Analyzing time-varying spectral characteristics of speech with function-on-scalar regression

Puggaard, R.

Journal of Phonetics 95 (2022) 101191

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/Phonetics

Research Article

Analyzing time-varying spectral characteristics of speech with function-on-scalar regression

Rasmus Puggaard-Rode

Leiden University Centre for Linguistics, Netherlands

ARTICLE INFO

Article history:
Received 22 December 2021
Received in revised form 15 August 2022
Accepted 14 September 2022

Keywords:
Functional data analysis
Function-on-scalar regression
Stop releases
Affrication
Danish
Corpus phonetics

ABSTRACT

The acoustic characteristics of noise from fricatives and stop releases are difficult to analyze. The spectral characteristics of such noise are multi-dimensional, and popular methods for analyzing them typically rely on reducing this complex information to one or a few discrete numbers, such as spectral moments or coefficients of discrete cosine transformations. In this paper, I propose using function-on-scalar regression models as a method for analyzing and mass-comparing spectra with minimal reduction of the complexity in the signal. The method is further useful for analyzing how spectra change as a function of time. The usefulness of this method is demonstrated with a corpus analysis of Danish aspirated stop releases, using the DanPASS corpus. The analysis finds that /t/ releases are invariably affricated; /k/ releases are highly affected by coarticulatory context; and /p/ releases are almost always dominated by aspiration in the latter half of the release, but are affricated in the first half in certain contexts.

1. Introduction

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 would be formants; for obstruent consonants, examples would be spectral moments and coefficients of discrete cosine transformations. The goal of this paper is to demonstrate function-on-scalar regression (FOSR; Reiss et al., 2010; Greven & Scheipl, 2017a; Bauer et al., 2018) as a method for taking the entire spectrum into account when analyzing sources of variance in the acoustic signal. Rather than relying on discrete measurements, FOSR allows for the use of complete spectra as dependent 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 & Kavitskaya, 2018) and generalized additive models (Nance & Kirkham, 2020), and 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 characteristics of speech spectra.[1]

As a case study, I focus on the spectral characteristics of Danish stop releases, how they change over time, and how they are affected by various contextual variables. It is well-established that the aspirated alveolar stop /t/ in Standard Danish is usually strongly affricated. This was pointed out early on by Otto Jespersen (1899: 335), who noted that foreigners would often perceive Danish /t/ as an affricate [ts] – particularly before high front vowels. He maintained that /t/ was still 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. He assumed that /t/ was most advanced in this sound change, followed by /k/, and finally /p/. This paper uses function-on-scalar regression models to explore the outcome of Jespersen's prediction more than a century later. This is not straightforward:

E-mail address: r.p.hansen@hum.leidenuniv.nl

[1] Wood (2017: 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.

https://doi.org/10.1016/j.wocn.2022.101191
0095-4470/© 2022 The Author(s). Published by Elsevier Ltd.

the boundary between an aspirated stop and an affricated one is fuzzy, as is the boundary between an affricated stop and a proper affricate. I approach this question by looking holistically and dynamically at time-varying spectral characteristics throughout stop releases, and how they vary, in a corpus of spontaneous spoken Danish.

The results of the case study show that /t/ releases are invariably affricated, but that aspiration also plays an important role in /t/ releases. Affrication is also found in /p k/ releases, but only plays an important role in certain phonetic contexts, which vary by place of articulation. The state of affairs does not seem to have changed much since Otto Jespersen's time: /t/ is still heavily affricated, and affrication remains less prominent in /p k/.

In the following subsection, I introduce a general problem in acoustic phonetic research: there are multiple competing methods for measuring frication noise, which all rely on boiling down the information available in the spectrum to a small number of variables. Subsequently, I argue that smoothing-based approaches to dynamic data analysis, and FOSR in particular, may be solutions to this problem. In Section 2, I introduce the case, viz. the releases of aspirated Danish stops and their position on the blurry aspirated–affricated continuum, as well as the data used to explore this case. In Section 3, I describe three statistical models, one for each of /p t k/, and the results of these models are presented and interpreted. In Section 4, I discuss the advantages and challenges of using FOSR for analyzing spectral variance, and in Chapter 5, a brief conclusion is given.

1.1. 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 & 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 & Halle, 1956; Strevens, 1960).

Forrest et al. (1988) popularized treating the complex spectrum as a probability mass function, and analyzing it by calculating four moments: 1) the 'mean frequency', or 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 distinguished fairly well between places of articulation in stop bursts, and that particularly COG, skewness, and kurtosis distinguished fairly well between places of articulation in alveolar and postalveolar 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. Subsequent studies testing this have not been particularly stable (see e.g. Shadle & Mair, 1996), but COG in particular has remained a very popular measure in the analysis of spectral properties of fricatives often without taking into account higher moments. 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 very full picture of the shape of the distribution, and spectra with quite different shapes may have the same COG. As an example, Gordon et al. (2002: 152ff.) find nearly identical mean COG in Western Aleut [x χ], but the spectral shapes are otherwise quite distinct.

A number of other candidate measures have been proposed in the literature for analyzing frication, mainly for determining the precise place of articulation of 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 – the frequency with the highest amplitude within a 3000–7000 Hz band – captures the fairly subtle difference between labialized and non-labialized /s/ in adolescents.

Another proposed method is using cepstral coefficients derived from a discrete cosine transform of the spectrum (DCT; Watson & 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 yield more correct classifications of place of articulation than spectral moments in both voiceless stops (Bunnell et al., 2004) and fricatives (Spinu & 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 have often been taken at fixed points in time, such as the midpoint or some pre-determined range around the midpoint of fricatives or affricates; 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 these sounds; affricates are inherently dynamic, and Reidy (2016a) shows 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 is a relatively small initial portion of the release (see e.g. Chodroff & Wilson, 2014).

Summing up, most existing approaches to quantifying frication reduce the complex time-varying information from spectra to something more manageable. This is very reasonable, because 1) many statistical methods frequently used in linguistics cannot necessarily 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 into account complex dynamic information are increasingly being used, as discussed below, and this paper 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 if we have not tested complex models. Studies mentioned above have shown 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 [ç ʃ] in German (which are currently undergoing a merger) are more readily apparent when using a measure which takes more of the spectrum into account (i.e., using DCT coefficients rather than moments).

All the above-mentioned measures are derived from the spectrum rather than directly from the waveform, so it follows that the method used for spectral estimation may have an influence on the results. The most common method is fast Fourier transformation (FFT). FFT is very efficient, but in some ways not particularly suitable for acoustic data. The Fourier basis is periodic, making FFT highly suitable for periodic data, such as voiced portions of speech, but less suitable for noisy, randomly generated data, such as voiceless portions of speech (Ramsay & Silverman, 2005: ch. 3.4). For noisy data, it is theoretically preferable to use a lower variance spectral estimate such as that provided by multitaper spectral estimation (Blacklock, 2004). Note however that Reidy (2015) showed that the spectral moments derived from FFT spectra and multitaper spectra may be practically equivalent.

1.2. Smoothing approaches to the analysis of dynamic data

In the past years, following Baayen's (2008) popularization of mixed-effects regression models in linguistics, there has been a rapid increase in the use of sophisticated statistical techniques in linguistics. A general problem in the field has been the analysis of dynamically varying data, in particular data from time series. If some measure – say, COG – 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; variation as a function of time is often nonlinear. A solution to this problem is the use of smoothed curves. Given a number of data points associated with e.g. a time series, a smoothing function (see de Boor, 2001; Wood et al., 2016) can be used to approximate a continuous curve corresponding to the data's non-linear variation as a function of time. Smoothing involves reducing the observations to a number of basis functions (or 'knots'), and using a penalizing smoothing parameter to determine the smoothness or 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 in the signal; conversely, combining too few basis functions with a high smoothing penalty will likely lead to underfitting, resulting in curves that omit relevant information in the signal.

Generalized additive (mixed) models (GAMMs) have quickly become popular in linguistics (see e.g. Baayen et al., 2017; 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 e.g. EEG registration (Baayen et al., 2018; Voeten, 2020: ch. 5), geo-linguistic variation (e.g. Wieling et al., 2011, 2014; Puggaard, 2021; Puggaard-Rode, forthc.), and speech spectra (Nance & Kirkham, 2020) dynamically.

Functional data analysis (FDA; Ramsay & Silverman, 2005; Ramsay et al., 2009; Gubian et al., 2015; Pouplier et al., 2017) has overall had less influence on statistical modeling in linguistics. FDA is a family of statistical methods which extend existing methods to account for functional data. In practice, this means that *curves* can be used as input variables rather than discrete values. An example is the functional extension of principal components analysis (FPCA), which can be used to determine the primary sources of variation in curves. Gubian et al. (2015) use FPCA to jointly analyze how F1 and F2 pattern in the realization of diphthongs and hiatuses in Spanish, respectively, and Puggaard-Rode (forthc.: ch. 6) combines GAMMs and FPCA to analyze how speech spectra of stop release midpoints vary geographically in traditional regional varieties of Danish.

1.3. Function-on-scalar regression

Functional regression models are suitable when one or more variables are of a functional nature (Bauer et al., 2018). 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. This could e.g. be useful for researchers seeking to predict reaction times from pitch contours (e.g. Cutler, 1976); pitch contours consist of complex functional data, which will otherwise have to be either simplified or tightly controlled in the experimental set-up. 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. This, in contrast, could be useful for researchers seeking to predict the shape of pitch contours from a range of predictor variables, such as pragmatic context or duration. It is likewise useful for modeling how the spectrum of a speech sound is affected by e.g. contextual variables, as I will show below.

There are several approaches to modeling function-on-scalar data (an overview is given in Greven & 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). The model can be summarized with the formula below, from Bauer et al. (ibid.: 353).

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

$g(\cdot)$ is a pre-specified link function mapping the predictor to the functional domain; in the case of a Gaussian model, this is simply the identity. The expected value $\mathbb{E}(\cdot)$ of each observation $i = 1, \ldots, n$ 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 the sum of 1) the functional intercept $\beta_0(t)$, 2) 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 3) 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), as well as handling of autocorrelated residuals (Baayen et al., 2018), and which 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).

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`. 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. An advantage of using `refund` rather than `mgcv` to fit FOSR models is that dependent and independent functional variables can be explicitly included in model formulas, allowing the user to conceptualize a problem as FOSR. In a model of spectral variance formalized in `mgcv`, the response variable would have to be an amplitude measure, whereas in a model formalized in `refund`, the response variable can be the spectral shape, which is conceptually more satisfactory. Note that this has no influence on how the models are fitted ‘under the hood’; from a computational perspective, they are the same (e.g. Morris, 2017). In Appendix A, I show how a simplified version of the models presented in Section 4.1 are fitted with `pffr`, and compare this to how that same model would be fitted with the `bam` function in `mgcv`. Furthermore, the model fitting and selection procedure is fully documented in annotated form in Puggaard-Rode (2022).

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 nonlinear variables in GAMMs, and the results of this test are also reported in the output of `pffr`; however, researchers should exercise caution in interpreting the resulting *p*-values, 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, but holds for essentially all frequentist models with very high sample sizes.

In any case, *p*-values and associated measures of nonlinear 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 such as this one 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 nonlinear 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 at a significant computational cost.

2. The case: Stop releases in Danish

It is not a goal of this paper to determine whether /p t k/ are phonetic affricates in Danish. The boundary between an aspirated stop and an affricated one is fuzzy, as is the boundary between an affricated stop and a proper affricate. In the end, 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. I focus on the following broad questions, which are more readily answerable:

- How do the spectral characteristics of Danish stop releases vary across time?
- How are their time-varying characteristics 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.

In the section below, I discuss the aspirate–affricate continuum from a theoretical perspective. Subsequently, I discuss the Danish stops /p t k/ and their position on this continuum. Finally, I introduce the corpus used for the study (DanPASS; Grønnum, 2009), and introduce the acoustic analysis of the data.

2.1. Aspirated stops, affricated stops, and affricates

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 noise generated at the area around the vocal folds continually excites the vocal tract. The different stages of an aspirated stop release are shown on a spectrogram in Fig. 1; visualizations of stops in this section are all from the corpus used for the case study (see Section 2.3.1).

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 5000 Hz to be excited. In velars, the turbulent air

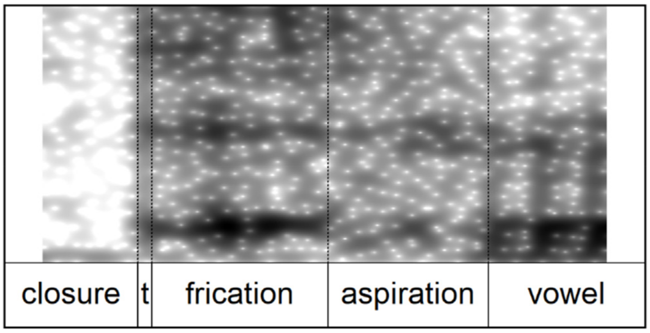


Fig. 1. Different stages of an aspirated stop release, exemplified in a token of /k/ before a low back vowel. 't' is short for 'transient'. The spectrogram shows frequencies from 0 to 8000 Hz.

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 2000 Hz; 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. Examples of aspirated bilabial, alveolar, and velar stops are shown in Fig. 2; /k/ is shown before a high front vowel and a low back vowel to visualize the clearly variable spectral characteristics in these environments.

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 (as in vowels) 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 & 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 & Blumstein, 1978; Blumstein & 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, making frication all the more important as a place cue. Frication is also usually more acoustically salient in aspirated stops relative to unaspirated 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, 1975, 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 & Hamann, 2006; Hall et al., 2006).

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)[2]; 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.

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.

2.2. Danish aspirated stops

In Section 1, I mentioned Otto Jespersen – his early observation that /p t k/ were becoming increasingly affricated, and his claim that they would eventually develop into affricates. Today, more than a century after Jespersen's observations, the affrication of /t/ has been established several times over, has been shown to be exceptionless, and is taken for granted in the literature. Fischer-Jørgensen (1954: 50) reports that the frequency range of the aspiration noise of /t/ is "exactly the same" as for /s/ throughout most of the release. While it is common for the initial burst noise of stops to have a similar frequency range to homorganic fricatives, this usually makes up a smaller portion of aspirated releases. Brink and Lund (1975) tracked the development of affrication in /t/ across more than a century of recordings from Copenhagen, and showed that it went from a widespread phenomenon in the mid-19th century to an exceptionless phenomenon in the mid-20th century. While earlier sources would often transcribe /t/ with both affrication and aspiration (e.g. Brink & Lund, 1975; Basbøll & Wagner, 1985), [t^{s}] has emerged as the standard transcription in the last few decades (e.g. Grønnum, 1998). Recently, Schachtenhaufen (2022) has proposed acknowledging the sound as a pure affricate and transcribing it as simply [ts].

[2] In a traditional binary feature account, affricates are often represented with both [–continuant] and [+continuant] (e.g. Sagey, 1986).

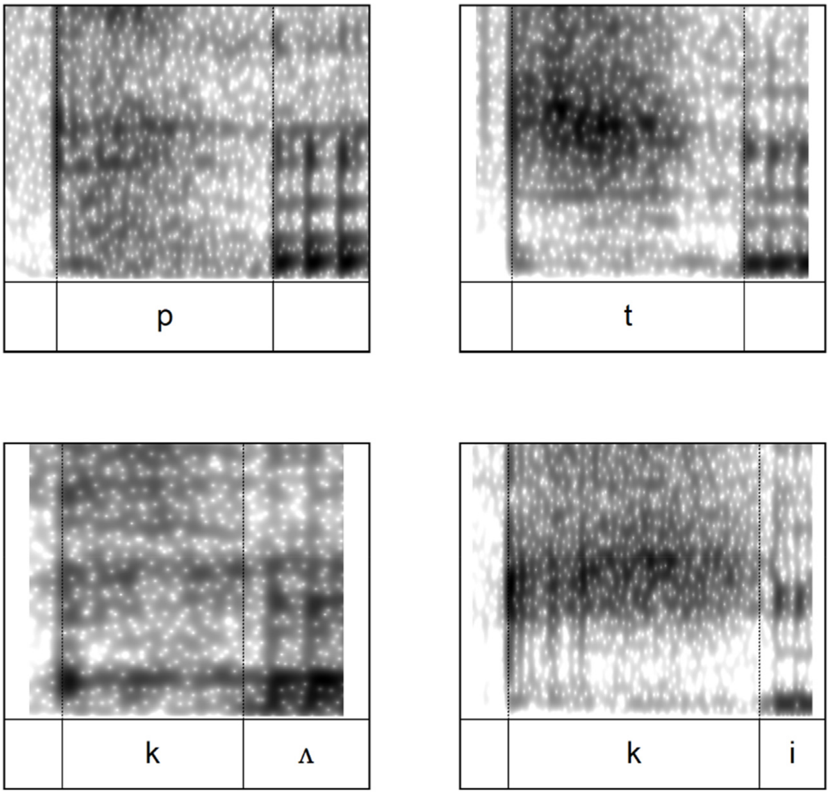


Fig. 2. Examples of aspirated stops at different places of articulation. The spectrograms show frequencies from 0 to 8000 Hz.

However, both frication and aspiration are often clearly present in /t/ releases, as exemplified in Fig. 3.

While there is consensus about the affricated status of /t/, possible affrication patterns in /p k/ have never been investigated. The most straightforward explanation for this is that no one ever noticed salient affrication in /p k/. This could either be because there truly is no affrication in /p k/, or because labial and dorsal frication are simply less salient than coronal frication. On the one hand, since /p t k/ show class behavior in other matters (e.g. phonotactics; Vestergaard, 1967), 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 a large-scale study of American English /p t k/.

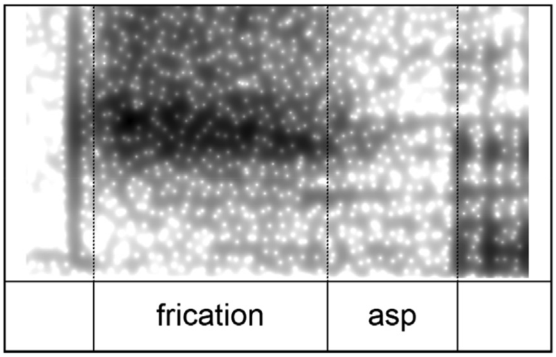


Fig. 3. Examples of a Danish /t/ release with a clear aspiration phase. The spectrogram shows frequencies from 0 to 8000 Hz.

The timing of gestures in Danish aspirated stops is seemingly different from comparable Germanic languages. In Icelandic and Swedish, peak glottal opening is achieved relatively early during the closure of aspirated stops (Pétursson, 1976; Löfqvist, 1980); also in English and German, 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 & Weismer, 1983; Braunschweiler, 1997). This ensures that supraglottal air pressure is high at the time of the release. In Danish, however, peak glottal opening typically falls just after the release of the stop (Frøkjær-Jensen et al., 1971), and closure duration is shortest in unaspirated stops (Fischer-Jørgensen, 1969, 1972). 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. This can partially explain why a constriction is retained for relatively long during Danish stop releases, since higher air pressure at the time of release in itself causes quicker movement of the articulators as the muscular tension forming the constriction is released. 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 constriction following the release.

From a phonological perspective, Danish /p t k/ show similar behavior. Their phonotactic behavior is similar to that of unaspirated stops (Vestergaard, 1967), and they all show the same patterns of positional allophony, with loss of aspiration (or affrication) after /s/ and medially before schwa, and loss of release syllable-finally (although they are optionally released phrase-finally; Grønnum, 2005). 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 following examples; etymologies are from DSL (2018).[3]

[sɑːˀ]	*tsar*	‘czar’	from Russian [tsarʲ]
[suˈkʰiːni]	*zucchini*	‘zucchini’	from Italian [tsukˈkino]
[sɛn]	*zen*	‘zen’	from Japanese [dzen]
[ˈsyʁ̯ek]	Zürich	‘Zurich’	from German [ˈtsyːʁɪç]
[suˈnɑːmi]	tsunami	‘tsunami’	from Japanese [tsɨnami]

In a study of Danish speakers’ productive acquisition of Standard Chinese coronal obstruents (Puggaard, 2020), 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/.

2.3. Methods and materials

2.3.1. The DanPASS corpus

The data for this study comes from the Danish Phonetically Annotated Spontaneous Speech (DanPASS) corpus (Grønnum, 2009, 2016). This corpus was established to obtain high-quality recordings of unscripted Danish speech. The recordings are of single speakers or pairs of speakers solving unscripted tasks. The corpus has already served as the basis for studies on plosive reduction (Pharao, 2011), the relationship between vowel height and voice onset time (Mortensen & Tøndering, 2013), and intervocalic stop voicing (Puggaard-Rode et al., 2022a), as well as a number of other studies; see Grønnum (2016) for a full list.

The corpus consists of monologues recorded in 1996, and dialogues recorded in 2004. Only the monologues are used in this study, since these are more simple to analyze. These consist of 171 minutes of speech from 18 speakers. 13 were men and 5 were women, and they were between 20 and 68 years old at the time of recording (mean = 29 years). Each speaker contributed a mean of 9m27s (range 6m13s – 15m49s). Age is not taken into account in the statistical modeling of the data, so the heterogeneity in age across speakers should not be a problem, especially since we have no reason to assume that there was significant change in the realization of /p t k/ across the speakers’ age span. The gender imbalance should also not be a problem, since most mixed modeling frameworks (including the one used here) do not assume balanced variables (e.g. Wood, 2017: ch. 2).

The recordings are segmented at the levels of prosodic phrase, word, and syllable, and transcribed both orthographically, phonemically, phonetically, and prosodically, as well as coded for morphology and parts-of-speech. Technical details about the recordings can be found in Grønnum (2009). The monologues consist of speakers performing three different tasks. In the *network* task (Swerts & Collier, 1992), they describe various shapes and colors. In the *city* task (Swerts, 1994), they describe routes drawn on a city map. In the *house* task (Terken, 1984), they describe how to build a house model using provided building blocks.

2.3.2. Acoustic analysis

The initial acoustic analysis was done using Praat (Boersma, 2001; Boersma & Weenink, 2019). An automated script was used to locate all aspirated stops (i.e. members of /p t k/) in simple onsets in the DanPASS monologues, and combine them into a single sound file with a subset of the (relevant) original annotations. This was an edited version of the script used by Puggaard-Rode et al. (2022b) written by Dirk Jan Vet. 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 taken as the beginning of the release and the first signs of periodicity taken as the end (following Francis et al., 2003). If multiple bursts were present, the final one was chosen (following Cho & Ladefoged, 1999). This process was partially automatized by a Praat script searching for sudden increases in amplitude, but required extensive manual correction. An example of a segmented /t/ release can be seen in Fig. 4. 205 tokens were excluded during this process if there was no discernible closure. The distribution of stops by phonemic category is shown in Table 1, along with the mean duration of release for stressed and unstressed tokens. This is equivalent to positive voice onset time (VOT). Note that in some cases, the mean VOT values differ quite dramatically from those reported by Mortensen and Tøndering (2013), also on the basis of the DanPASS corpus; this is likely because they follow Fischer-Jørgensen and Hutters (1981) in going by the onset of higher formants rather

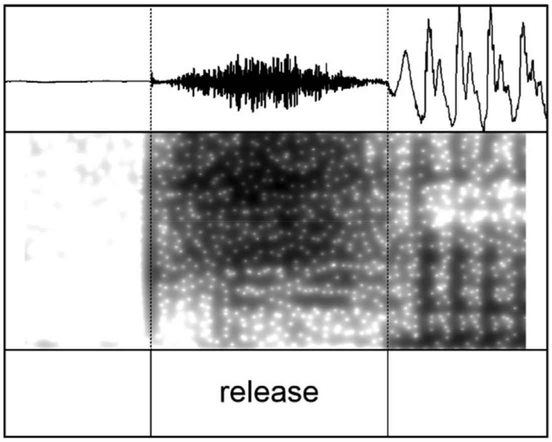


Fig. 4. Example of a segmented /t/ token.

Table 1
Number of aspirated stops included in the study, along with mean VOT values.

Phoneme	Number	Mean duration (stressed), ms	Mean duration (unstressed), ms
/p/	642	57	41
/t/	850	79	68
/k/	842	59	46

[3] 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/.

than the first signs of periodicity, which leads to higher overall values, particularly for /k/. Fischer-Jørgensen and Hutters (1981) recommend this landmark because it yields relatively stable VOT values across vowel types; however, voicing is clearly and consistently present in the waveform and spectrogram before the onset of higher formants, meaning that this landmark is not suitable for an analysis of spectral variance where we would like to avoid the presence of intrusive voicing and (in this case) the first formant.

Subsequently, a Praat script (available in Puggaard-Rode, 2022) was used to extract the release duration and information about the phonetic context from each stop. The phonetic context in question is four binary variables describing the height, backness, and rounding of the following vowel, as well as stress. For this purpose, [i y u ɪ ʏ ʊ e ø o] are all defined as high vowels. Danish has a very complex vowel system, and these vowels all have a mean F1 below 400 Hz in modern Standard Danish (Grønnum, 1995; Juul et al., 2016). [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 stop was split 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. The recordings are filtered to include a frequency range of 500–12,000 Hz. Frequencies below 500 Hz were filtered away to avoid a potential influence of intrusive voicing or low frequency ambient noise. Frequencies above 12,000 Hz were filtered because they rarely play a significant role in speech. In fact, 12,000 Hz is a relatively high cut-off point compared to similar studies, but was chosen due to a study by Pharao and Maegaard (2017) on sociolinguistic variation in Danish /t/ showing that mean COG for fronted realizations of /t/ can go above 6000 Hz, suggesting that very high frequencies may occasionally play a role. For each time step, the four first spectral moments were also extracted; the spectral moments are not used for this analysis, but are published alongside the other data used for the analysis (Puggaard-Rode, 2022).

Multitaper spectra were generated for each time step using R (R Core Team, 2020; RStudio Team, 2021).[4] The study relies on multitaper spectra for both theoretical reasons (outlined in Section 1.1) and practical reasons. From a practical perspective, multitaper spectra consist of fewer frequency bins than FFT spectra, making their use in statistical models less computationally expensive. The parameters used for spectral estimation (the number of tapers $K = 8$ and the time-bandwidth parameter $nW = 4$) were set to match previous studies by Romeo et al. (2013) and Reidy (2015).

Three tokens of /k/ were excluded because their total release duration was below 10 ms, and the algorithm used to generate the spectra does not function for sound files shorter than 0.5 ms. The multitaper spectra are the dependent variable in the statistical analysis; each consists of a vector of amplitude values by frequency ranges. The study uses raw amplitude on the W/m^2 scale rather than the more commonly used transformed amplitude on the decibel scale; models were tested on both scales and yielded practically similar results, but the results are somewhat easier to interpret when using raw values. This is discussed in more detail in Puggaard-Rode (forthc.: chs. 5–6). The frequency ranges differ in size depending on the duration of the time step; longer time steps have more fine-grained spectra. For each spectrum, the amplitude values were standardized,[5] since plenty of non-linguistic factors can lead to deviations in overall amplitude level. Only the frequency range between 500–10,000 Hz was used for the statistical analysis of /t/ spectra, and 500–8000 Hz for /p k/, since the minor activity above these limits seemed to be essentially random noise, which interfered with the clarity of the results.

3. The study

This section covers the statistical analysis and interpretation of the results. I first discuss the model specification, present the results for each stop in turn, and link the results to their presumed underlying articulatory mechanisms.

3.1. Model specification

All statistics were calculated using R (R Core Team, 2020; RStudio Team, 2021) and a number of add-on packages.[6] All code is freely available in annotated form online (Puggaard-Rode, 2022); this includes various residual and autocorrelation plots. 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/), which 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 sparsely distributed data, i.e. when the number of observations per function differs (Wood, 2017b). Time step is included as a non-linear independent variable, smoothed with thin plate regression splines (Wood, 2003) with 16 basis functions to ensure quite high granularity in the temporal dimension. Smoothing penalization parameters were automatically selected using fast restricted maximum likelihood estimation (Wood, 2011). The residuals for the models are reasonably normally distributed, 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 & 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

[4] This was done using the add-on packages `tuneR` (Ligges, 2021) and `multitaper` (Rahim, 2014; Rahim & Burr, 2020), with convenience functions published with the package `nzilbb.labbcat` (Fromont, 2021) based on code from Reidy (2013, 2016b). The code is published in Puggaard-Rode (2022).

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

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

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 above the lag-1 autocorrelation in a corresponding model with no correction.[7] Autocorrelation along the frequency domain in the AR(1)-corrected models is negligible and short-range (between 0.06–0.16 at lag-1). Note that all models display moderate negative autocorrelation at higher lags, which is stable across different values for ρ (between 0.14–0.39 at lag-8). This is demonstrated in Fig. 5, which shows autocorrelation plots for the model of /t/ releases from four different models with different parameter settings for ρ. Note that the models for both /k/ and /p/ show less autocorrelation (both positive and negative) than the model for /t/.

Another method for accounting for autocorrelated errors is the use of functional random intercepts, potentially with smoothing parameters set using splines based on functional principal components (Scheipl et al., 2015; Greven & Scheipl, 2017a; Bauer et al., 2018). 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 AR(1)-correction is necessarily somewhat ad hoc. The latter approach can also be implemented in `pffr`, but at a significant computational cost. It is less computationally heavy to account for autocorrelation using another spline basis, such as thin plate regression splines. Recall that these are used to model variation along the time domain; Scheipl et al. (2015: appendix C.1) show that this approach can be used to simultaneously model a predictor variable and minimize autocorrelation along a functional domain (in this case the time domain). They further suggest that remaining residual structure can be accounted for by including scalar random intercepts, in this case corresponding to an intercept for each individual stop token. These are not included in the final models here, but I show in the accompanying code that adding such a scalar random intercept adds a significant computational load while accounting for very little variance in this case (Puggaard-Rode, 2022).

The model further includes 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. These are contrast coded (see Schad et al., 2020), such that absence of the feature in question is numerically coded as $-½$ and presence as +½; the speaker sex variable was (randomly) coded as $-½$ female, +½ male. Contrast coding of 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 effects, by-speaker random slopes were also included (except for speaker sex, which logically cannot vary by-speaker).[8] Including interaction effects (e.g. backness × rounding) would be possible, but I have opted against doing so here, since it would make the results unnecessarily complicated for an exploratory study such as this one. The model formulae can be expressed as follows, where all variables are standardized (or pseudo-normalized via contrast-coding):

$$\begin{aligned}\mathrm{amplitude}_{ij}(\mathrm{F}) = {} & \alpha(\mathrm{F}) + \gamma(\mathrm{t}_{ij},\ \mathrm{F}) + \mathrm{sex}(\mathrm{t}_{i},\ \mathrm{F}) + \mathrm{stress}(\mathrm{t}_{ij},\ \mathrm{F}) \\ & + \mathrm{height}(\mathrm{t}_{ij},\ \mathrm{F}) + \mathrm{backness}(\mathrm{t}_{ij}, \mathrm{F}) \\ & + \mathrm{roundness}(\mathrm{t}_{ij},\ \mathrm{F}) + \mathrm{speaker}_j\ \gamma(\mathrm{t}_{ij},\ \mathrm{F}) \\ & + \mathrm{speaker}_j\ \mathrm{stress}(\mathrm{t}_{ij}\ \mathrm{F}) + \mathrm{speaker}_j\ \mathrm{height}(\mathrm{t}_{ij},\ \mathrm{F}) \\ & + \mathrm{speaker}_j\ \mathrm{backness}(\mathrm{t}_{ij},\ \mathrm{F}) \\ & + \mathrm{speaker}_j\ \mathrm{roundness}(\mathrm{t}_{ij},\ \mathrm{F}) + \rho e_{i-1} + E_{ij}(\mathrm{F})\end{aligned}$$

where *i* indexes each observation, *j* indexes each speaker, *F* is frequency, *t* is time, α(F) denotes the smooth functional intercept, $\gamma(\mathrm{t}_{ij},\ \mathrm{F})$ denotes the non-linear time variable, and E_{ij}(F) denotes the functional residual error. The individual error of an observation e_i is further modulated by the error of the preceding observation e_{i-1} by a factor of ρ; this is the AR(1) process described above (Baayen et al., 2018). In Appendix A, I describe how a simplified model is fitted with `pffr`, and compare this to how a similar model would be fitted as a GAMM with `mgcv`.

As discussed in Section 5.4, I do not report *p*-values for the FOSR models, as they likely reflect the very large number of observations (almost 550,000 normalized amplitude values for the largest model) rather than practical significance. I do report the rest of the model summary, which is similar to summaries of GAMMs fitted with `mgcv`. I do not include random effects in the summaries, but they are included in the accompanying code (Puggaard-Rode, 2022). Model summaries include *estimated degrees of freedom* (*edf*), which reflect the linearity of a variable, with a low *edf* near 1 indicating that the variable is near-linear; *reference degrees of freedom* (*ref. df*), which reflect the complexity of fitting a variable; and *F*-values, which reflect how much variation in the data is accounted for by a variable. As such, *ref.df* and *F* combined reflect the fitting–complexity tradeoff of including a variable, and these are usually used to calculate *p*-values, following the procedure described by Wood (2013; 2017: ch. 6.12). As this study is largely exploratory, I take *F*-values as a proxy for the influence of individual variables, and do not otherwise touch on statistical significance in the traditional sense. This may not be satisfactory for all kinds of studies, and I return to the issue of null-hypothesis significance testing in Section 4.

I primarily 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 (including variation along the time domain) are kept at zero. These are not very telling

[7] Autocorrelation remained somewhat too high when ρ corresponded exactly to lag-1, which is why ρ is increased here.

[8] 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 sparsely distributed data.

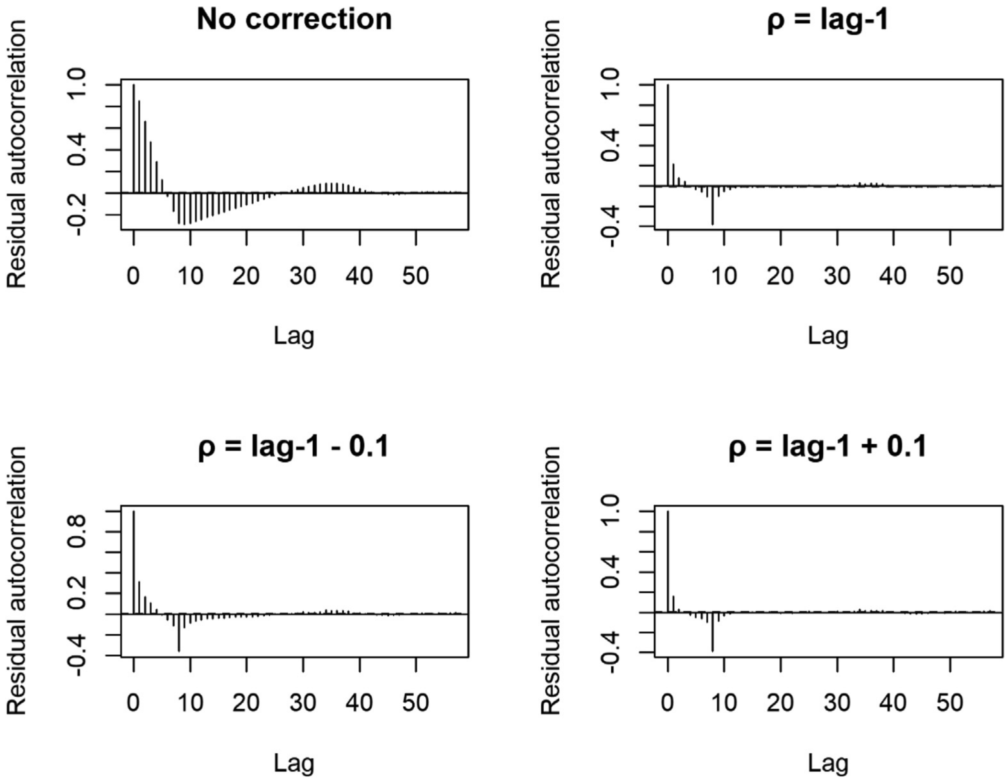


Fig. 5. Autocorrelation along the frequency domain in models of /t/ with no correction (top left), AR(1) model with ρ = lag-1 (top right), AR(1) model with ρ = 0.1 below lag-1 (bottom left), AR(1) model with ρ = 0.1 above lag-1 (bottom right).

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 variation in 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, with darker shading indicating higher energy. 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 require separate plots for the upper and lower limits. Plotted 95% confidence intervals for the multidimensional variables are included in Appendix B. These plots demonstrate the uncertainty associated with each fitted effect, and I will refer to these plots throughout the paper.

3.2. Results

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

3.2.1. /t/

The model of /t/ releases has a high effect size of R^2 = 0.53. The functional intercept (see Fig. 6) shows an energy peak around 3500–5000 Hz, with comparatively little energy elsewhere, particularly above 8000 Hz. Since all the binary variables in the statistical model are contrast coded, the intercept reflects a grand weighted mean across time with all contextual variables kept at zero. Since the intercept summarizes a dynamic series of events, it is not in itself very meaningful. In the spectro-temporal fits (Figs. 7–8), 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 Fig. 7. Initially, energy is skewed towards the higher end of the spectrum, with a fairly strong energy peak in the 3500–5500 Hz range (as in the intercept spectrum, Fig. 6), but also reasonably equal distribution of energy in the 5500–8000 Hz range. Increased energy above the main peak gradually tapers off, and in the final three-fourths of releases, energy is broadly distributed below 5000 Hz, including at the lowest frequencies visualized (500 Hz); this is comparable to the concrete example of a /t/ release shown in Fig. 3, which shows a similar development over time. The main effect of time is quite robust, as visible from 95% confidence intervals (see Appendix B.1).

Spectro-temporal fits for each direction of the individual variables are shown in Fig. 8. Table 2 shows the model summary. Fig. 8 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 relative to high vowels, the figures show a relatively early increase in energy at low

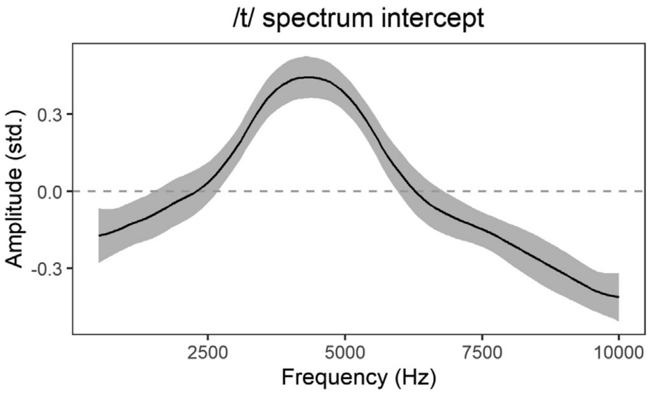


Fig. 6. Intercept spectrum for /t/.

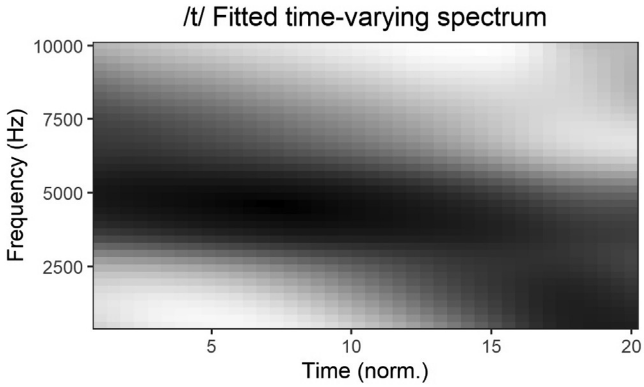


Fig. 7. Fitted time-varying spectrum of /t/ (main effect of time).

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)[9] begin dominating relatively early. Women show strong initial energy in frequencies above 5000 Hz, and although lower frequencies come into play late in the release, frequencies up to 5000 Hz are excited throughout the release. The effect of *sex* is robust (see Appendix B.1) and associated with a large *F*-score.

Lower frequencies start dominating towards the end of the release in unstressed syllables, and much earlier in stressed syllables; this core effect is quite stable, but there is significant uncertainty associated with the size of the effect (see Appendix B.1). Lower frequencies also dominate relatively late before *high* vowels, and frequencies above 6000 Hz are also more excited at the beginning of the release in this context. This is a strong effect associated with a high *F*-score. Again, these two effects are stable, but the size of the effects is relatively uncertain, and there is significant uncertainty associated with how vowel height otherwise affects the release. 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 a particularly low *F*-score for the *roundness* variable, and they are both associated with significant uncertainty.

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 3500–5000 Hz in the middle of the release is a consistent feature across all variables.

3.2.2. /k/

The model of /k/ releases has a high effect size of $R^2 = 0.56$. Recall that figures here do not extend above 8000 Hz. The intercept spectrum (see Fig. 9) shows almost evenly distributed energy below 4000 Hz, with small peaks around 500 Hz and just below 4000 Hz, and overall decreasing energy above 4000 Hz.

There is no strong main effect of time; there is little variance across the time domain in Fig. 10, and the variance that we do see is associated with significant uncertainty (see Appendix B.2). The associated *F*-score is also relatively small. Spectro-temporal fits for each direction of the individual variables are shown in Fig. 11. Table 3 shows the model summary.

There is a noticeable *sex* difference. There is little energy at lower frequencies during the first half of the release for female speakers, and more activity at frequencies above 4000 Hz. Lower frequencies become dominant in the last quarter of the release for female speakers, whereas for male speakers, they are seemingly dominant throughout the release. The *F*-score for this variable is high, and the patterns are quite robust (see Appendix B.2).

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 500–1000 Hz, while unstressed syllables have more energy at the higher band around 3500–4000 Hz, although lower frequencies gradually become dominant in the latter half of the release. Note, however, that the associated *F*-score is modest, and the variable is associated with significant uncertainty (see Appendix B.2).

Before *high* vowels, there is a lot of high frequency energy between 3000–5000 Hz 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 *F*-score. Non-back vowels and non-round vowels show roughly the same patterns as high vowels, although with slightly varying temporal alignment. The *backness* variable in particular is associated with a very large *F*-score. High frequency noise lasts somewhat longer for non-round vowels than non-back vowels. The *F*-score associated with the *roundness* variable is also large. All effects associated with features of the following vowel are robust.

[9] 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.

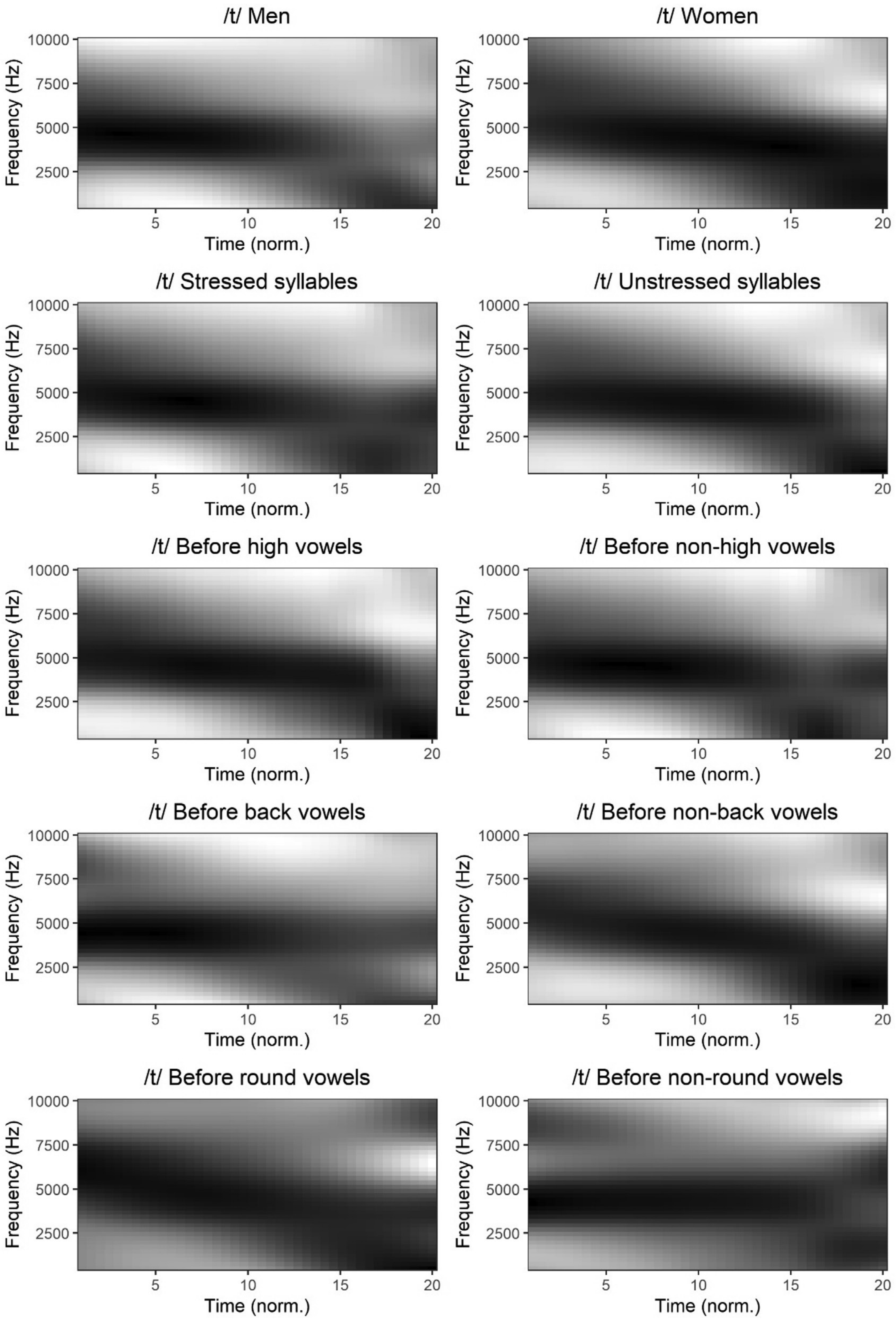


Fig. 8. Spectro-temporal fits of /t/ for each direction of the individual variables.

Table 2
Summary of /t/ model.

	edf	*ref.df*	*F*
Intercept	29.82	31	170.57
Time	49.02	57.82	18.51
Sex	31.45	39.16	40.16
Stress	41.2	52.08	12.37
Height	58.81	70.37	24.21
Backness	29.52	36.22	14.68
Roundness	24.26	29.56	8.95

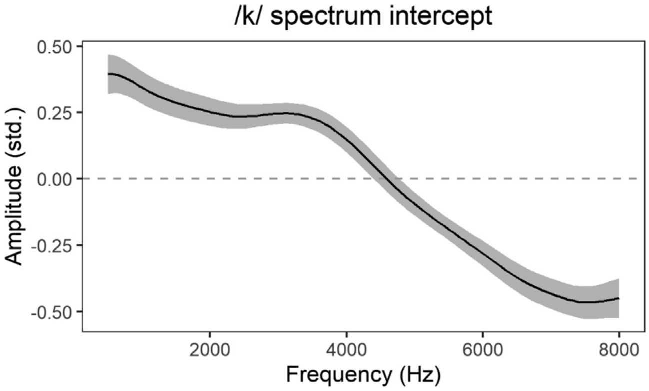


Fig. 9. Intercept spectrum for /k/.

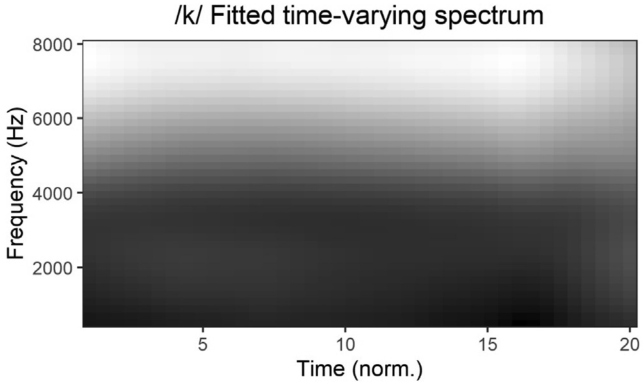


Fig. 10. Fitted time-varying spectrum of /k/ (main effect of time).

3.2.3. /p/

The model of /p/ releases has a very high effect size of R^2 = 0.7. The intercept spectrum (see Fig. 12) shows most energy in the lowest frequencies, with energy gradually reducing at higher frequencies. Assuming that the more diffuse distribution of noise towards the end of the release is not linguistically substantial, there is only a very marginal main effect of time (see Fig. 13), although what we do see is quite robust (see Appendix B.3).

Some of the by-variable time-varying characteristics of /p/ are clearer, as shown in Fig. 14. Table 4 shows the model summary.

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; the overall pattern is relatively robust, but the magnitude of the pattern is associated with significant uncertainty (see Appendix B.3). 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. 5000 Hz). The *stress* variable is quite strong and robust. A similar pattern is found before high vowels, with lower frequencies dominating relatively late in the release. The *height* variable has a relatively high *F*-score, and is also quite robust. To a lesser extent, the same pattern is found before non-back vowels. There is no obvious influence of round vowels, and this variable is associated with significant uncertainty (see Appendix B.3).

3.3. Linking the results to underlying articulatory mechanisms

In the preceding section, 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 2.2 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 partially determined by contextual factors like 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 2000 Hz. The results here, however, show two primary patterns of energy distribution: much higher resonance frequencies around 4000 Hz, or resonance frequencies centered around the lower end of the spectrum. I presume that

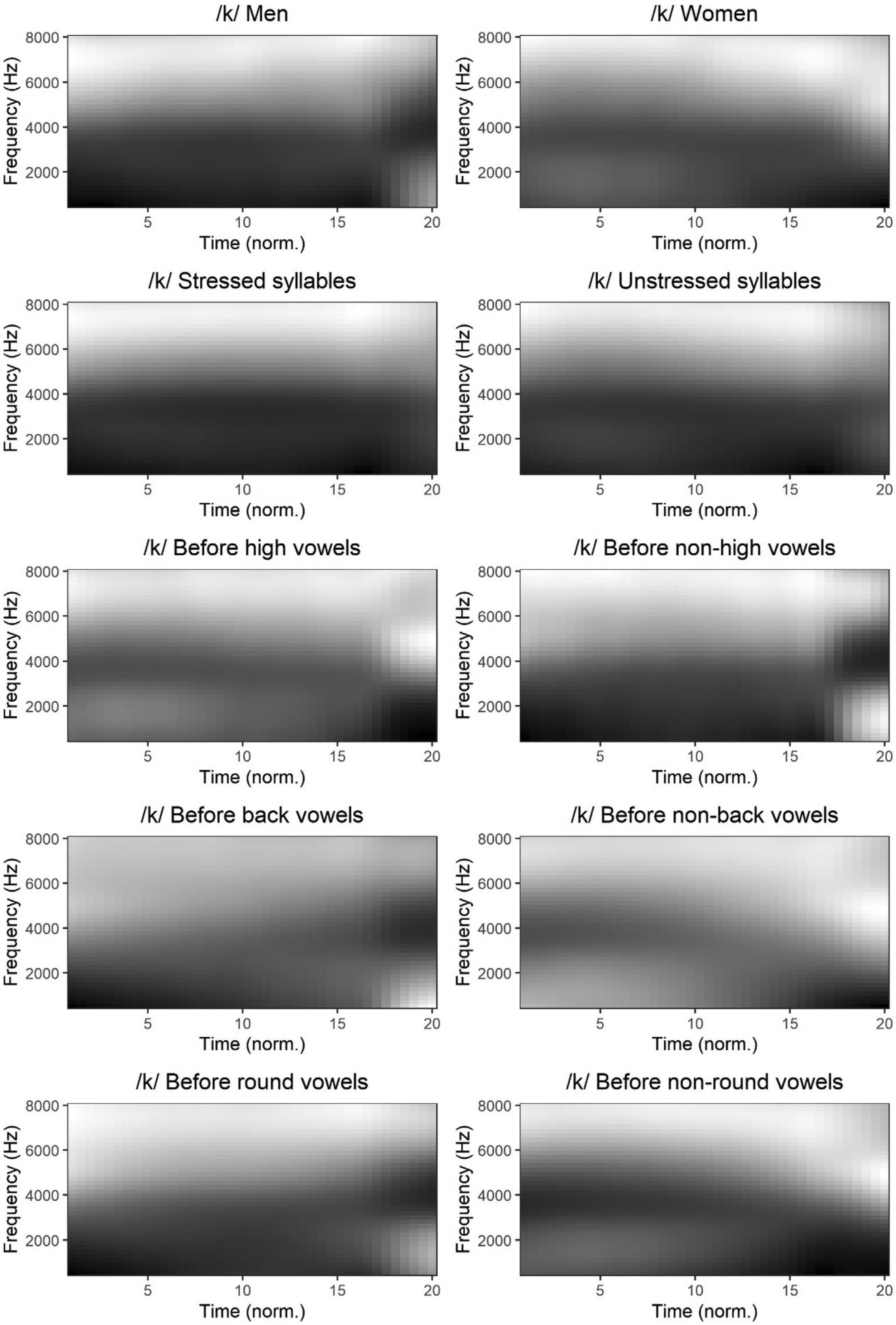


Fig. 11. Spectro-temporal fits of /k/ for each direction of the individual variables.

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

Table 3
Summary of /k/ model.

	edf	*ref.df*	*F*
Intercept	17.31	18	191.77
Time	48	57.97	13.8
Sex	42.41	51.38	59.74
Stress	18.88	24.31	5.75
Height	55.03	62.98	58.87
Backness	45.66	54.54	384.85
Roundness	21.23	26.19	77.41

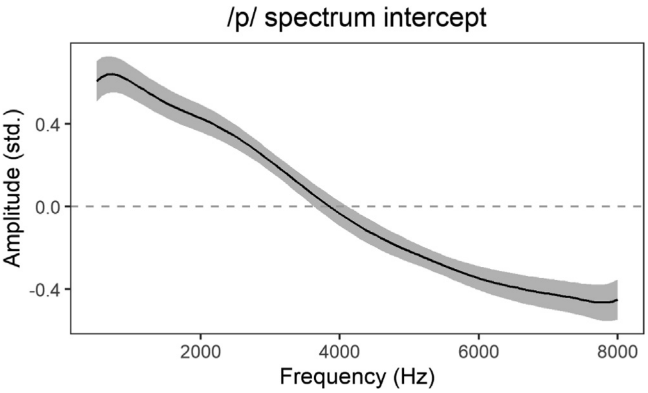


Fig. 12. Intercept spectrum for /p/.

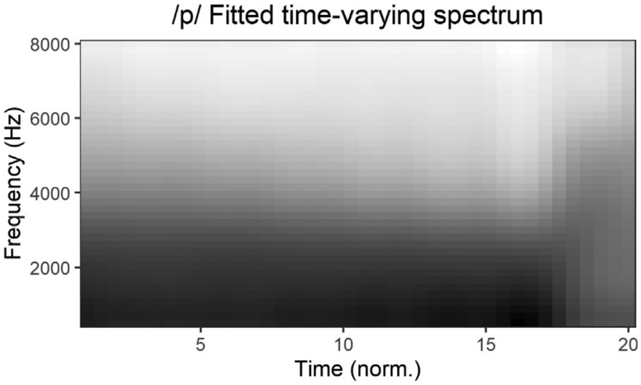


Fig. 13. Fitted time-varying spectrum of /p/ (main effect of time).

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 velar stops 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/.

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.

4. Discussion: Function-on-scalar regression and the spectrum

This paper 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 compute the uncertainty associated with each, 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, 2022).

FOSR modeling of spectra will generally involve highly multidimensional data, especially if the temporal dimension is also taken into account. 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 are provided in Appendix B and which I have referred to throughout. These are computationally efficient, and visualize the uncertainty associated with fitted

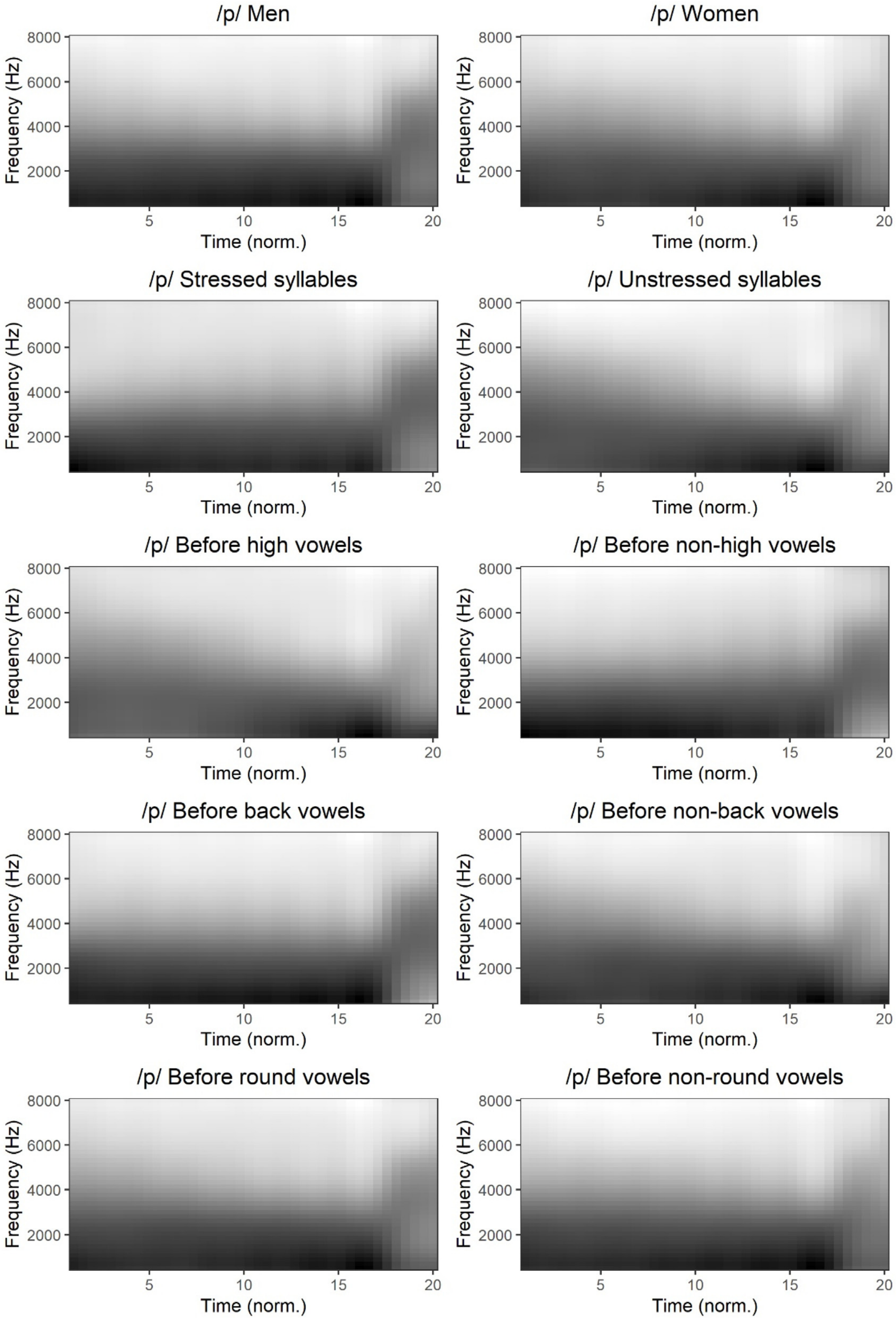


Fig. 14. Spectro-temporal fits of /p/ for each direction of the individual variables.

effects along the functional domain(s). 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 possi-

Table 4
Summary of /p/ model.

	edf	*ref.df*	*F*
Intercept	15.8	16	116.22
Time	68.86	72.91	35.1
Sex	27.21	34.96	8.37
Stress	17.85	22.28	43.62
Height	14.25	17.65	35.83
Backness	35.29	43.56	11.3
Roundness	13.08	16.9	2.57

ble to readily quantify the uncertainty related to the results (see e.g. Vasishth et al., 2018). 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 further 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 & 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 decibel scale for amplitude, and the equivalent rectangular bandwidth (ERB) scale, Bark scale, or mel scale for frequency.[10] Note that the latter two scales have been profitably used in combination with DCT in previous studies (Bukmaier & Harrington, 2016; Jannedy & Weirich, 2017).

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 4000 Hz would suddenly be excited immediately before the onset of voicing in a stop–vowel sequence, regardless of phonetic context. I can see three possible explanations for this: 1) the spectral characteristics of aspiration are highly variable, making it impossible for the model to make precise predictions, 2) the pseudo-centralization (contrast coding) of categorical variables sometimes causes the model to infer patterns that are not meaningful for one pole 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. Conclusion

This paper has introduced function-on-scalar regression as a method for analyzing speech spectra and how they vary over time. 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 already used to working with spectrograms. I showed how this tool can be fruitfully applied in the analysis of Danish stop releases, how they 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 affrication gradually turning to 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. As the precise point of occlusion in velar stops is known to be largely determined by the following vowel, these also have a great influence on release noise in /k/, particularly in the first portion of the release. The acoustic characteristics of /p/ releases show less prominent coarticulatory context effects, which mainly affect the first half of the release.

Availability of data and code

All code and data except actual sound files are available in the *DataverseNL* repository (Puggaard-Rode 2022). Sound files are available online, but are password protected (see Grønnum 2016). Praat scripts and annotated R code are also shared.

Acknowledgments

This paper has greatly benefitted from the expertise of three anonymous peer reviewers as well as associate editor Jeff Mielke. Bert Botma, Janet Grijzenhout, Sam Kirkham, and Ruben Schachtenhaufen all provided helpful comments on previous versions of the manuscript. I also thank the *Sounds of Language and Speech* audience for giving feedback on an earlier version of this study, and Morgan Sonderegger as well anonymous peer reviewers from a tangentially related paper submitted to *Journal of the International Phonetic Association* for pointing me in the direction of this approach to spectral analysis. Any remaining faults are my own.

[10] 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.

Appendix A. Fitting FOSR models with pffr and bam

In this appendix, I will show how a relatively simple FOSR model is fitted with `refund::pffr`, and then show how a corresponding GAMM could be fitted with `mgcv::bam`.

Assume that we want to fit a Gaussian FOSR model with spectra is the response variable, *time* as a dynamic variable, *stress* as a binary variable that varies dynamically over time, and by-speaker random slopes for both the main effect of time and the time-varying effect of stress. We smooth the data with P-splines over the frequency domain, and thin plate regression splines over the time domain, using 6 basis functions for the frequency domain, 16 basis functions for the time domain, and 20 basis functions for the functional intercept. We also include an AR(1) model with the ρ-parameter set at 0.8. This model can be formulated as.

$$\text{amplitude}_{ij}(\text{F}) = \alpha(\text{F}) + \gamma(\text{t}_{ij},\ \text{F}) + \text{stress}(\text{t}_{ij},\ \text{F}) + \text{speaker}_j\ \gamma(\text{t}_{ij},\ \text{F}) + \text{speaker}_j\ \text{stress}(\text{t}_{ij},\ \text{F}) + 0.8\,\text{e}_{\text{i}-1} + E_{ij}(\text{F})$$

This model is specified as follows:

```
refund::pffr(Y ~ s(timestep, k = 16) +
        s(speaker, timestep, bs="re") +
        s(timestep, k = 16, by = stress) +
        s(speaker, timestep, by = stress, bs="re"),
        data = df, ydata = y_df,
        bs.yindex = list(bs="ps", k = 6, m = c(2,1)),
        bs.int = list(bs="ps", k = 20, m = c(2,1)),
        rho = 0.8)
```

`pffr` includes both a scalar intercept and a functional intercept by default, so it only has to be explicitly specified if these should for some reason not be included. The response variable is always denoted `Y`. Non-linear variables are constructed with `s()`, which should be familiar from `mgcv`; also as in `mgcv`, they are smoothed with thin plate regression splines unless another spline basis is specified with the `bs`-parameter. The number of basis functions for non-linear variables are given with the `k`-parameter. Random effects are also constructed with `s()`, with the smoothing basis parameter set as `bs="re"`. The data have to be stored in separate data frames: one with information about the covariates and just one observation per function, called by `data`, and one with all observations along the functional domain, called by `ydata`. `bs.yindex` is used to tweak the smoothing parameters for the functional domain; the default is to use cubic P-splines `(bs="ps")` with 5 basis functions, and first order difference penalties; here, the number of basis functions is increased to 6.[11] `bs.int` is used to tweak the smoothing parameters for the global functional intercept; here, the default is 8 basis functions, which I increase to 20. `rho` sets the ρ-parameter to be used for an AR(1) model. `pffr` fits Gaussian models by default, but other link functions can be set with the `family`-parameter.

As mentioned in Section 1.3, `pffr` uses the `mgcv` computation engine, meaning that the same model can technically be fitted as a GAMM. The formula passed on by `pffr` for the above model is the following:

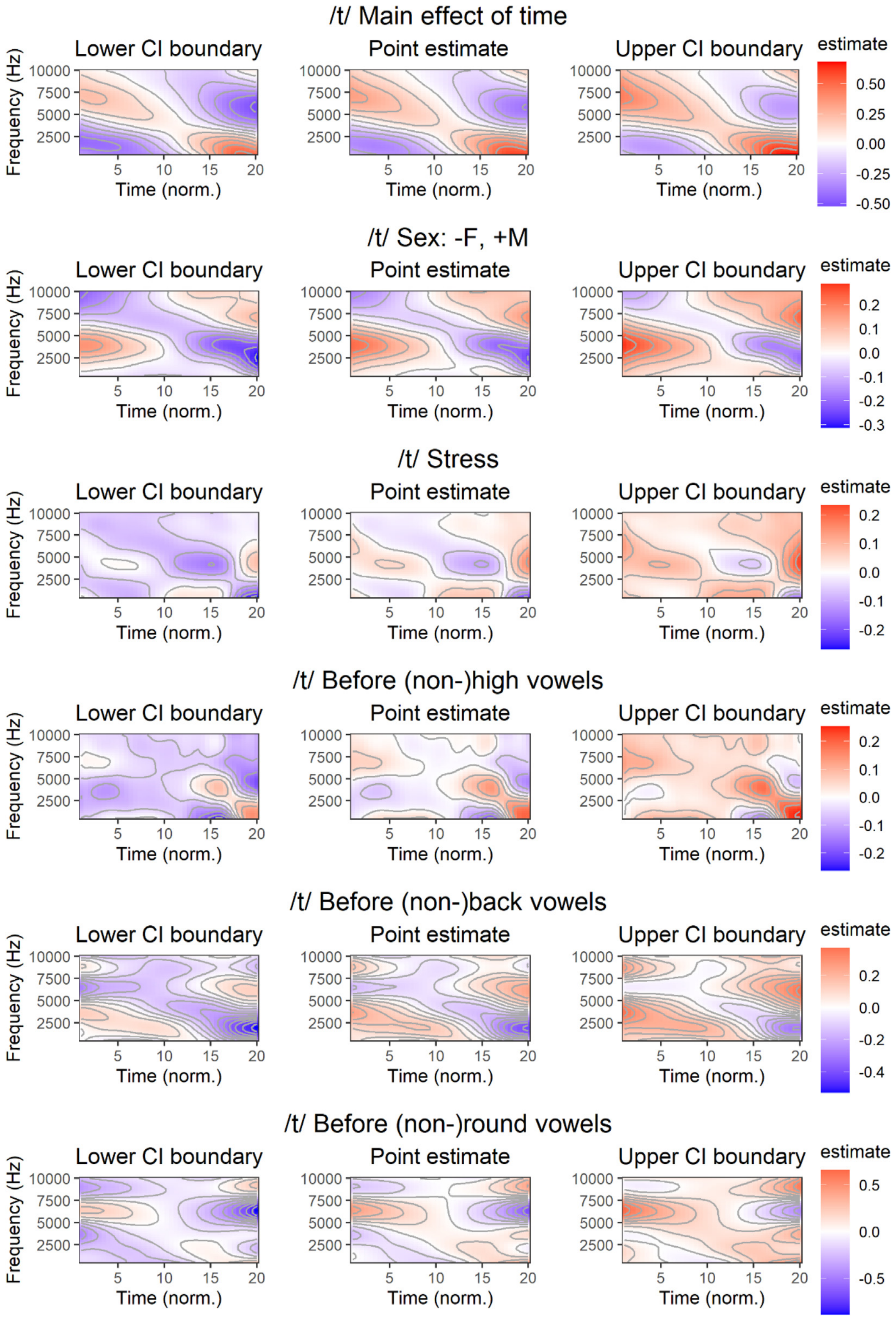
```
mgcv::bam(Y ~ s(x = yindex.vec, bs = "ps", k = 20,
  m = c(2, 1)) + ti(timestep, k = c(16, 6), bs = c("tp",
  "ps"), d = c(1, 1), yindex.vec, mc = c(TRUE, FALSE),
  m = c(2, 1)) + ti(speaker, timestep, bs = c("re", "ps"),
  d = c(2, 1), yindex.vec, mc = c(TRUE, FALSE), m = c(2, 1),
  k = c(25, 6)) + ti(timestep, k = c(16, 6), by = stress,
  d = c(1, 1), yindex.vec, mc = c(TRUE, FALSE), bs = c("tp",
  "ps"), m = c(2, 1)) + ti(speaker, timestep, by = stress,
  bs = c("re", "ps"), d = c(2, 1), yindex.vec,
  mc = c(TRUE, FALSE), m = c(2, 1), k = c(25, 6)),
  data = pffrdata, method = "fREML", chunk.size = 10000,
  rho = 0.8)
```

The obvious upshot is that the relatively simple `pffr`-formula expands to a very complex `bam`-formula. `pffr` constructs a new data frame `pffrdata`, including a variable `yindex.vec` which gives observations along the functional domain. The functional intercept is constructed with an `s()`-term using the parameter settings we gave in `bs.int` above, and the other non-linear variables are modeled with pure interactions, constructed with `ti()`-terms, each of which potentially employ multiple smoothing bases, order penalties, marginal centering constraints (the `mc`-parameter), and marginal basis dimensions (the `d`-parameter).

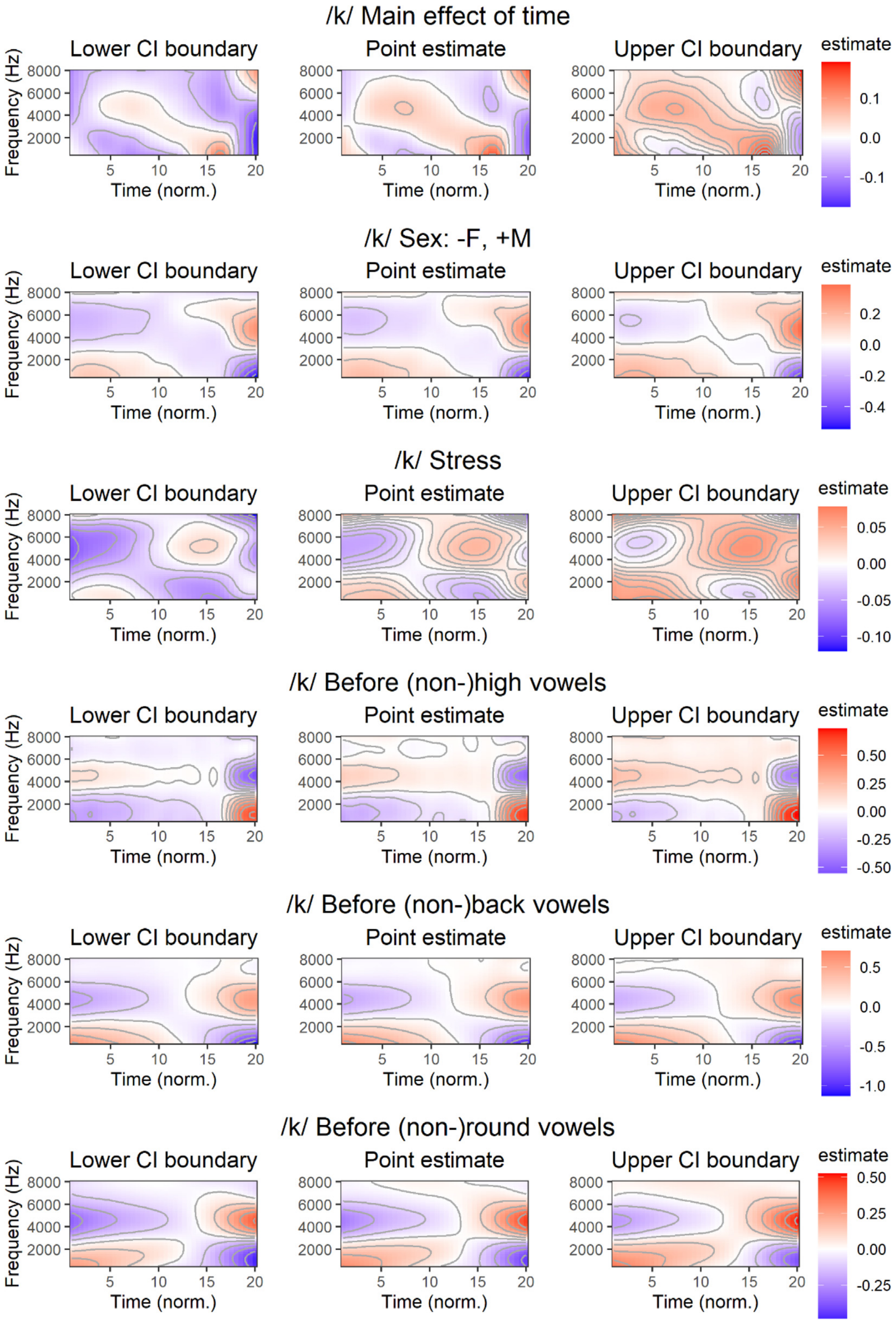
Appendix B. 95% confidence intervals for two-dimensional dynamic variables

This appendix provides plots with 95% confidence intervals for two-dimensional dynamic variables, following the method proposed by Marra and Wood (2012). Unlike the spectro-temporal fits shown in the article, these do not particularly look like spectrograms, and include both poles of binary variables in one plot. For the plot showing the main effect of time, red shading indicates a higher fitted value relative to the intercept, and blue shading indicates a lower fitted value relative to the intercept. For binary variables, red shading indicates a higher fitted value for the positive pole, and blue shading indicates a higher fitted value for the negative pole.

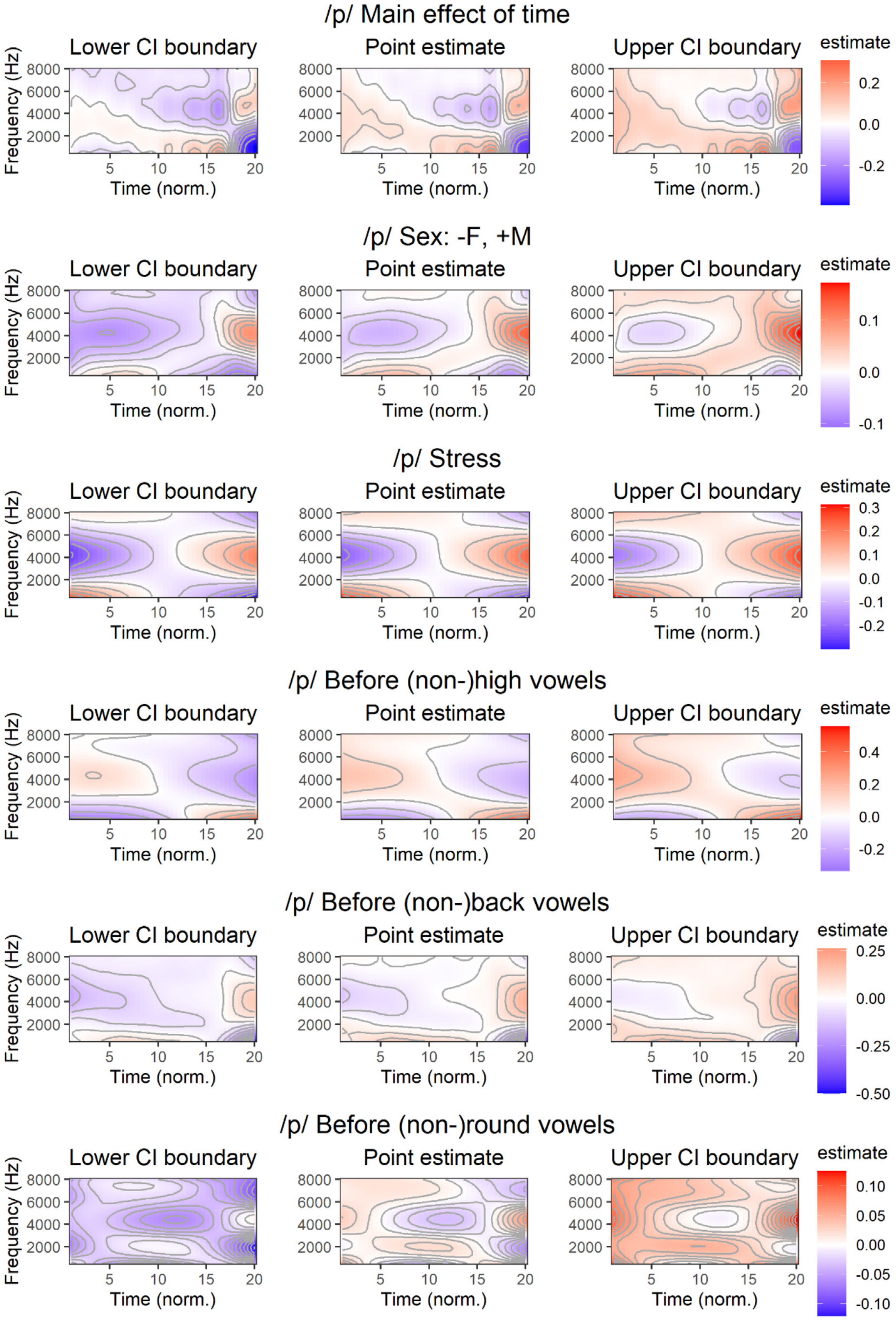
[11] The `m=c(2,1)` parameter specifies that the splines are cubic and use first order difference penalties.



Appendix B.1. 95% confidence intervals of two-dimensional variables in the model of /t/ releases. (19,28).



Appendix B.2. 95% confidence intervals of two-dimensional variables in the model of /k/ releases.



Appendix B.3. 95% confidence intervals of two-dimensional variables in the model of /p/ releases.

References

Baayen, R. H. (2008). *Analyzing linguistic data. A practical introduction to statistics using R*. Cambridge University Press. https://doi.org/10.1017/CBO9780511801686.

Baayen, R. H., van Rij, J., de Cat, C., & Wood, S. N. (2018). Autocorrelated errors in experimental data in the language sciences. Some solutions offered by generalized additive mixed models. In D. Speelman, K. Heylen, & D. Geeraerts (Eds.), *Mixed-effects regression models in linguistics* (pp. 49–69). Springer. https://doi.org/10.1007/978-3-319-69830-4_4.

Baayen, R. H., Vasishth, S., Kliegl, R., & Bates, D. (2017). The cave of shadows. Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language, 94*, 206–234. https://doi.org/10.1016/j.jml.2016.11.006.

Basbøll, H., & Wagner, J. (1985). *Kontrastive Phonologie des Deutschen und Dänischen. Segmentale Wortphhonologie und -phonetik*. Max Niemeyer. https://doi.org/10.1515/9783111358345.

Bauer, A. (2021). FoSIntro. Convenience functions for function-on-scalar regression. (R package version 1.0.3). https://github.com/bauer-alex/FoSIntro.

Bauer, A., Scheipl, F., Küchenhoff, H., & Gabriel, A.-A. (2018). An introduction to semiparametric function-on-scalar regression. *Statistical Modelling, 18*(3/4), 346–364. https://doi.org/10.1177/1471082X17748034.

Bauer, A., Scheipl, F., Küchenhoff, H., & Gabriel, A.-A. (2021). Registration for incomplete non-Gaussian functional data. (Unpublished manuscript.) https://arxiv.org/abs/2108.05634.

Biswas, M. (2022). *Hypothesis testing in function on scalar regression*. (PhD dissertation, North Carolina State University). https://lib.ncsu.edu/resolver/1840.20/39364.

Blacklock, O.S. (2004). *Characteristics of variation in production of normal and disordered fricatives, using reduced-variance spectral methods*. (PhD dissertation, University of Southampton). https://eprints.soton.ac.uk/420069.

Blumstein, S. E., & Stevens, K. N. (1979). Acoustic invariance in speech production. Evidence from measurements of the spectral characteristics of stop consonants. *Journal of the Acoustical Society of America, 66*(4), 1001–1017. https://doi.org/10.1121/1.383319.

Blumstein, S. E., & Stevens, K. N. (1980). Perceptual invariance and onset spectra for stop consonants in different vowel environments. *Journal of the Acoustical Society of America, 67*(2), 648–662. https://doi.org/10.1121/1.383890.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5*(9/10), 341–345 https://fon.hum.uva.nl/paul/papers/speakUnspeakPraat_glot2001.pdf.

Boersma, P., & Weenink, D. (2019). Praat. Doing phonetics by computer. (Version 6.0.5.5). http://fon.hum.uva.nl/praat.

de Boor, C. (2001). *A practical guide to splines* (2nd ed.). Springer.

Braunschweiler, N. (1997). Integrated cues of voicing and vowel length in German. A production study. *Language and Speech, 40*(4), 353–376. https://doi.org/10.1177/002383099704000403.

Brink, L., & Lund, J. (1975). *Dansk rigsmål. Lydudviklingen siden 1840 med særligt henblik på sociolekterne i København*. Gyldendal.

Bukmaier, V., & Harrington, J. (2016). The articulatory and acoustic characteristics of Polish sibilants and their consequences for diachronic change. *Journal of the International Phonetic Association, 46*(3), 311–329. https://doi.org/10.1017/S0025100316000062.

Bunnell, H. T., Polikoff, J., & McNicholas, J. (2004). Spectral moment vs. bark cepstral analysis of children's word-initial voiceless stops. In *8th international conference on spoken language processing*. https://doi.org/10.21437/Interspeech.2004-81.

Carignan, C., Hoole, P., Kunay, E., Pouplier, M., Joseph, A., Voit, D., ... Harrington, J. (2020). Analyzing speech in both time and space. Generalized additive mixed models can uncover systematic patterns of variation in vocal tract shape in real-time MRI. *Laboratory Phonology, 11*(2). https://doi.org/10.5334/labphon.214.

Cederbaum, J., Pouplier, M., Hoole, P., & Greven, S. (2016). Functional linear mixed models for irregularly or sparsely sampled data. *Statistical Modelling, 16*(1), 67–88. https://doi.org/10.1177/1471082X15617594.

Cho, T., & Ladefoged, P. (1999). Variation and universals in VOT. Evidence from 18 languages. *Journal of Phonetics, 27*(2), 207–229. https://doi.org/10.1006/jpho.1999.0094.

Chodroff, E., & Wilson, C. (2014). Burst spectrum as a cue for the stop voicing contrast in American English. *Journal of the Acoustical Society of America, 136*(5), 2762–2772. https://doi.org/10.1121/1.4896470.

Chodroff, E., & Wilson, C. (2018). Predictability of stop consonant phonetics across talkers. Between-category and within-category dependencies among cues for place and voice. *Linguistics Vanguard, 4*(s2). https://doi.org/10.1515/lingvan-2017-0047.

Cutler, A. (1976). Phoneme-monitoring reaction time as a function of preceding intonation contour. *Perception & Psychophysics, 20*(1), 55–60. https://doi.org/10.3758/BF03198706.

DSL = Det Danske Sprog- og Litteraturselskab. (2018). *Den danske ordbog*. https://ordnet.dk/ddo.

Fant, C. G. M. (1960). Acoustic theory of speech production with calculations based on X-ray studies of Russian articulations. *Mouton*. https://doi.org/10.1515/9783110873429.

Fischer-Jørgensen, E. (1954). Acoustic analysis of stop consonants. *Le Maître Phonetique, 32*(69), 42–59 https://jstor.org/stable/44705403.

Fischer-Jørgensen, E. (1969). Voicing, tenseness and aspiration in stop consonants, with special reference to French and Danish. *Annual Report of the Institute of Phonetics, University of Copenhagen, 3*, 63–114 https://tidsskrift.dk/ARIPUC/article/view/130755.

Fischer-Jørgensen, E. (1972). Tape cutting experiments with Danish stop consonants in initial position. *Annual Report of the Institute of Phonetics, University of Copenhagen, 6*, 104–168 https://tidsskrift.dk/ARIPUC/article/view/130910.

Fischer-Jørgensen, E., & Hutters, B. (1981). Aspirated stop consonants before low vowels. A problem of delimitation. *Annual Report of the Institute of Phonetics, University of Copenhagen, 15*, 77–102 https://tidsskrift.dk/ARIPUC/article/view/131752.

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

Francis, A. L., Ciocca, V., & Yu, J. M. C. (2003). Accuracy and variability of acoustic measures of voicing onset. *Journal of the Acoustical Society of America, 113*(2), 1025–1032. https://doi.org/10.1121/1.1536169.

Fromont, R. (2021). nzilbb.labbcat. Accessing data stored in "LaBB-CAT" instances. (R package version 1.0-1). https://CRAN.R-project.org/package=nzilbb.labbcat.

Frøkjær-Jensen, B., Ludvigsen, C., & Rischel, J. (1971). A glottographic study of some Danish consonants. In L. L. Hammerich, R. Jakobson, & E. Zwirner (Eds.), *Form and substance. Phonetic and linguistic papers presented to Eli Fischer-Jørgensen* (pp. 123–140). Akademisk Forlag.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press. https://doi.org/10.1017/CBO9780511790942.

Goldsmith, J., Scheipl, F., Huang, L., Wrobel, J., Di, C., Gellar, J., Harezlak, J., McLean, M.W., Swihart, B., Xiao, L., Crainicineau, C., & Reiss, P.T. (2021). refund. Regression with functional data. (R package version 0.1-24). https://CRAN.R-project.org/package=refund.

Gordon, M., Barthmaier, P., & Sands, K. (2002). A cross-linguistic acoustic study of voiceless fricatives. *Journal of the International Phonetic Association, 32*(2), 141–174. https://doi.org/10.1017/S0025100302001020.

Greven, S., & Scheipl, F. (2017a). A general framework for functional regression modelling. *Statistical Modelling, 17*(1/2), 1–35. https://doi.org/10.1177/1471082X16681317.

Greven, S., & Scheipl, F. (2017b). Rejoinder. *Statistical Modelling, 17*(1/2), 100–115. https://doi.org/10.1177/1471082X16689188.

Grønnum, N. (1995). Danish vowels. Surface contrast versus underlying form. *Phonetica, 52*(3), 215–220. https://doi.org/10.1159/000262173.

Grønnum, N. (1998). Illustrations of the IPA. Danish. *Journal of the International Phonetic Association, 28*(1/2), 99–105. https://doi.org/10.1017/S0025100300006290.

Grønnum, N. (2005). *Fonetik og fonologi. Almen og dansk* (3rd ed.). Akademisk Forlag.

Grønnum, N. (2009). A Danish phonetically annotated spontaneous speech corpus (DanPASS). *Speech Communication, 51*(7), 594–603. https://doi.org/10.1016/j.specom.2008.11.002.

Grønnum, N. (2016). DanPASS. *Danish Phonetically Annotated Speech* https://danpass.hum.ku.dk.

Gubian, M., Torreira, F., & Boves, L. (2015). Using Functional Data Analysis for investigating multidimensional dynamic phonetic contrasts. *Journal of Phonetics, 49*, 16–40. https://doi.org/10.1016/j.wocn.2014.10.001.

Guion, S. G. (1998). The role of perception in sound change of velar palatalization. *Phonetica, 55*(1/2), 18–52. https://doi.org/10.1159/000028423.

Hall, T. A., & Hamann, S. (2006). Towards a typology of stop assibilation. *Linguistics, 44*(6), 1195–1236. https://doi.org/10.1515/ling.2006.039.

Hall, T. A., Hamann, S., & Żygis, M. (2006). The phonetic motivation for phonological stop assibilation. *Journal of the International Phonetic Association, 36*(1), 59–81. https://doi.org/10.1017/S0025100306002453.

Hanson, H. M., & Stevens, K. N. (2003). Models of aspirated stops in English. In M.-J. Solé, D. Recasens, & J. Romeo (Eds.), *Proceedings of the 15th international congress of phonetic sciences* (pp. 783–786). https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2003/p15_0783.html.

Hock, H. H. (1991). *Principles of historical linguistics* (2nd ed.). Mouton de Gruyter. https://doi.org/10.1515/9783110219135.

Hoole, P., Pompino-Marschall, B., & Dames, M. (1984). Glottal timing in German voiceless occlusives. In M. P. R. Van den Broecke & A. Cohen (Eds.), *Proceedings of the tenth international congress of phonetic sciences* (pp. 399–403). Foris. https://doi.org/10.1515/9783110884685-059.

Hughes, G. W., & Halle, M. (1956). Spectral properties of fricative consonants. *Journal of the Acoustical Society of America, 28*(2), 303–310. https://doi.org/10.1121/1.1908271.

Iskarous, K., & Kavitskaya, D. (2018). Sound change and the structure of synchronic variability. Phonetic and phonological factors in Slavic palatalization. *Language, 94*(1), 43–83. https://doi.org/10.1353/lan.2018.0001.

Jaeger, J. J. (1983). The fortis/lenis question. Evidence from Zapotec and Jawoñ. *Journal of Phonetics, 11*(2), 177–189. https://doi.org/10.1016/S0095-4470(19)30814-9.

Jakobson, R., Fant, C. G. M., & Halle, M. (1951). *Preliminaries to speech analysis. The distinctive features and their correlates*. MIT Press.

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

Jespersen, O. (1899). *Fonetik. En systematisk fremstilling af læren om sproglyd*. Det Schubotheske Forlag.

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *Journal of the Acoustical Society of America, 108*(3), 1252–1263. https://doi.org/10.1121/1.1288413.

Juul, H., Pharao, N., & Thøgersen, J. (2016). Moderne danske vokaler. *Danske Talesprog, 16*, 35–72.

Kewley-Port, D. (1982). Measurements of formant transitions in naturally produced stop consonant–vowel syllables. *Journal of the Acoustical Society of America, 72*(2), 379–389. https://doi.org/10.1121/1.388081.

Kewley-Port, D. (1983). Time-varying features as correlates of place of articulation in stop consonants. *Journal of the Acoustical Society of America, 73*(1), 322–335. https://doi.org/10.1121/1.388813.

Kewley-Port, D., Pisoni, D. B., & Studdert-Kennedy, M. (1983). Perception of static and dynamic cues to place of articulation in initial stop consonants. *Journal of the Acoustical Society of America, 73*(5), 1779–1793. https://doi.org/10.1121/1.388813.

Kim, H. (2001). A phonetically based account of phonological stop assibilation. *Phonology, 18*(1), 81–108. https://doi.org/10.1017/S095267570100402X.

Knief, U., & Forstmeier, W. (2021). Violating the normality assumption may be the lesser of two evils. *Behavior Research Methods, 53*, 2576–2590. https://doi.org/10.3758/s13428-021-01587-5.

Koenig, L. L., Shadle, C. H., Preston, J. L., & Mooshammer, C. R. (2013). Toward improved spectral measures of /s/. Results from adolescents. *Journal of Speech, Language, and Hearing Research, 56*(4), 1175–1189. https://doi.org/10.1044/1092-4388(2012/12-0038).

Komsta, L., & Novomestky, F. (2015). moments. Moments, cumulants, skewness, kurtosis and related tests. (R package version 0.14). https://CRAN.R-project.org/package=moments.

Kopp, G. A., & Green, H. C. (1946). Basic phonetic principles of visible speech. *Journal of the Acoustical Society of America, 18*(1), 74–89. https://doi.org/10.1121/1.1916345.

Kühberger, A., Fritz, A., Lermer, E., & Scherndl, T. (2015). The significance fallacy in inferential statistics. *BMC Research Notes, 8*(84). https://doi.org/10.1186/s13104-015-1020-4.

Ligges, U. (2021). tuneR. Analysis of music and speech. (R package version 1.3.3.1). https://CRAN.R-project.org/package=tuneR.

Lin, Y.-H. (2011). Affricates. In M. van Oostendorp, C.J. Ewen, E.V. Hume, & K. Rice (eds.), *The Blackwell companion to phonology. Volume I: General issues and segmental phonology* (pp. 367–390). Wiley-Blackwell. https://doi.org/10.1002/9781444335262.wbctp0016.

Lisker, L. (1957). Closure duration and the intervocalic voiced–voiceless distinction in English. *Language, 33*(1), 42–49. https://doi.org/10.2307/410949.

Löfqvist, A. (1975). A study of subglottal pressure during the production of Swedish stops. *Journal of Phonetics, 3*(3), 175–189. https://doi.org/10.1016/S0095-4470(19)31366-X.

Löfqvist, A. (1976). Closure duration and aspiration for Swedish stops. *Phonetics Laboratory/Department of General Linguistics, Lund University Working Papers*, 13, 1–40. https://journals.lub.lu.se/LWPL/article/view/17004.

Löfqvist, A. (1980). Interarticulator programming in stop production. *Journal of Phonetics, 8*(4), 475–490. https://doi.org/10.1016/S0095-4470(19)31502-5.

Lombardi, L. (1990). The nonlinear organization of the affricate. *Natural Language and Linguistic Theory, 8*, 375–425. https://doi.org/10.1007/BF00135619.

Marra, G., & Wood, S. N. (2012). Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics, 39*(1), 53–74. https://doi.org/10.1111/j.1467-9469.2011.00760.x.

Marron, J. S., Ramsay, J. O., Sangalli, L. M., & Srivastava, A. (2015). Functional data analysis of amplitude and phase variation. *Statistical Science, 30*(4), 468–484. https://doi.org/10.1214/15-STS524.

Morris, J. S. (2017). Comparison and contrast of two general functional regression modelling frameworks. *Statistical Modelling, 17*(1–2), 59–85. https://doi.org/10.1177/1471082X16681875.

Mortensen, D. R. (2012). The emergence of obstruents after high vowels. *Diachronica, 29*(4), 434–470. https://doi.org/10.1075/dia.29.4.02mor.

Mortensen, J., & Tøndering, J. (2013). The effect of vowel height on voice onset time in stop consonants in CV sequences in spontaneous Danish. In *Proceedings of Fonetik 2013. The XXVIth annual phonetics meeting* (pp. 49–52). https://static-curis.ku.dk/portal/files/55957319/Mortensen_Tondering_Fonetik2013.pdf.

Mücke, D., Grice, M., & Cho, T. (2014). More than a magic moment. Paving the way for dynamics of articulation and prosodic structure. *Journal of Phonetics, 44*, 1–7. https://doi.org/10.1016/j.wocn.2014.03.001.

Nance, C., & Kirkham, S. (2020). The acoustics of three-way lateral and nasal palatalisation contrasts in Scottish Gaelic. *Journal of the Acoustical Society of America, 147*(4), 2858–2872. https://doi.org/10.1121/10.0000998.

Ohala, J. J. (1992). What's cognitive, what's not, in sound change. In G. Kellerman & M. D. Morrissey (Eds.), *Diachrony within synchrony. Language history and cognition* (pp. 309–355). Peter Lang. http://www.linguistics.berkeley.edu/~ohala/papers/what's_cognitive.pdf.

Ouni, S. (2014). Tongue control and its implication in pronunciation training. *Computer Assisted Language Learning, 27*(5), 439–453. https://doi.org/10.1080/09588221.2012.761637.

Pétursson, M. (1976). Aspiration et activité glottale. Examen expérimental à partir de consonnes islandaises. *Phonetica, 33*(3), 169–198. https://doi.org/10.1159/000259721.

Pharao, N. (2011). Plosive reduction at the group level and in the individual speaker. In *Proceedings of the 17th international congress of phonetic sciences* (pp. 1590–1593). https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2011.

Pharao, N., & Maegaard, M. (2017). On the influence of coronal sibilants and stops on the perception of social meanings in Copenhagen Danish. *Linguistics, 55*(5), 1141–1167. https://doi.org/10.1515/ling-2017-0023.

Plummer, A. R., & Reidy, P. F. (2018). Computing low-dimensional representations of speech from socio-auditory structures for phonetic analysis. *Journal of Phonetics, 71*, 355–375. https://doi.org/10.1016/j.wocn.2018.09.008.

Pouplier, M., Cederbaum, J., Hoole, P., Marin, S., & Greven, S. (2017). Mixed modeling for irregularly sampled and correlated functional data. Speech science applications. *Journal of the Acoustical Society of America, 142*(2), 935–946. https://doi.org/10.1121/1.4998555.

Pouplier, M., Hoole, P., Cederbaum, J., Greven, S., & Pastätter, M. (2014). Perceptual and articulatory factors in German fricative assimilation. In S. Fuchs, M. Grice, A. Hermes, L. Lancia, & D. Mücke (Eds.), *Proceedings of the 10th international seminar on speech production* (pp. 332–335). https://www.phonetik.uni-muenchen.de/~hoole/pdf/Poupliereta l_issp2014.pdf.

Puggaard, R. (2020). The productive acquisition of dental obstruents by Danish learners of Chinese. In D. Chen & D. Bell (eds.), *Explorations of Chinese theoretical and applied linguistics* (pp. 168–195). Cambridge Scholars. https://hdl.handle.net/1887/3146417.

Puggaard, R. (2021). Modeling regional variation in voice onset time of Jutlandic varieties of Danish. In H. Van de Velde, N. H. Hilton, & R. Knooihuizen (Eds.), *Language variation. European perspectives VIII* (pp. 79–110). John Benjamins. https://doi.org/10.1075/silv.25.04pug.

Puggaard-Rode, R. (Forthcoming). *Stop! Hey, what's that sound? The representation and realization of Danish stops*. (Submitted PhD dissertation, Leiden University).

Puggaard-Rode, R. (2022). Replication data for: Analyzing time-varying spectral characteristics of speech with function-on-scalar regression. *DataverseNL*. https://doi.org/10.34894/DYJT4V.

Puggaard-Rode, R., Horslund, C. S., & Jørgensen, H. (2022). The rarity of intervocalic voicing of stops in Danish spontaneous speech. *Laboratory Phonology, 13*(11). https://doi.org/10.16995/labphon.6449.

Puggaard-Rode, R., Horslund, C. S., Jørgensen, H., & Vet, D. J. (2022). Replication data for: The rarity of intervocalic voicing of stops in Danish spontaneous speech. *DataverseNL*. https://doi.org/10.34894/OSGTR8.

Pya, N., & Wood, S. N. (2016). A note on basis dimension selection in generalized additive modelling. (Unpublished manuscript). https://arxiv.org/abs/1602.06696.

R Core Team. (2020). R. A language and environment for statistical computing. (Version 4.0.3.). https://www.R-project.org/.

Rahim, K. J. (2014). Ap*plications of multitaper spectral analysis to nonstationary data*. (PhD dissertation, Queen's University). https://hdl.handle.net/1974/12584.

Rahim, K. J., & Burr, W. S. (2020). multitaper. Spectral analysis tools using the multitaper method. (R package version 1.0-15). https://CRAN.R-project.org/package=multitaper.

Ramsay, J. O., Hooker, G., & Graves, S. (2009). *Functional data analysis with R and MATLAB*. Springer. https://doi.org/10.1007/2F978-0-387-98185-7.

Ramsay, J. O., & Silverman, B. W. (2005). *Functional data analysis* (2nd ed.). Springer. https://doi.org/10.1007/b98888.

Reidy, P. F. (2013). An introduction to random processes for the spectral analysis of speech data. *Ohio State University Working Papers in Linguistics*, 60, 67–116. https://linguistics.osu.edu/research/pubs/papers.

Reidy, P. F. (2015). A comparison of spectral estimation methods for the analysis of sibilant fricatives. *Journal of the Acoustical Society of America: Express Letters, 137*(4), 248–254. https://doi.org/10.1121/1.4915064.

Reidy, P. F. (2016a). Spectral dynamics of sibilant fricatives are contrastive and language specific. *Journal of the Acoustical Society of America, 140*(4), 2518–2529. https://doi.org/10.1121/1.4964510.

Reidy, P. F. (2016b). spectRum. Compute periodogram and multitaper spectral estimates of waveform data. (R convenience functions). https://github.com/patrickreidy/spectRum.

Reiss, P. T., Huang, L., & Mennes, M. (2010). Fast function-on-scalar regression with penalized basis expansions. *International Journal of Biostatistics, 6*(1), art. 28. https://doi.org/10.2202/1557-4679.1246.

van Rij, J., Vaci, N., Wurm, L. H., & Feldman, L. B. (2020a). Alternative quantitative methods in psycholinguistics. Implications for theory and design. In V. Pirrelli, I. Plag, & W. U. Dressler (Eds.), *Word knowledge and word usage. A cross-disciplinary guide to the mental lexicon* (pp. 83–126). Mouton de Gruyter. https://doi.org/10.1515/9783110440577-003.

van Rij, J., Wieling, M., Baayen, R. H. & van Rijn, H. (2020b). itsadug. Interpreting time series and autocorrelated data using GAMMs. (R package version 2.4). https://CRAN.R-project.org/package=itsadug.

Roettger, T. B. (2019). Researcher degrees of freedom in phonetic research. *Laboratory Phonology, 10*(1). https://doi.org/10.5334/labphon.147.

Romeo, R., Hazan, V., & Pettinato, M. (2013). Developmental and gender-related trends of intra-talker variability in consonant production. *Journal of the Acoustical Society of America, 134*(5), 3781–3792. https://doi.org/10.1121/1.4824160.

RStudio Team. (2021). RStudio. Integrated development environment for R. (Version 1.4.1717). http://rstudio.com.

Sagey, E. C. (1986). *The representation of features and relations in non-linear phonology.* (PhD dissertation, Massachusetts Institute of Technology). https://hdl.handle.net/1721.1/15106.

Sawashima, M. (1970). Glottal adjustments for English obstruents. *Status Report on Speech Research, 21*(22), 187–200 https://files.eric.ed.gov/fulltext/ED044679.pdf#page=189.

Schachtenhaufen, R. (2022). Ny dansk fonetik. Books-on-demand. https://udtaleordbog.dk/ipa.php.

Schad, D. J., Vasishth, S., Hohenstein, S., & Kliegl, R. (2020). How to capitalize on a priori contexts in linear (mixed) models. A tutorial. *Journal of Memory and Language, 110*. https://doi.org/10.1016/j.jml.2019.104038.

Scheipl, F., Gertheiss, J., & Greven, S. (2016). Generalized functional additive mixed models. *Electronic Journal of Statistics, 10*(1), 1455–1492. https://doi.org/10.1214/16-EJS1145.

Scheipl, F., Staicu, A.-M., & Greven, S. (2015). Functional additive mixed models. *Journal of Computational and Graphical Statistics, 24*(2), 477–501. https://doi.org/10.1080/10618600.2014.901914.

Shadle, C. H. (1991). The effect of geometry on source mechanisms of fricative consonants. *Journal of Phonetics, 19*(3–4), 409–424. https://doi.org/10.1016/S0095-4470(19)30332-8.

Shadle, C. H., & Mair, S. J. (1996). Quantifying spectral characteristics of fricatives. In *Fourth international conference on spoken language processing* (pp. 1521–1524). International Speech Communication Association. https://doi.org/10.1109/ICSLP.1996.607906.

Sóskuthy, M. (2017). Generalised additive mixed models for dynamic analysis in linguistics. A practical introduction. (Unpublished article). https://arxiv.org/abs/1703.05339.

Sóskuthy, M. (2021). Evaluating generalised additive mixed modelling strategies for dynamic speech analysis. *Journal of Phonetics, 84*. https://doi.org/10.1016/j.wocn.2020.101017.

Spinu, L., & Lilley, J. (2016). A comparison of cepstral coefficients and spectral moments in the classification of Romanian fricatives. *Journal of Phonetics, 57*, 40–58. https://doi.org/10.1016/j.wocn.2016.05.002.

Stathopoulos, E. T., & Weismer, G. (1983). Closure duration of stop consonants. *Journal of Phonetics, 11*(4), 395–400. https://doi.org/10.1016/S0095-4470(19)30838-1.

Stevens, K. N. (1971). Airflow and turbulence noise for fricative and stop consonants. Static considerations. *Journal of the Acoustical Society of America, 50*(4B), 1180–1191. https://doi.org/10.1121/1.1912751.

Stevens, K. N. (1993a). Modelling affricate consonants. *Speech Communication, 13*(1/2), 33–43. https://doi.org/10.1016/0167-6393(93)90057-R.

Stevens, K. N. (1993b). Models for the production and acoustics of stop consonants. *Speech Communication, 13*(3/4), 367–375. https://doi.org/10.1016/0167-6393(93)90035-J.

Stevens, K. N. (1998). *Acoustic phonetics*. MIT Press. https://doi.org/10.7551/mitpress/1072.001.0001.

Stevens, K. N., & Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America, 64*(5), 1358–1368. https://doi.org/10.1121/1.382102.

Stevens, K. N., Manuel, S. Y. & Matthies, M. (1999). Revisiting place of articulation measures for stop consonants. Implications for models of consonant production. In J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville & A. C. Bailey (eds.), *14th international congress of phonetic sciences* (pp. 1117–1120). https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS1999/p14_1117.html.

Stoel-Gammon, C., Williams, K. A., & Buder, E. (1994). Cross-language differences in phonological acquisition. Swedish and American /t/. *Phonetica, 51*(1–3), 146–158. https://doi.org/10.1159/000261966.

Strevens, P. (1960). Spectra of fricative noise in human speech. *Language and Speech, 3*(1), 32–49. https://doi.org/10.1177/002383096000300105.

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

Swerts, M. (1994). *Prosodic features of discourse units*. (PhD dissertation, Eindhoven University of Technology). https://doi.org/10.6100/IR411593.

Swerts, M., & Collier, R. (1992). On the controlled elicitation of spontaneous speech. *Speech Communication, 11*(4–5), 463–468. https://doi.org/10.1016/0167-6393(92)90052-9.

Terken, J. M. B. (1984). The distribution of pitch accents in instructions as a function of discourse structure. *Language and Speech, 27*(3), 269–289. https://doi.org/10.1177/002383098402700306.

Vasishth, S., Nicenboim, B., Beckman, M. E., Li, F., & Kong, E. J. (2018). Bayesian data analysis in the phonetic sciences. A tutorial introduction. *Journal of Phonetics, 71*, 147–161. https://doi.org/10.1016/j.wocn.2018.07.008.

Vestergaard, T. (1967). Initial and final consonant combinations in Danish monosyllables. *Studia Linguistica, 21*(1), 37–66. https://doi.org/10.1111/j.1467-9582.1967.tb00547.x.

Voeten, C. C. (2020). *The adoption of sound change. Synchronic and diachronic processing of regional variation in Dutch*. (PhD dissertation, Leiden University). https://doi.org/1887/137723.

Volkmann, A., Stöcker, A., Scheipl, F., & Greven, S. (2021). Multivariate functional additive mixed models. *Statistical Modelling*. https://doi.org/10.1177/1471082X211056158.

Watson, C. I., & Harrington, J. (1999). Acoustic evidence for dynamic formant trajectories in Australian English vowels. *Journal of the Acoustical Society of America, 106*(1), 458–468. https://doi.org/10.1121/1.427069.

Wickham, H. (2016). *ggplot2. Elegant graphics for data analysis*. Springer. https://doi.org/10.1007/978-0-387-98141-3.

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., & Dunnington, D. (2021). ggplot2. Create elegant data visualizations using the grammar of graphics. (R package version 3.3.5). https://CRAN.R-project.org/package=ggplot2.

Wieling, M. (2018). Analyzing dynamic phonetic data using generalized additive mixed modeling. A tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics, 70*, 86–116. https://doi.org/10.1016/j.wocn.2018.03.002.

Wieling, M., Montemagni, S., Nerbonne, J., & Baayen, R. H. (2014). Lexical differences between Tuscan dialects and Standard Italian. Accounting for geographic and sociodemographic variation using generalized additive mixed modeling. *Language, 90*(3), 669–692. https://doi.org/10.1353/lan.2014.0064.

Wieling, M., Nerbonne, J., & Baayen, R. H. (2011). Quantitative social dialectology. Explaining linguistic variation geographically and socially. *Plos One, 6*(9). https://doi.org/10.1371/journal.pone.0023613.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society B, 65*(1), 95–114. https://doi.org/10.1111/1467-9868.00374.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society B, 73*(1), 3–36. https://doi.org/10.1111/j.1467-9868.2010.00749.x.

Wood, S. N. (2013). On p-values for smooth components of an extended generalized additive model. *Biometrika, 100*(1), 221–228. https://doi.org/10.1093/biomet/ass048.

Wood, S. N. (2017a). *Generalized additive models. An introduction with R* (2nd ed.). CRC Press. https://doi.org/10.1201/9781315370279.

Wood, S. N. (2017b). P-splines with derivative based penalties and tensor product smoothing of unevenly distributed data. *Statistics and Computing, 27*, 985–989. https://doi.org/10.1007/s11222-016-9666-x.

Wood, S. N. (2021). mgcv. Mixed GAM computation vehicle with automatic smoothness estimation. (R package version 1.8-36). https://CRAN.R-project.org/package=mgcv.

Wood, S. N., Goude, Y., & Shaw, S. (2015). Geralized additive models for large data sets. *Journal of the Royal Statistical Society C, 64*(1), 139–155. https://doi.org/10.1111/rssc.12068.

Wood, S. N., Li, Z., Shaddick, G., & Augustin, N. H. (2017). Generalized additive models for gigadata. Modeling the U.K. Black Smoke network daily data. *Journal of the American Statistical Association, 112*(519), 1199–1210. https://doi.org/10.1080/01621459.2016.1195744.

Wood, S. N., Pya, N., & Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association, 111*(516), 1548–1563. https://doi.org/10.1080/01621459.2016.1180986.