. Jakobson et al. 1951) or [continuant] (e.g. Lombardi 1990);[3] see Lin (2011) for an overview of how affricates have been modeled in phonological theory. If an occlusive with a lot of frication behaves like an aspirated stop to all extents and purposes, it should probably be considered an aspirated stop at the phonological level; there will be no need to posit a [continuant] feature. If it patterns with fricatives, or shows other forms of exceptional behavior, those would be grounds for considering it an affricate at the phonological level.

On these grounds, Standard Danish /t/ should certainly be considered an aspirated stop. The phonotactic behavior of /t/ is similar to that of other stops (Vestergaard 1967), and /t/ shows the same patterns of positional allophony as /p k/, with truncated release after /s/ and in weak position (see Chapter 3), and loss of release syllable-finally (although optional release phrase-finally; Grønnum 2005: 49). Furthermore, when loan words with alveolar affricates are nativized and adapted to Danish phonology, the affricate is generally reanalyzed as /s/ rather than /t/, as in the examples in (3);[4] etymologies are from DSL (2018).

[3] In binary feature accounts, affricates are often represented with both [-continuant] and [+continuant] (e.g. Sagey 1986).

[4] A counterexample is *tzatziki*, which is nativized as [tʰætˈsiki] (DSL 2018); here, the first /ts/ is reanalyzed as /t/, and the second as ambisyllabic /t.s/.

(3)	[sɑːˀ]	*tsar*	'czar'	from Russain [tsarʲ]
	[suˈkʰiːni]	*zucchini*	'zucchini'	from Italian [tsukˈkino]
	[sɛn]	*zen*	'zen'	from Japanese [dzen]
	['syɐ̯ek]	*Zürich*	'Zurich'	from German ['tsyːʁɪç]
	[suˈnɑːmi]	*tsunami*	'tsunami'	from Japanese [tsɨnami]

In a study of Danish speakers' productive acquisition of Standard Chinese coronal obstruents (Puggaard 2020c), it was further shown that the most common error in the production of (non-aspirated) /ts/ is realizing it with no closure phase, i.e. similar or identical to /s/. Native speakers of Danish do not map Standard Chinese /ts/ to their native /t/ phoneme. They do, however, tend to map Standard Chinese /tsʰ/ to their native /t/ phoneme, further cementing that both affrication and aspiration are crucial cues to Danish /t/.

From a phonetic perspective, Stevens (1993a) defines affricates as sounds which have two separate constrictions formed by the primary articulator. The anterior constriction forms a complete closure, while the posterior one forms a close approximation. In affricates, frication noise is generated at this posterior constriction, while in stops, frication noise is generated directly at the point of occlusion. This distinction is difficult to extend to acoustics or to gauge impressionistically. In practice, most decisions about stop–affricate category membership is likely based on intuition; a sound is categorized as an affricate if frication lasts for more than a certain proportion of the release. It is therefore not a goal of this chapter to determine whether /p t k/ are phonetic affricates in Danish; such a decision can only be made with targeted articulatory studies comparing Danish with other languages with clear-cut stop–affricate distinctions. This is rather an exploratory study aimed at better understanding the distribution of spectral properties in Danish stop releases.

5.3 Measuring frication

It has long been established that frication at different places of articulation (whether in fricatives, stop releases, or otherwise) has distinct spectral properties (see Kopp and Green 1946). A classic method for

differentiating places of articulation in frication is locating peaks and valleys in spectral energy distribution, essentially by 'eyeballing' spectrograms (e.g. Hughes and Halle 1956; Strevens 1960).

Forrest et al. (1988) popularized treating the spectrum as a probability mass function, and analyzing it by calculating four moments: 1) the 'mean frequency', also known as center of gravity (COG); 2) standard deviation (SD), 3) skewness, and 4) kurtosis. COG reflects the mean distribution of energy across the spectrum; SD reflects how much the energy deviates from the mean; skewness reflects how much the energy distribution is skewed relative to the mean, and in which direction; kurtosis reflects the peakedness of the energy distribution. Forrest et al. found that spectral moments distinguish fairly well between places of articulation in stop bursts, and that particularly COG, skewness, and kurtosis distinguish fairly well between places of articulation in alveolar and post-alveolar fricatives; Stoel-Gammon et al. (1994), on the other hand, found that SD is particularly stable in determining the difference between dental and alveolar stop bursts. The results of subsequent studies have overall not been particularly stable (see e.g. Shadle and Mair 1996), but COG remains a very popular measure in the analysis of spectral properties of fricatives, often without taking into account other moments; an example is Gordon et al. (2002). This is problematic, since spectra often correspond to functions that are far from normally distributed. The mean value from a non-normal distribution does not give a clear picture of the shape of the distribution, and spectra with quite different shapes may have very similar COG.

A number of other measures have been proposed for analyzing frication, mainly for determining the precise place of articulation in fricatives. Jongman et al. (2000) find that the different places of articulation in English fricatives are distinguished fairly well using the average location of the spectral peak. Koenig et al. (2013) show that the mid-frequency spectral peak, i.e. the frequency with the highest amplitude within a 3–7 kHz band, captures the fairly subtle difference between labialized and non-labialized alveolar fricatives in adolescents.

Another proposed method is using cepstral coefficients derived from a discrete cosine transform of the spectrum (DCT; Watson

and Harrington 1999). DCT reduces the high dimensionality of the spectrum to (typically) four discrete values, corresponding to the amplitude of half-cycle cosine waves derived from the spectrum. DCT0 reflects the mean amplitude of the spectrum; DCT1 reflects the linear slope; DCT2 reflects the curvature; and DCT3 reflects the amplitude at higher frequencies. In a comparison of /ʃ ç/ in different varieties of German, Jannedy and Weirich (2017) show that DCT-based classification more closely approximates the perception of these sounds than classification based on spectral moments, and DCT coefficients have been shown to outperform spectral moments in classification of place of articulation in both voiceless stops (Bunnell et al. 2004) and fricatives (Spinu and Lilley 2016). While DCT coefficients give a fuller picture of spectral shape than spectral moments, they are also more difficult to interpret.

Measurements such as the ones discussed above are often taken at static or normalized points in time, such as the midpoint (or some pre-determined range around the midpoint) of fricatives or affricates (for examples, see e.g. Jongman et al. 2000; Liu and Jongman 2013). Mücke et al. (2014) refer to these points in time as 'magic moments'. Magic moments give us a limited picture of the acoustic nature of sounds; affricates are inherently dynamic, and Reidy (2016a) shows that even sibilant coronal fricatives vary dynamically throughout their time course in language-specific ways. Spectral properties of stop releases are usually measured only at the burst, which in aspirated stops corresponds to a relatively small initial portion of the release (see e.g. Chodroff and Wilson 2014).

Summing up, most approaches to quantifying frication reduce the complex time-varying information in spectra to something more manageable. This is very reasonable, because 1) many popular statistical methods in linguistics cannot handle variables with high dimensionality, and 2) it is a goal in itself to propose the simplest possible model of how language works with the highest possible explanatory value. With regards to 1), statistical models which can take complex dynamic information into account are increasingly being used, as discussed in the next section; this chapter demonstrates how FOSR can be used to model time-varying spectral information with little reduction of dimensionality. With regards to 2), deciding on a model

of language which balances simplicity and explanatory value can simply not be done without first testing very complex models. Studies mentioned above have demonstrated how some patterns can only be uncovered by increasing dimensionality. For example, Reidy (2016a) shows that the language-specific nature of spectral dynamics in fricatives only becomes apparent when measuring spectral properties at several timepoints, and Jannedy and Weirich (2017) show that the spectral differences between [ʃ ç] in German are more readily apparent when using a measure that takes more of the spectrum into account (i.e. using DCT coefficients rather than moments).

5.4 Smoothing approaches to analyzing dynamic data

In the past years, following Baayen's (2008) popularization of mixed-effects regression models in linguistics, the field has seen a rapid increase in the use of sophisticated statistical techniques. A problem with linear models is the analysis of dynamically varying data, in particular data from time series. If some measure varies as a function of time, then a linear model by necessity assumes that the variation follows a straight line. As Sóskuthy (2017) demonstrates for formants, this is a poor assumption: variance as a function of time is often non-linear. A solution to this problem is using smoothed curves. Given a number of data points associated with e.g. a time series, a smoothing function can be used to approximate a continuous curve corresponding to the data's non-linear variation as a function of time (see de Boor 2001; Wood et al. 2016). Smoothing involves reducing the observations to a number of basis functions (or 'knots'), and using a penalizing smoothing parameter to determine the wiggliness of the resulting curve (see Gubian et al. 2015). Combining too many basis functions with a low smoothing penalty will lead to overfitting, resulting in curves that include irrelevant information; conversely, combining too few basis functions with a high smoothing penalty will likely lead to underfitting, resulting in curves that omit relevant information.

Generalized additive (mixed) models (GAMMs) have rapidly become very popular in linguistics (see e.g. Wood 2017a; Wieling 2018;

van Rij et al. 2020a; Sóskuthy 2021). These are similar to linear mixed-effects models, but allow for the inclusion of smooth effects. They are typically used for time series analysis, but have also been used to analyze the dynamics of e.g. EEG registration (Baayen et al. 2018; Voeten 2020: ch. 5), geo-linguistic variation (e.g. Wieling et al. 2011, 2014; see also Chapter 6), and speech spectra (Nance and Kirkham 2020).

Functional data analysis (FDA; Ramsay and Silverman 2005; Ramsay et al. 2009; Gubian et al. 2015; Pouplier et al. 2017) has overall been less influential in linguistics. FDA refers to a family of statistical methods which extend existing methods to account for functional data. In practice, this means that smoothed curves can be used as input variables in statistical models, in addition to discrete values. An example of this is the functional extension of principal component analysis (FPCA), which can be used to determine the primary sources of variance in curves. For example, Gubian et al. (2015) use FPCA to jointly analyze how *F*1 and *F*2 pattern in the realization of diphthongs and hiatuses in Spanish, respectively. I return to FPCA in Section 6.7, where I use the method to determine the primary modes of variation in noisy spectra.

Functional regression models are suitable when one or more of the analyzed variables are of a functional nature. If an independent variable is functional and the response variable is constant over the functional domain, this can be modeled with scalar-on-function regression; if the response variable is functional and all independent variables are constant over the functional domain, this is suitably modeled with function-on-scalar regression (Bauer et al. 2018). There are several approaches to modeling function-on-scalar data (an overview is given in Greven and Scheipl 2017b: 110ff.). Here, I will focus on the implementation presented by Scheipl et al. (2015, 2016), Greven and Scheipl (2017a) and Bauer et al. (2018). For the mathematically inclined, the model can be summarized as in (4), from Bauer et al. (2018: 353).

$$g(\mathbb{E}(Y_i(t)|\chi_i, E_i(t))) = \beta_0(t) + \sum_{r=1}^{R} f_r(\chi_{ri}, t) + E_i(t) \tag{4}$$

$g(\cdot)$ is a pre-specified link function mapping the predictor to the functional domain. The expected value $\mathbb{E}(\cdot)$ of each observation *i = 1, ..., n* of the response variable *Y* as a function of *t* conditional on a set of covariates χ and residual functional error $E(t)$ corresponds to a functional intercept $\beta_0(t)$, as well as *R* covariate effects $f_r(\cdot)$, each of which form a subset χ_r of the full covariate set and may vary over the functional domain *t*, and the residual functional error $E(t)$.

Functional regression models and GAMMs are conceptually very similar. GAMMs are often fitted using the R package `mgcv` (Wood 2017a, 2021), which allows for significant flexibility in the selection of spline bases (Wood 2017a: ch. 5), residual error distributions (Wood et al. 2016), and smoothing parameter estimation methods (Wood 2011; Wood et al. 2015), and handling of autocorrelated residuals (Baayen et al. 2018), and can handle very large data sets (Wood et al. 2017). Wood (2017a: 290ff.) gives a number of examples of how functional regression models can be implemented in `mgcv`. Perhaps for this reason, the framework for functional regression modeling I adhere to here is sometimes referred to as (generalized) functional additive mixed modeling (Scheipl et al. 2015, 2016). A disadvantage of GAMMs is that they cannot take functional response variables. If I wanted to model spectral variance with GAMMs, I would have to use an amplitude measure as the response variable, and model the variation in amplitude across the time and frequency domains. This is conceptually not very satisfactory: the spectral shape is our variable of interest, not the individual amplitude levels.

Functional additive regression models are implemented in the `pffr` function of the R package `refund` (Goldsmith et al. 2021). This function uses the `mgcv` computation engine, and inherits the same flexibility as GAMMs fitted with `mgcv`, but allows for dependent and independent variables to be functional. The syntax is also similar to `mgcv`, except there are several more term constructors for including various kinds of variables; most of these are not discussed here. A fully reproducible example of the model fitting and selection procedure using `pffr` is given in Puggaard-Rode (2022a), where I also decompose the code.

Functional regression models are usually high-dimensional and the number of underlying data points is often very high. This can make traditional significance tests unreliable, as these are highly affected

by sample size (see e.g. Kühberger et al. 2015 and references therein). Wood (2013) proposes an *F*-test for calculating significance of non-linear variables in GAMMs, and the results of this test are also reported in the output of `pffr`; however, researchers should exercise caution in interpreting these results, as even tiny effects will appear highly significant if the sample size is sufficiently large. This is also the case for likelihood ratio tests of nested models. For this reason, I do not report *p*-values in this chapter. This issue is not specific to FOSR models; if the same data was fitted to a GAMM, these concerns would still hold.

In any case, *p*-values and associated measures of non-linear effects can only tell us if there *is* an effect, they cannot tell us much about the nature of that effect. A more suitable way to explore non-linear effects in exploratory studies is to visualize them. If the goal is hypothesis testing, Bauer et al. (2018) propose several different solutions. Marra and Wood (2012) propose a method for calculating confidence intervals of non-linear effects; this method can be used to quantify and visualize the uncertainty associated with non-linear fitted effects along the functional grid. Bauer et al. (2018) propose a more precise bootstrap-based method for calculating confidence intervals, but this precision comes with a significant computational cost.

5.5 Methods and materials

5.5.1 Acoustic analysis

As in the study of intervocalic voicing presented in Chapter 4, this study relies on the monologues from the DanPASS corpus (Grønnum 2009; see Section 4.5.1). The initial acoustic analysis was done in Praat (Boersma 2001; Boersma and Weenink 2021). An automated script was used to locate all aspirated stops (i.e. members of /p t k/) in simple onset position in the DanPASS monologues, and combine them into a single sound file with a subset of the original annotations. The script is a modified version of the one used in Chapter 4 written by Dirk Jan Vet (see Puggaard-Rode et al. 2022b). This located a total of 2,539 stops. The release phases of the stops were segmented primarily on the basis of the waveform, with the burst used to demarcate the beginning of the release and the first signs of periodicity used to demarcate the end

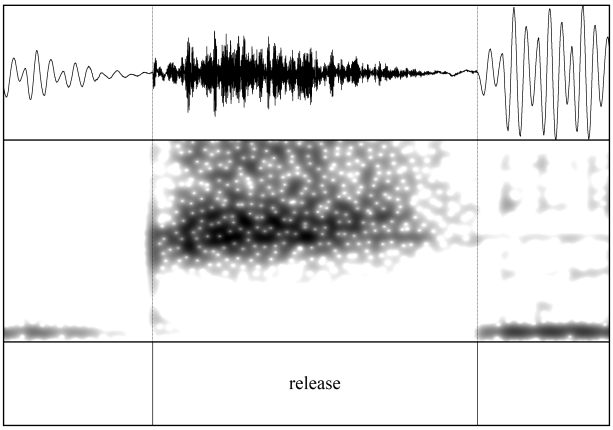

Figure 5.1: *Example of a segmented* /t/ *release.*

of the release (following Francis et al. 2003). If multiple bursts were present, the final one was chosen (following Cho and Ladefoged 1999). This process was partially automated with a Praat script searching for sudden increases in amplitude, but the results required extensive manual correction. An example of a segmented /t/ release is shown in Figure 5.1. 205 tokens were excluded during this process if there was no discernible closure phase. The distribution of stops by phoneme is shown in Table 5.1, along with the mean duration of the release for stressed and unstressed tokens, which is equivalent to positive voice onset time (VOT).

In some cases, the mean VOT values differ quite dramatically from those reported by Mortensen and Tøndering (2013) on the basis of DanPASS (see Section 2.3.1). This is likely because Mortensen and Tøndering follow Fischer-Jørgensen and Hutters (1981) in considering the onset of higher formants to be the relevant landmark for measuring VOT rather than the first signs of periodicity; this strategy leads to higher overall values, particularly for /k/. The VOT measurements are discussed in more detail, and compared to similar measurements from traditional regional varieties of Danish, in Section 6.5.2.1.

A Praat script was subsequently used to extract the release duration and information about the phonetic context for each token. The

Table 5.1: *Number of aspirated stops included in the study, along with mean VOT values. First and third quantiles are given in parentheses.*

Phoneme	Number	Mean duration (stressed), ms	Mean duration (unstressed), ms
/p/	642	57 (42–70)	41 (27–50)
/t/	850	79 (60–92)	68 (53–79)
/k/	842	59 (43–69)	46 (36–54)

phonetic context is coded four binary variables concerning vowel HEIGHT, BACKNESS, and ROUNDING, as well as STRESS. For this purpose, as motivated in Section 4.4.1.1, [i y u ɪ ʏ ʊ e ø o] are all defined as high vowels. [u ʊ o ʌ ɔ ɑ ɒ] are the relevant back vowels, and [y u ʊ ø o œ ɔ ɶ ɒ] are the relevant rounded vowels.

Each release was divided into 20 equally long time steps. This is too coarse-grained to tease apart very dynamic sequences, such as the segue from initial transient to frication, but should be fine-grained enough to capture gross changes in affrication and aspiration. The recordings were filtered to include a frequency range between 500–12,000 Hz. Frequencies below 500 Hz were removed to avoid a potential influence of intrusive voicing or low frequency ambient noise. Frequencies above 12 kHz were removed because they rarely play a role in speech. In fact, 12 kHz is a relatively high cut-off point compared to other comparable studies; this is motivated a study on sociolinguistic variation in Danish /t/ which showed that mean COG for fronted realizations of /t/ can go above 6 kHz, suggesting that very high frequencies may occasionally play a role in /t/ releases (Pharao and Maegaard 2017). For each time step, the four first spectral moments were also extracted; the spectral moments are not used in the analysis, but are available alongside the other data used for the analysis (Puggaard-Rode 2022a).

Multitaper spectra for each time step were generated in R (R Core Team 2021; RStudio Team 2022).[5] Compared to spectra computed using fast Fourier transformation (FFT), such as those computed in Praat, multitaper spectra provide a lower variance spectral estimate which make them suitable for spectra that are noisy and highly dynamic (Blacklock 2004; Reidy 2015). 3 tokens of /k/ were excluded because the total duration of their release was below 10 ms, and the algorithm used to generate the spectra does not work for sound files shorter than 0.5 ms. The dependent variables in the statistical models are the multitaper spectra; each of these consists of a vector of amplitude values along the frequency domain. The frequency ranges differ in size depending on the duration of the time step; longer time steps result in more fine-grained spectra, and thus smaller frequency ranges. Within each spectrum, the amplitude measurements were standardized,[6] since plenty of non-linguistic factors can lead to deviations in overall amplitude level. Note that the multitaper spectral analysis returns intensity values in watt per square meter (W/m^2) rather than amplitude values in the more common logarithmic decibel (dB) scale. I use the W/m^2 scale for this study, as statistical results proved similar across scales, but visualizations are more readily interpretable when using the W/m^2 scale. Only the frequency range between 500–10,000 Hz was used for the statistical analysis of /t/ spectra, and 500–8,000 Hz for /p k/, since the minor activity above these limits seemed to be essentially random noise, and interfered with the clarity of the results.

[5]This was done using the add-on packages `tuneR` (Ligges 2021) and `multitaper` (Rahim 2014; Rahim and Burr 2020), with convenience functions based on code from Reidy (2013, 2016b).

[6]The amplitude measurements were standardized by subtracting the mean and dividing by two standard deviations, following Gelman and Hill (2006).

5.5.2 Statistical analysis

All statistics were calculated in R (R Core Team 2021; RStudio Team 2022) with a number of add-on packages.[7] Separate FOSR models were fitted for each stop with multitaper spectra as the dependent variables. The spectra are smoothed using P-splines with the number of basis functions for the global intercept set as the mean number of amplitude observations per spectrum (corresponding to 32 for /t/, 19 for /k/, and 17 for /p/). This seems to strike a good balance between signal and noise. For the functional responses, 6 basis functions were used for the /t/ model and 5 for the /k/ and /p/ models, guided by the selection procedure proposed by Pya and Wood (2016). P-splines are useful for data sampled on uneven grids (Wood 2017b). Normalized time is included as a non-linear independent variable, smoothed with thin plate regression splines (Wood 2003) with 16 basis functions to ensure high granularity in the temporal dimension. Smoothing penalization parameters were automatically selected using fast restricted maximum likelihood estimation (fREML; Wood 2011). The residuals for the models are reasonably normally distributed,[8] although for the /p/ model, they are somewhat leptokurtic (kurtosis = 5.45); however, Gaussian models with a high number of observations should be quite robust to violations of normality (e.g. Knief and Forstmeier 2021).

A major advantage of GAMMs is the ability to account for autocorrelated residual error (Baayen et al. 2018; Wieling 2018); for example, measurements taken at adjacent steps in a time series are likely to be correlated simply because they are adjacent, which adds unwanted structure to the model residuals. This also applies to adjacent amplitude values in the frequency domain. One way to correct for this is by setting a ρ-parameter, often corresponding to the autocorrelation at 'lag-1', i.e. the mean correlation between adjacent measurements. This correction, called an AR(1) model, can also be included in FOSR models. AR(1) models are included in all models with ρ set at 0.1 below the

[7] As mentioned above, `refund` (Goldsmith et al. 2021) was used to fit FOSR models. `mgcv` (Wood 2017a, 2021), `itsadug` (van Rij et al. 2020b), and `moments` (Komsta and Novomestky 2015) were used for health checks of the resulting models. `ggplot2` (Wickham 2016; Wickham et al. 2021) was used for visualizations, with added convenience functions from `FoSIntro` (Bauer et al. 2018; Bauer 2021).

[8] See the supplementary data (Puggaard-Rode 2022a) for various residual plots.

lag-1 autocorrelation in a corresponding model with no correction.[9] Autocorrelation along the functional domain in the AR(1)-corrected models is moderate and short-range, and autocorrelation along the temporal domain is relatively moderate and short-range even without correction. Another method for accounting for autocorrelated errors is the use of functional random intercepts, with smoothing parameters set using splines based on functional principal components (Greven and Scheipl 2017a; Bauer et al. 2018; for an introduction to the latter concept, see Section 6.7.1). Pouplier et al. (2017) argue in favor of the latter approach because 1) the influence of random effects can then be more readily decomposed, and 2) the basis for the correction is computed directly from the data, while the parameter setting used for AR1-correction is necessarily somewhat *ad hoc*. The latter approach can also be implemented in `pffr`, but at a significant computational cost.

The models further include by-category smooths for a number of independent binary variables: speaker SEX, following vowel HEIGHT, BACKNESS, and ROUNDING, as well as STRESS. The influence of speaker sex on the spectral profile has not been discussed much above, but is also included here, since previous studies have shown a gender effect on the spectral profile of fricatives (e.g. Stuart-Smith 2007). I am interested only in how these variables affect the time-varying characteristics of spectra, so no main effects were included for these variables. The binary variables are contrast coded (see Schad et al. 2020 and Section 4.5.3), such that absence of the feature in question is coded numerically as -½ and the presence as +½; the SEX variable is (randomly) coded as -½ female, +½ male. Contrast coding categorical variables is similar to centralizing continuous variables, and ensures that the global intercept corresponds to a weighted global mean, which makes the final results much easier to interpret. For each of these

[9]The reason for setting ρ lower than the autocorrelation at lag-1 is that all models show some degree of negative autocorrelation at higher lags, which is exacerbated when ρ is increased; see more details in Puggaard-Rode (2022a).

effects, by-speaker random slopes are also included (except for SEX, which logically cannot vary by-speaker).[10]

As discussed in Section 5.4, I do not report *p*-values for the FOSR models, as they likely reflect the number of observations rather than practical significance. Instead, I explore the model fits through two types of plots: 1) *Spectrum intercepts*, which visualize the functional intercepts of the models, corresponding to an average release spectrum when all other variables are kept at zero. These are not very telling in themselves, but are important for interpreting other effects. The spectrum intercepts are plotted with 95% confidence intervals, computed in the manner proposed by Marra and Wood (2012). 2) Spectro-temporal fits, which visualize the spectrum across time. The interpretation of these is similar to spectrograms; they are 'flipped' spectra, with normalized time along the x-axes, frequency along the y-axes, and greyscale shading indicating differences in fitted amplitude along the time–frequency domains. These visualizations reflect the effect size of different variables. The plots of the main effect of time are computed by combining the functional intercept with the fitted effect of time; the plots of other variables are computed by combining the functional intercept, the fitted (main) effect of time, and the fitted time-varying effect of the variable in question. This means that if the model finds no noticeable effect of time, there will be no noticeable change along the horizontal dimension; if there is no noticeable effect of a particular variable, the plot associated with this variable will be similar or identical to the plotted main effect of time. Since these plots are two-dimensional, visualizing 95% confidence intervals would require separate plots for the upper and lower limits; in order to keep the number of visualizations manageable, I do not include these here, but they can be found online in Puggaard-Rode (2022a). These plots demonstrate the uncertainty associated with each fitted effect. I will refer to these plots only when they show that a variable is associated with a great deal of uncertainty.

[10]Using factor smooths instead of random slopes would have given a more thorough estimation of the by-speaker variation in the data (Baayen et al. 2018; Wieling 2018; Sóskuthy 2021), but unfortunately these cannot currently be fitted with data along uneven grids.

Another way to get an indication of the fitting–complexity trade-off of including an individual variable is by comparing the minimized smoothing parameter selection scores (fREML scores) of a nested model without that variable (van Rij 2016; van Rij et al. 2020a). fREML scores are conceptually similar to information criteria like the Akaike Information Criterion (AIC): a lower fREML score indicates a better model fit.[11] For each variable in each model, I report the increase in fREML score of a nested model 1) without that variable and its associated random slope, and 2) without its associated random slope only. If a variable is associated with a large fREML decrease, this means that including the variable results in a much better model fit, i.e. the variable is very influential. This gives an indication of the relative effect size of each variable, and (in the case of random effects) how much of this can be accounted for with by-speaker variation. Note, however, that it is only meaningful to compare fREML scores within the same model, and not across models; fREML scores can be taken as a proxy for relative effect size, not for statistical significance.

5.6 Results

The results of the three different models will be presented in separate sections below, starting with the model for /t/.

5.6.1 /t/

The model of /t/ releases has a high effect size of R^2 = 0.54. The functional intercept (see Figure 5.2) shows an energy peak around 3.5–5 kHz, with comparatively little energy elsewhere, particularly above 8 kHz. Recall that the intercept summarizes the grand weighted mean over a dynamic series of events, so it is not in itself very meaningful. In the spectro-temporal fits, any changes on the horizontal dimension are a result of spectral characteristics changing as a function of time.

The /t/ model shows a strong main effect of time in the expected direction, as shown in Figure 5.3. Initially, energy is skewed towards the higher end of the spectrum, with fairly strong energy around the

[11] AIC does not provide a reliable test for smooth variables (van Rij 2016).

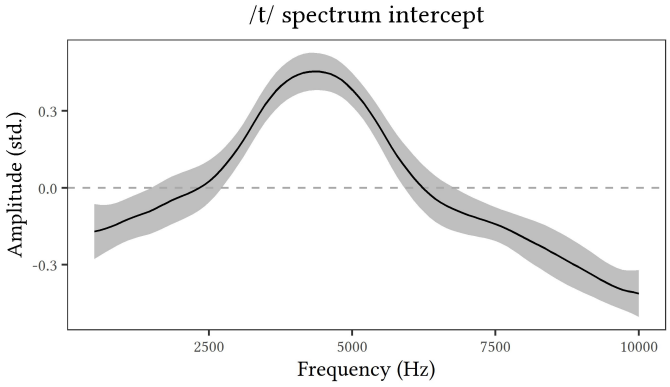

Figure 5.2: *Functional intercept for the model of* /t/ *releases.*

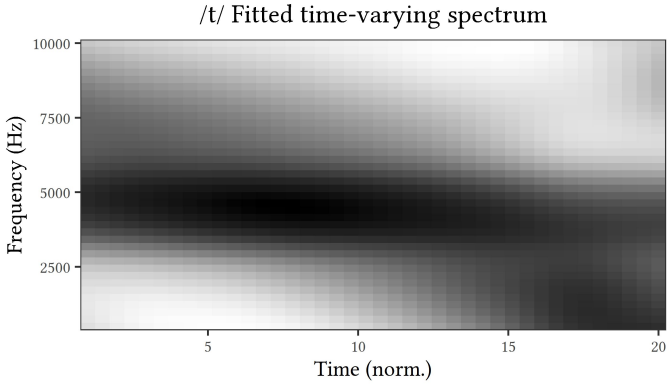

Figure 5.3: *Fitted time-varying spectrum of* /t/ *(main effect of time).*

intercept but also reasonably equal distribution of energy in the 5.5–8 kHz range. Increased energy above the main peak gradually tapers off, and in the final three-fourths of releases, energy is broadly distributed below 5 kHz, including at the lowest frequencies visualized (500 Hz).

Spectro-temporal fits for each direction of the individual variables are shown in Figure 5.4. Table 5.2 shows the reduction in fREML score associated with each variable. Figure 5.4 reflects a residual issue with this modeling technique. In contexts where we expect reduced affrication and earlier onset of aspiration, as in e.g. non-high vowels

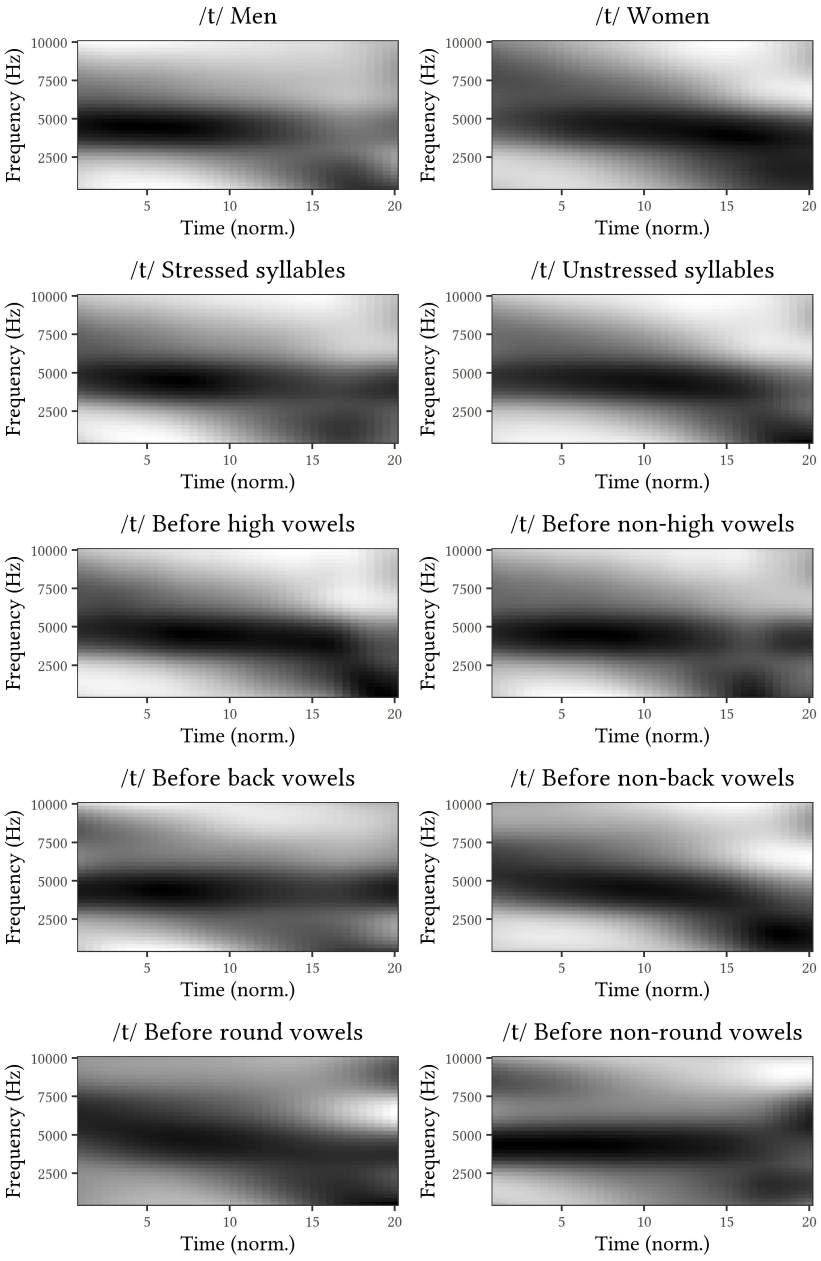

Figure 5.4: *Spectro-temporal fits of* /t/ *for each direction of the individual variables.*

Table 5.2: *fREML score reduction associated with each variable in the* /t/ *model.*

Variable	**fREML reduction**
SEX	1,539
STRESS	1,875
STRESS (random slope only)	1,546
HIGH VOWEL	2,264
HIGH VOWEL (random slope only)	1,384
BACK VOWEL	913
BACK VOWEL (random slope only)	648
ROUND VOWEL	286
ROUND VOWEL (random slope only)	174

relative to high vowels, the figures show a relatively early increase in energy at low frequencies, but also tend to show a sudden final increase in energy at higher frequencies. There is no linguistic reason to expect this, and it is consistent across models; I assume that this is a technical issue that does not reflect the data or the linguistic reality.

Overall, men show relatively little energy above the peak in the intercept spectrum, and lower frequencies (indicative of aspiration)[12] begin dominating relatively early. Women show strong initial energy in frequencies above 5 kHz, and although lower frequencies come into play late in the release, frequencies up to 5 kHz are excited throughout the release. The effect of SEX is strong and associated with a large reduction in fREML score.

Lower frequencies start dominating towards the end of the release in unstressed syllables, and much earlier in stressed syllables. STRESS is an influential variable, although much of its influence is due to the by-speaker random slope. Lower frequencies also dominate relatively late before high vowels, and frequencies above 6 kHz are also more excited at the beginning of the release in this context. This is a very

[12] As mentioned in Section 5.2, during aspiration, low-frequency noise is generated at or near the glottis, and the turbulent airstream excites the resonant frequencies of the oral cavity. The dominance relationship between these sources may differ, but in both cases, the primary frequencies being excited are well below those excited during alveolar frication.

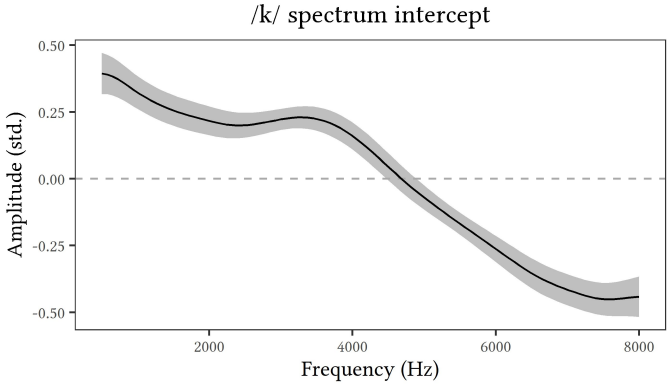

Figure 5.5: *Functional intercept for the model of* /k/ *releases.*

strong effect, which is relatively stable across speakers (i.e. the random slope contributes fairly little.) Lower frequencies dominate relatively early before back vowels and round vowels. In both of these contexts, there is also a coarticulatory effect at the start of the release: relatively high frequencies are excited before round and non-back vowels. These variables are less influential, with the fitting–complexity trade-off being relatively poor for the ROUND VOWEL variable in particular.

It is interesting that none of these variables are particularly influential around the middle portion of the release; they may affect whether particularly high frequencies are excited around the start of the release, and whether/when lower frequencies begin to dominate near the end of the release, but high energy in frequencies around 3.5–5 kHz in the middle of the release is a consistent feature across all variables.

5.6.2 /k/

The model of /k/ releases has a high effect size of $R^2 = 0.57$. Recall that the frequency range for the models of /k/ and /p/ does not extend above 8 kHz. The functional intercept (see Figure 5.5) shows almost evenly distributed energy below 4 kHz, with small peaks around 500 Hz and

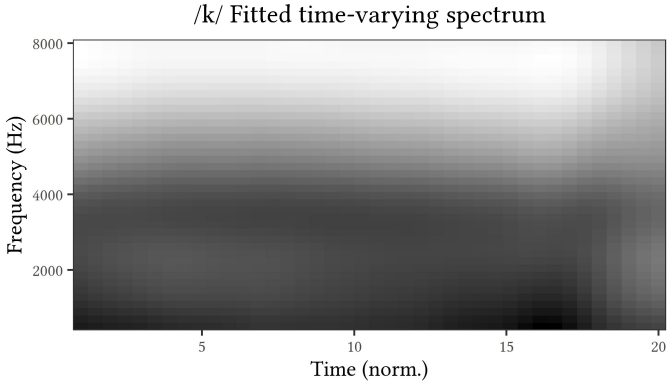

Figure 5.6: *Fitted time-varying spectrum of* /k/ *(main effect of time).*

Table 5.3: *fREML score reduction associated with each variable in the* /k/ *model.*

Variable	**fREML reduction**
SEX	1,708
STRESS	577
STRESS (random slope only)	507
HIGH VOWEL	6,082
HIGH VOWEL (random slope only)	4,785
BACK VOWEL	17,829
BACK VOWEL (random slope only)	13,502
ROUND VOWEL	3,620
ROUND VOWEL (random slope only)	3,231

just below 4 kHz, and linearly decreasing energy between approx. 4–7 kHz.

There is no strong main effect of time; there is little variance in the time domain in Figure 5.6, and the variance that we do see is associated with significant uncertainty (as evidenced by the 95% confidence intervals shown in Puggaard-Rode 2022a). Spectro-temporal fits for each direction of the individual variables are shown in Figure 5.7. Table 5.3 shows the reduction in fREML score associated with each variable.

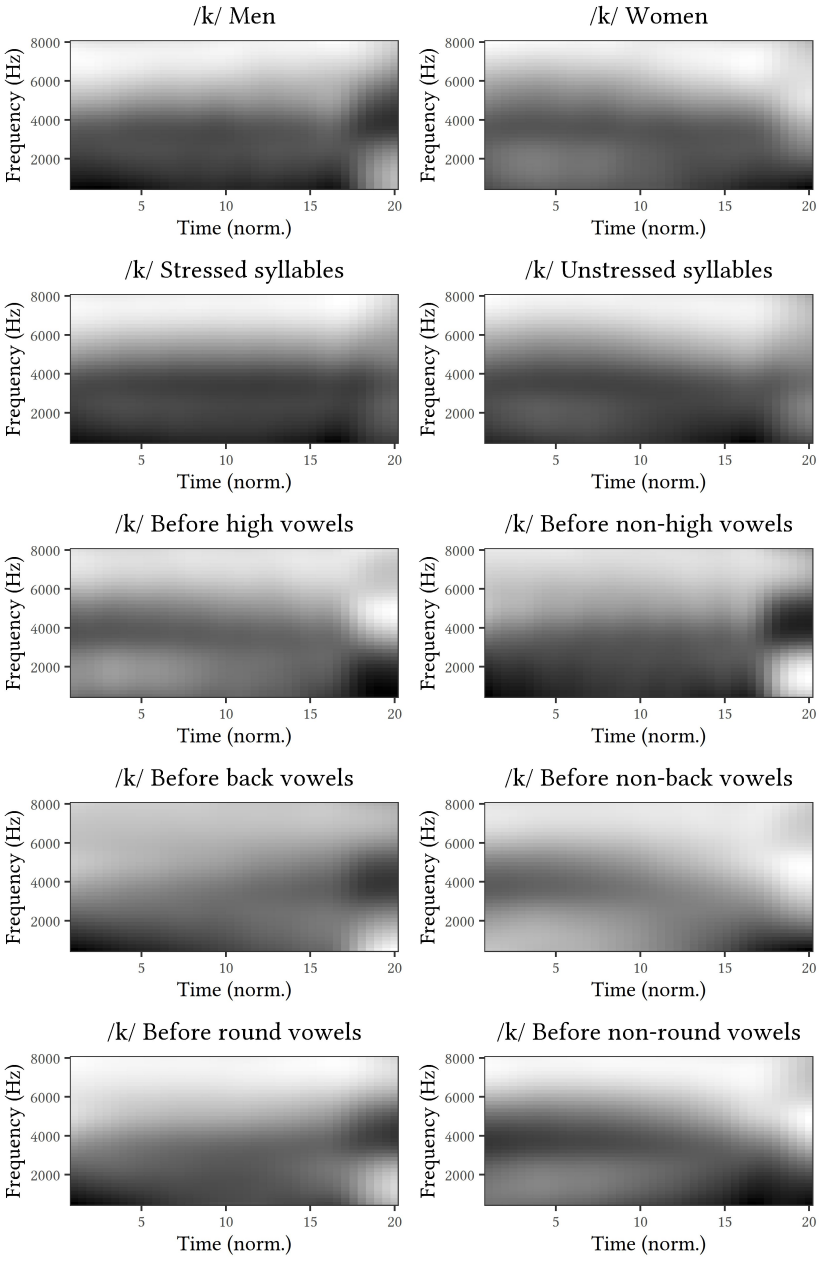

Figure 5.7: *Spectro-temporal fits of* /k/ *for each direction of the individual variables.*

There is a noticeable sex difference, although the associated reduction in fREML is modest. There is little energy at lower frequencies during the first half of the release for female speakers, and more activity at frequencies above 4 kHz. Lower frequencies becomes dominant in the last quarter of the release for female speakers, whereas for male speakers, they are seemingly dominant throughout the release.

As expected, phonetic context effects have a clear influence on the /k/ spectral trajectory, particularly those effects that reflect properties of the following vowel. Stressed syllables have somewhat more energy at the lower band around 500–1,000 Hz, while unstressed syllables have more energy at the higher band around 3.5–4 kHz, although lower frequencies gradually become dominant in the latter half of the release. The size of this effect is modest, and mostly comes down to by-speaker variation; it is also associated with significant uncertainty, as evidenced by 95% confidence intervals (see Puggaard-Rode 2022a).

Before high vowels, there is a lot of high frequency energy between 3–5 kHz during the first half of the release, with more diffuse distribution of energy before the onset of low-frequency noise towards the end of the release; low frequency energy overall dominates releases before non-high vowels. This variable is associated with a large fREML reduction. Non-back vowels and non-round vowels show roughly the same patterns as high vowels, although with slightly varying temporal alignment. The BACK VOWEL variable in particular is associated with a very large fREML reduction. High frequency noise lasts somewhat longer for non-round vowels than non-back vowels. The fREML reduction associated with the ROUND VOWEL variable is also relatively large, although largely a result of by-speaker variation.

5.6.3 /p/

The model of /p/ releases has a very high effect size of $R^2 = 0.71$. The functional intercept (see Figure 5.8) shows most energy in the lowest frequencies around 500 Hz, with energy gradually reducing at higher frequencies. Assuming that the more diffuse distribution of noise towards the end of the release is not linguistically relevant, there is only a very marginal main effect of time (see Figure 5.9).

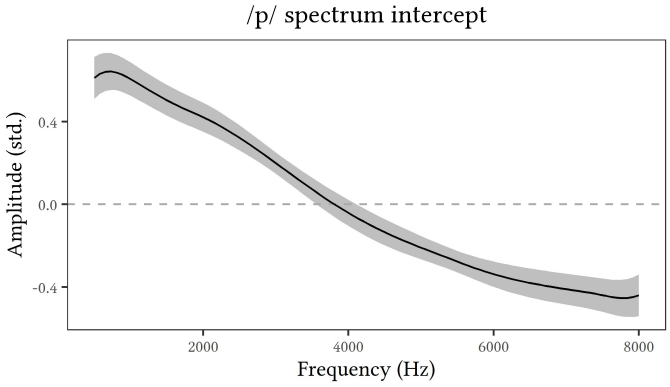

Figure 5.8: *Functional intercept for the model of* /p/ *releases.*

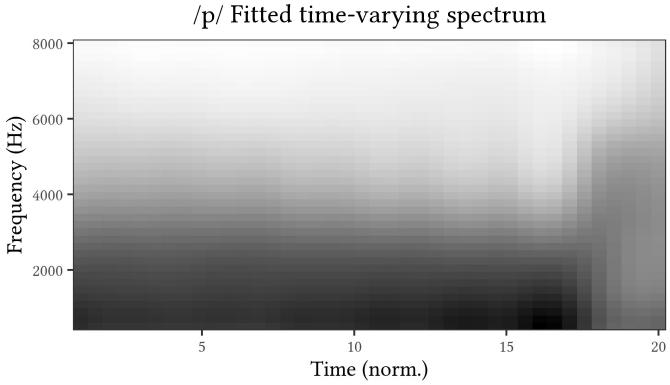

Figure 5.9: *Fitted time-varying spectrum of* /p/ *(main effect of time).*

There are clearer by-variable time-varying characteristics of /p/, as shown in Figure 5.10. Table 5.4 shows the reduction in fREML score associated with each variable. Compared to the other models, random slopes account for a large proportion of the variance in /p/ releases. There are modest signs of higher frequencies being excited more in the first half of releases produced by women, but not by men. The SEX effect is, however, quite weak, and associated with a great deal of uncertainty, as evidenced by 95% confidence intervals (Puggaard-Rode 2022a).

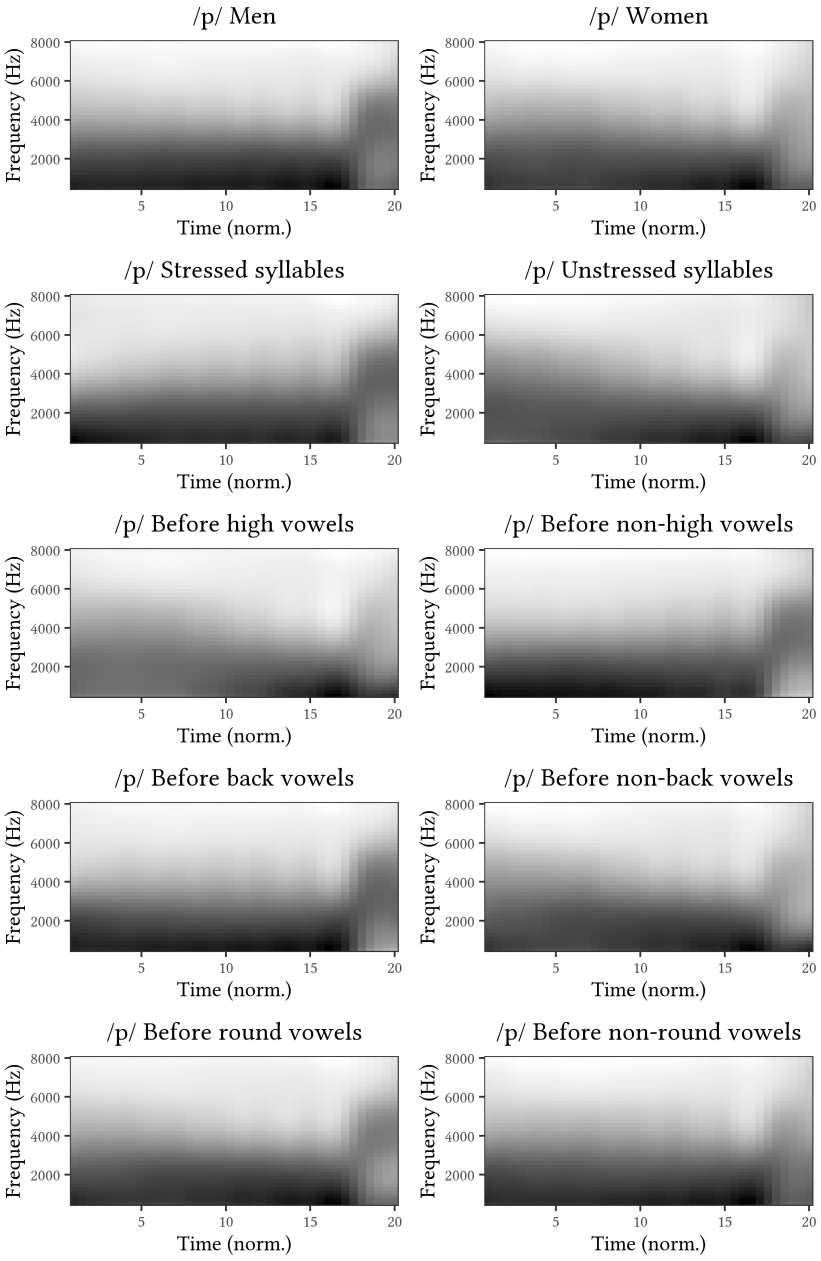

Figure 5.10: *Spectro-temporal fits of* /p/ *for each direction of the individual variables.*

Table 5.4: *fREML score reduction associated with each variable in the* /p/ *model.*

Variable	fREML reduction
SEX	189
STRESS	1,531
STRESS (random slope only)	1,248
HIGH VOWEL	1,123
HIGH VOWEL (random slope only)	1,035
BACK VOWEL	911
BACK VOWEL (random slope only)	804
ROUND VOWEL	494
ROUND VOWEL (random slope only)	460

During the first portion of the release, unstressed tokens have a broader distribution of energy throughout the spectrum, and more energy at higher frequencies (above approx. 5 kHz). The STRESS variable is quite strong, and relative to other variables, quite robust across speakers. A similar pattern is found before high vowels, with lower frequencies dominating relatively late in the release. To a lesser extent, the same pattern is found before non-back vowels. Both of these effects are associated with large fREML reductions, but largely due to by-speaker variation. There is no obvious influence of round vowels, and this variable is associated with significant uncertainty, as evidenced by 95% confidence intervals. The modest fREML reduction associated with the ROUND VOWEL variable is almost exclusively due to by-speaker variation.

5.7 Discussion

5.7.1 Contextual variation in stop releases

In Section 5.6 above, I described the patterns of energy distribution that are visible in the spectro-temporal fits in prose. In this section, I aim to provide a link between those representations and the articulatory mechanisms that presumably underlie them. This discussion is necessarily somewhat speculative, but relies on established knowledge about

the articulation–acoustics link, and about the articulation of Danish specifically.

While all stops show diffuse patterns of energy distribution towards the end of the release, only /t/ clearly shows a strong main effect of time, with a gradual downward trend in energy distribution over time. During the first half of the release, high frequencies are excited, often above and beyond what is necessarily expected for an alveolar constriction. During the second half of the release, lower frequency energy consistent with a glottal noise source gradually becomes dominant. As mentioned in Section 5.2 above, there is reason to assume that oral air pressure is not particularly high at the time of release in Danish aspirated stops, which provides both an aerodynamic reason and a functional–phonological motivation for why the constriction is maintained somewhat longer than in comparable 'aspiration languages': there is no high air pressure to ensure that the constriction is quickly released, and to ensure a salient burst. Nevertheless, contrary to the general conception in literature, alveolar constriction usually does not dominate the entire release.

The relative timing of the shift in dominance from an alveolar noise source to a glottal one is largely determined by contextual factors, in particular stress and vowel height. Speaker sex also plays a role. Stop releases in stressed syllables show a larger proportion of aspiration. In other words, phonetic reduction mainly targets the aspiration in /t/ releases, not the frication. Features of the following vowel affect the relative timing of the dominance shift much more than they affect the distribution of energy during the first half of the release, although high and round vowels do show coarticulatory effects lasting throughout the release. The linguistic upshot is that lengthy alveolar frication is an invariant feature of /t/ releases in Modern Standard Danish, but the proportion of alveolar frication varies; some degree of aspiration is almost always observed.

Stevens' (1998) model of velar stop releases suggested that the velar frication excites low resonance frequencies mostly below 2 kHz. The results here, however, show two primary patterns of energy distribution: much higher resonance frequencies around 4 kHz, or resonance frequencies centered around the lower end of the spectrum. I presume that the former represents a velar noise source – likely fronted,

since a fronted velar constriction leads to a shorter distance between the constriction and the hard palate, which the turbulent air stream partially impinges on – and that the latter corresponds primarily to a glottal noise source. However, it may be difficult to tease apart a noise source in the back portion of the velum and a glottal noise source. The dominant noise source is mostly contextually determined. The main effect of time is marginal, although low-frequency aspiration is overall dominant during the final portion of the release. Before high vowels and non-back vowels in particular, noise at higher frequencies is dominant during the first part of the release. If the following vowel is high, the tongue dorsum logically remains fairly close to the velum throughout the release, causing a dominant dorsal noise source, the characteristics of which vary on the basis of other vowel features. The point of occlusion varies by backness of the following vowel, such that the outgoing air impinges more directly on the hard palate before front vowels, causing more salient noise at higher frequencies. The energy from the palatal noise source is dampened by lip rounding, which increases the size of the oral cavity. The linguistic upshot is that coarticulation has a major influence on spectral characteristics throughout /k/ releases; this is in line with the general observation that the point of occlusion in /k/ is prone to coarticulatory variation (e.g. Ouni 2014).

/p/ releases also vary in whether there is a primary glottal noise source (a strong energy peak at lower frequencies), or whether there is a primary labial noise source (no strong energy peak at lower frequencies). There is no strong main effect of time. In unstressed syllables, before high vowels, and to some extent before non-back vowels, energy is more broadly distributed throughout the spectrum, indicating a dominant labial noise source. /p/ releases vary relatively little compared to /t k/, and much of the variance found in the data is the result of by-speaker variation.

These results confirm the observation that /t/-affrication in Modern Standard Danish is invariant. Generally, however, /t/ affrication does not last throughout the release; aspiration is also an important component of /t/ releases, especially in stressed position. There is also a frication component in /p k/ releases, but under many conditions, these releases are dominated by a glottal noise source. During a /t/ release,

the outgoing air impinges on a hard surface – the teeth – immediately downstream of the preceding occlusion. This is not the case for either /p/ or /k/; the lips constitute a soft surface, and the hard palate is further removed from the velar occlusion. As such, it is well-understood why an alveolar noise source dominates a glottal one more readily than corresponding bilabial or velar noise sources.

5.7.2 Function-on-scalar regression and the spectrum

This chapter has introduced the use of FOSR in the analysis of speech spectra and their variance as a function of time. This method shows a lot of promise. It allows us to get around the problem of choosing one or a few discrete measures to represent the spectrum, all of which come with their own set of methodological problems. In a sense, analyzing these models is similar to the classical technique of 'eyeballing' spectrograms, but in a way that allows the user to more efficiently and reliably find systematic patterns of variation in the data, to tease apart various influences on the results, and to filter out by-speaker variation. Some lingering issues remain with the method; some specific to this study, and some inherent to the field. I will briefly address a few of these.

As with any kind of quantitative phonetic study, there are significant researcher degrees of freedom involved in FOSR modeling of spectra (see Roettger 2019). Token selection, spectral estimation, smoothing procedure, low-level software implementation, as well as several other factors all have a potentially non-trivial influence on the results. There is no easy remedy to this, but transparent reporting and motivation of all these choices goes a long way. I have aimed to do that here, and the actual code used to implement the analysis is available in annotated form (Puggaard-Rode 2022a).

FOSR modeling of spectra quickly leads to highly multidimensional data, especially if the temporal dimension is also taken into account, and this makes the use of traditional methods for significance testing problematic. I do not consider this to be an issue in the current study. For one, the study is largely exploratory, and the research questions are not necessarily suitable for null hypothesis significance testing. With that said, there are methods for testing the stability of the results. This

includes the 95% confidence intervals proposed by Marra and Wood (2012), which I have occasionally referred to here, and include in the online appendix to this chapter (Puggaard-Rode 2022a); this method is implemented for FOSR visualization in the `FoSIntro` package in R (Bauer 2021). Additionally, there are functional implementations of discriminant analysis and regression trees which may be used to explore the generalizability of results, and fully Bayesian implementation of the analysis would make it possible to readily quantify the uncertainty related to the results (see e.g. Vasishth et al. 2018b). This will hopefully be explored in future research, but is beyond the scope of the current study. The prospects of hypothesis testing in FOSR models is explored in a recent dissertation by Biswas (2022).

The implementation of FOSR in this study shares a problem with analyses based on e.g. spectral moments, mid-frequency peaks, and DCT: the Hz-based frequency scale and the W/m^2-based amplitude scale are 'physicalist' in nature, in that they represent the behavior of vibrations in the air, and not how these vibrations are perceived by the human ear (Plummer and Reidy 2018). I use the Hz scale here because it results in a model output which is more immediately interpretable for readers with experience with analyzing spectrograms; I use the W/m^2 scale because it results in more clearly interpretable patterns in the fitted models. It is, however, worth exploring in future studies how the results would be affected by combining perceptually motivated scales, such as the equivalent rectangular bandwidth (ERB) scale and the decibel scale.[13]

The most serious lingering issue is the diffuse patterns sometimes seen in the final time steps of the spectro-temporal visualizations. These cannot be considered linguistically meaningful; there is no linguistic reason why high frequencies above 4 kHz would suddenly be excited immediately before the onset of voicing in a stop–vowel sequence. I can see three possible explanations for this: 1) the spectral

[13] Alternatively, the positions of knots used for smoothing could be placed according to a (semi-)logarithmic scale, e.g. giving the model higher granularity in frequency regions where humans have greater perceptual acuity. This could potentially achieve a similar effect while keeping the 'physicalist' scales. In this study, the knots are equidistantly spaced, but `mgcv` and consequently `pffr` allow the user to specify knot locations freely.

characteristics of aspiration are highly variable, making it impossible for the model to make precise predictions, 2) the pseudo-centralization of categorical variables sometimes causes the model to infer patterns that are not meaningful for one value of variables, or 3) it is caused by phase variation. Regarding 2), consider /k/ before high and non-high vowels: the model finds a strong increase in low frequency energy in the final time steps before high vowels, which is linguistically meaningful, as the glottal noise source becomes dominant immediately before the onset of voicing. The model finds a corresponding increase in high frequencies and decrease in low frequencies in the final time steps before non-high vowels, which is *not* linguistically meaningful, but is the direct opposite of the meaningful finding before high vowels. A possible solution would be to fit the model without contrast-coded categorical variables, but this would make it impossible to interpret models' intercepts and main effects of time, which I believe would seriously harm the interpretability of the findings. Regarding 3), phase variation is a practical problem in functional data analysis, where lateral displacement in input curves can cause results to be blurred and distorted. Managing phase variation in the analysis of functional data is an area of active research (Marron et al. 2015; Bauer et al. 2021)

5.8 Conclusion

The study presented in this chapter is, to the extent of my knowledge, the first to use function-on-scalar regression to analyze sound spectra. This method forgoes the need to boil down the complex, multi-dimensional information in the spectrum to a few discrete values, and it forgoes the need to rely on 'magic moments' in time. By plotting the fit of a FOSR model, we can explore the systematic influences of different variables on the spectrum with visualizations that should be intuitively familiar to anyone used to working with spectrograms. I showed how this tool can be fruitfully applied in the analysis of Danish stop releases, how their spectral characteristics vary vary over time, and how they are affected by their phonetic environments.

The analysis finds that /t/, as expected from the literature, is invariably affricated, but also that the spectrum is very dynamic throughout /t/ releases, with dominant affrication gradually being replaced by dominant aspiration. Affrication dominates the majority of the spectrum, and much of the aspiration is lost in unstressed syllables. Coarticulatory context effects may affect the entirety of /t/ releases, and not just the final portion. Coarticulatory context effects greatly influence the spectra of /k/ releases, particularly in the first portion of the release. The acoustic characteristics of /p/ releases show a lot of by-speaker variation, but also coarticulatory context effects, mainly in the first half of the release.