entropy

MDPI

Article

Detrending the Waveforms of Steady-State Vowels

Marnix Van Soom * and Bart de Boer

Artificial Intelligence Laboratory, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium; bart.de.boer@ai.vub.ac.be

* Correspondence: marnix@ai.vub.ac.be

Received: 4 February 2020; Accepted: 11 March 2020; Published: 13 March 2020

Abstract: Steady-state vowels are vowels that are uttered with a momentarily fixed vocal tract configuration and with steady vibration of the vocal folds. In this steady-state, the vowel waveform appears as a quasi-periodic string of elementary units called pitch periods. Humans perceive this quasi-periodic regularity as a definite pitch. Likewise, so-called pitch-synchronous methods exploit this regularity by using the duration of the pitch periods as a natural time scale for their analysis. In this work, we present a simple pitch-synchronous method using a Bayesian approach for estimating formants that slightly generalizes the basic approach of modeling the pitch periods as a superposition of decaying sinusoids, one for each vowel formant, by explicitly taking into account the additional low-frequency content in the waveform which arises not from formants but rather from the glottal pulse. We model this low-frequency content in the time domain as a polynomial trend function that is added to the decaying sinusoids. The problem then reduces to a rather familiar one in macroeconomics: estimate the cycles (our decaying sinusoids) independently from the trend (our polynomial trend function); in other words, detrend the waveform of steady-state waveforms. We show how to do this efficiently.

Keywords: formant; steady-state; vowel; detrending; acoustic phonetics; source-filter theory; probability theory; uncertainty quantification; model averaging; nested sampling

Relation of This Work to the Conference Paper

We have already presented the main idea of this work in a preceding conference paper [1], albeit in a relatively obscure form. In this work we give an improved theory together with a more complete picture, by showing how that main idea, the detrending of steady-state vowel waveforms, can be derived heuristically from canonical source-filter theory in a simple way.

As this work examines only steady-state systems, we need only a quite limited set of concepts from acoustic phonetics to get by—which, in addition, are well defined by virtue of the assumed steady-state. To make this work self-contained, we introduce these concepts where needed, though more tersely compared to the conference paper. We refer the reader to the conference paper and the references given for more details.

1. Introduction

Formants are characteristic frequency components in human speech that are caused by resonances in the vocal tract (VT) during speech production and occur both in vowels and consonants (In the literature, the distinction between the physical resonance of the VT and the associated characteristic frequency in the resulting speech is often not made [2] (p. 179); as such, the term "formant" can mean both, and we will follow that custom here). In source-filter theory [3], the "standard model" of acoustic phonetics, the speech production process is modeled as a linear time-invariant system [4]. In a nutshell, the input to the system is the glottal source (i.e., the vibration of the vocal folds), the system's transfer

function describes the formants of the VT by assigning one conjugate pole pair to each formant, and the output is the speech signal. The speech signal is thus the result of filtering the glottal source signal with the VT formants.

The formants' bandwidth, frequency and relative intensity can be manipulated by us humans through changing the VT configuration (such as rounding the lips or closing the mouth) during speech. Measuring the formants of a given speech fragment is a routine preoccupation in the field of acoustic phonetics as formants can be said to carry basic information—above all to human listeners—about uttered phonemes [5], speaker sex, identity and physique [6], medical conditions [7], etc.

Accordingly, when formants are used as one of the pieces of information by a speech processing computer program trying to determine, say, the height of a speaker [8], it is desirable to acknowledge the uncertainty in the formant measurement, as this uncertainty propagates to the uncertainty about the speaker's height. While our contention that such uncertainty quantification is desirable stems mainly from a principled point of view [9], we argue that in critical cases such as forensic speaker identification [10], the ability to assign a degree of confidence to formant measurements—upon which further conclusions rest—is valuable, perhaps essential, and well worth the considerable extra computational effort required (As far as we know, while "there is a huge and increasing demand for [forensic speaker identification] expertise in courts" [10] (p. 255), uncertainty quantification for formant measurements is currently not in (widespread) use in forensics [10–12]. We are aware of several works on quantifying and discussing the nature of the variability and reliability of formant measurements that have been published quite recently [2,13–17]; this matter is discussed further in the conference paper under the umbrella term "the formant measuring problem"). In more routine circumstances one may simply take the error bars on the formant estimates as a practical measure of the computer program's trust in its own output. As with many things, the use for error bars or confidence intervals for formant measurements depends strongly on the application and available resources at hand.

The goal of this paper is a discussion of a simple pitch-synchronous linear model of steady-state vowels capable of quantifying the uncertainty of the formant measurements in a very straightforward way: by *inferring* them in the context of (Bayesian) probability theory [18]. The model, which works by effectively detrending the waveforms of steady-state vowels, is a generalization of previous work by others [19,20] and is similar in principle to [21]'s Bayesian method to infer the fundamental frequency. While the remainder of this Introduction sketches the background and rationale for the model in some detail, readers may wish to skip directly to Section 2 for its actual mathematical statement.

1.1. Background

Historically, formants and steady-state vowels (as opposed to vowels in general) are intrinsically connected because the concept of a formant was originally defined in terms of steady-state vowels—see Figure 1a. What we mean by steady-state vowels in this work is the steady-state portion (which may never be attained in some cases) of a vowel utterance. This is the time interval in which (a) the VT configuration can be taken to be approximately fixed, leading to essentially constant formants, and (b) the vocal folds are sustained in a reasonably steady vibration called the glottal cycle, during which the vocal folds open and close periodically. Because of (a) and (b) the vowel waveform appears as a quasi-periodic string of elementary units called pitch periods (We use "quasi-periodic" in the (colloquial) sense of Reference [22] (p. 75), i.e., designating a recurrent function of time for which the waveforms for successive periods are approximately the same. Examples are given in Figure 1a,c,d). In practice, steady-state vowels are identified simply by looking for such quasi-periodic strings, which typically consist of about 3 to 5 pitch periods [5]. Results from clinical trials [23] indicate that normal (non-pathological) voices are usually Type I [24], i.e., phonation results in nearly periodic sounds, which supports the notion that uttered vowels in normal speech typically reach a steady-state before "moving on" [25].

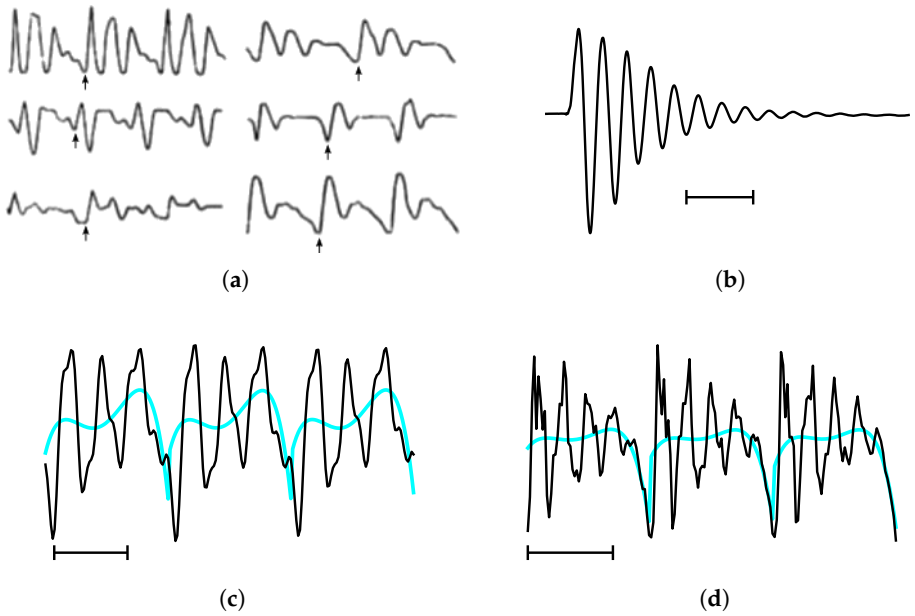


Figure 1. Several illustrations of pitch periods (a,c,d) and a related concept, the impulse response (b). The horizontal lines below the waveforms in (b–d) indicate a duration of 5 ms. The inferred trends in (c,d) are plotted in cyan. (**a**) In 1889, Hermann [26] used Fourier transforms of single pitch periods of steady-state vowels to calculate their spectra, and coined the term "formant" to designate the peak frequencies which were characteristic to the vowel [27] (p. ix). Shown here are examples of steady-state vowels from Hermann's work. Small vertical arrows indicate the start (glottal closing instant or GCI) of the second pitch period. Adapted from [27] (p. ix); originally from [28] (p. 40). (**b**) The impulse-like response produced when one of the authors excited his vocal tract by flicking his thumb against his larynx whilst mouthing "o". This is an old trick to emulate the impulse response of the vocal tract normally brought about by sharp GCIs (also known as glottal closures). The impulse responses triggered by sharp GCIs can be observed occasionally in the waveforms of vocal fry sections or as glottal stops /ʔ/ [27] (p. 49). (**c**) Three pitch periods taken from a synthesized steady-state instance of the vowel /ɤ/ at a fundamental frequency of 120 Hz and sampled at 8000 Hz. The trend inferred by our model is a fifth order polynomial. This example is discussed in Section 4.1. (**d**) Three pitch periods taken from a steady-state instance of the vowel /æ/ at a fundamental frequency of 138 Hz. The trend inferred by our model is a weighted combination of a 4th and 5th order polynomial. This example is discussed in Section 4.2. Source: [29], `bdl/arctic_a0017.wav`, 0.51–0.54 s, resampled to 8000 Hz.

During the steady-state, the pitch periods happen in sync with the glottal cycle [30]. The start of the pitch period coincides with the glottal *closing* instant (GCI), which causes a sudden agitation in the waveform and can be automatically detected quite reliably [31]. The duration between the GCIs, i.e., the length of the pitch periods, defines the fundamental frequency of the vowel, which for women is on the order of $(5\text{ ms})^{-1} = 200$ Hz and for men on the order of $(8\text{ ms})^{-1} = 125$ Hz [32,33]. The GCI pulse at the start of each pitch period is often so sharp that, recalling source-filter theory, the resulting response of the VT approximates the VT impulse response—see Figure 1b. In contrast, the glottal *opening* instant (GOI) excites the VT only weakly, which injects additional low-frequency content into the waveform roughly halfway through the pitch period.

1.2. The Pinson Model

The observation that the GCI pulse is often sufficiently sharp underlies Pinson's basic model [19] of the portion of the pitch period *between GCI and GOI* as being essentially the VT impulse response. The model was originally proposed to estimate formants during voiced speech directly in the time domain. The reason why the model does not address the whole pitch period is an additional

complication due to the GOI: when the glottis is open, the VT is coupled to the subglottal cavities (such as the lungs), which slightly increases the bandwidths of the formants, and thus slightly shifts the poles of the VT transfer function. If the "closed half" of the pitch period is characterized by Q formants with bandwidths $\boldsymbol{\alpha} = \{\alpha_1 \cdots \alpha_Q\}$ and frequencies $\boldsymbol{\omega} = \{\omega_1 \cdots \omega_Q\}$, then the VT transfer function H_P has Q conjugate pole pairs and up to $(2Q-1)$ zeros:

$$H_P(s, \boldsymbol{\alpha}, \boldsymbol{\omega}) = \frac{N(s)}{\prod_{k=1}^{Q}[s-(\alpha_k + i\omega_k)][s-(\alpha_k - i\omega_k)]} \qquad (\deg(N(s)) < 2Q), \tag{1}$$

where $N(s)$ is a polynomial of constrained degree in order that H_P be proper. (Note that, to emphasize the physical connection to resonances, we denote the bandwidth and frequency of the kth formant by (α_k, ω_k) and not by (B_k, F_k) as is more customary in acoustic phonetics.)

Pinson's model for the "closed half" of one pitch period is then just a bias term plus the associated impulse response of H_P (its inverse Laplace transform) constrained to live between GCI ($t = 0$) and GOI ($t = T_O$):

$$f_P(t, \boldsymbol{b}, \boldsymbol{\alpha}, \boldsymbol{\omega}) = b_1 + \sum_{k=1}^{Q} \left(b_{k+1} \cos \omega_k t + b_{k+1+Q} \sin \omega_k t\right) \exp\{-\alpha_k t\} \qquad (0 \le t < T_O). \tag{2}$$

Here the $(2Q+1)$ amplitudes $\boldsymbol{b} = \{b_1 \cdots b_{1+2Q}\}$ are determined by $N(s)$, $\boldsymbol{\alpha}$ and $\boldsymbol{\omega}$ through the partial fraction decomposition of Equation (1).

The model is then put to use by locating a steady-state vowel, choosing the central pitch period and sampling the "closed half" between GCI and GOI (which at the time had to be estimated by hand). Denoting the N samples by $\boldsymbol{d} = \{d_1 \cdots d_N\}$ sampled at times $\boldsymbol{t} = \{t_1 \cdots t_N\}$, estimates of the Q formants were found by weighted least-squares:

$$(\hat{\boldsymbol{b}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\omega}}) = \text{argmin}_{(\boldsymbol{b}, \boldsymbol{\alpha}, \boldsymbol{\omega})} \sum_{i=1}^{N} w_i^2 [d_i - f_P(t_i, \boldsymbol{b}, \boldsymbol{\alpha}, \boldsymbol{\omega})]^2.$$

Here w_i is an error-weighting function which deemphasizes samples close to GCI and GOI. The amplitude estimates $\hat{\boldsymbol{b}}$ can be used to reconstruct the fit to the data and determine $N(s)$ but are otherwise not used for the formant estimation.

Two features of Pinson's model are of interest here. The first one is that the bandwidth estimates obtained by this method seem to be (much) more reliable than those obtained by today's standard linear predictive coding (LPC) methods [13,14], when compared to bandwidths measured by independent methods [4,34,35]. The second one is the direct parametrization of the model function in Equation (2) by the formant bandwidths and frequencies $(\boldsymbol{\alpha}, \boldsymbol{\omega})$, which, as we explain in Section 3, transparently enables uncertainty quantification for their estimates $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\omega}})$ in a straightforward and transparent way—this is much harder in LPC-like methods.

1.3. The Proposed Model for a Single Pitch Period

The model for a single pitch period we propose in this paper is a simple generalization of Pinson's model in Equation (2) based on the empirical observation that the waveforms in pitch periods often seem to oscillate around a baseline or *trend*, which becomes more pronounced towards the end of the pitch period—see the examples in Figure 1c–d. In order to model this trend, we generalize the bias term in Equation (2) to an arbitrary polynomial of order $P-1$ and widen the scope of the model to the *full* pitch period of length T:

$$f(t, \boldsymbol{b}, \boldsymbol{\alpha}, \boldsymbol{\omega}) = \sum_{k=1}^{P} b_k t^{k-1} + \sum_{k=1}^{Q} \left(b_{k+P} \cos \omega_k t + b_{k+P+Q} \sin \omega_k t\right) \exp\{-\alpha_k t\} \qquad (0 \le t < T). \tag{3}$$

As before, the $\boldsymbol{b} = \{b_1 \cdots b_{P+2Q}\}$ are the model amplitudes, but now f is an implicit function of (P, Q), the *model orders*. P and Q will be subject to variation during model fitting, as opposed to the Pinson model where $P \equiv 1$ and Q was decided upon beforehand. The Bayesian approach in Section 3 avoids such a particular choice by using *model averaging* over all allowed values of (P, Q) to estimate the $2Q$ formant bandwidths and frequencies $(\boldsymbol{\alpha}, \boldsymbol{\omega})$.

As we will discuss in Section 2.4, the trend is caused by the weak excitation of the VT by the GOI. The trend is essentially a low-frequency byproduct of the glottal "open-close" cycle driving steady-state vowels (and is therefore absent at isolated glottal closure events, as in Figure 1b). The main innovation of the model is the assumption that the low-frequency content can be modeled adequately by superimposing a polynomial to the impulse response of the VT in the time domain. This reduces the problem of estimating formants to one frequently encountered in macroeconomics. This problem is the detrending of nonstationary time series such as business cycles for which one needs to estimate the cycles (in our case, parametrized by $\boldsymbol{\alpha}$ and $\boldsymbol{\omega}$) independently from the trend (in our case $\sum_{k=1}^{P} b_k t^{k-1}$) [36].

Since our proposed model for a single pitch period in Equation (3) is a generalization of Equation (2), it inherits the more reliable bandwidth estimation and the ability for straightforward uncertainty quantification from the Pinson model. In addition, the cumbersome labor of handpicking the "closed half" of the pitch period is eliminated by extending the scope of the model to full pitch periods, for which automatically estimated GCIs can be used. However, we pay for this convenience with a less precise model as we do not take into account the change in the formant bandwidths during the open part of the glottal cycle.

1.4. Outline

In Section 2 we state the model for a steady-state vowel, which is a chain of single pitch period models all sharing the same $(\boldsymbol{\alpha}, \boldsymbol{\omega})$ parameters, and discuss the origin of the trend.

In Section 3 we discuss how the formants and the uncertainty on their estimates are inferred using Bayesian model averaging and the nested sampling algorithm [37].

In Section 4 we apply the model to synthetic data and to real data.

Section 5 concludes.

2. A Pitch-Synchronous Linear Model for Steady-State Vowels

The model we propose is a simple variation of the standard linear model [38]:

$$\boldsymbol{d} = \boldsymbol{f}(\boldsymbol{t}, \boldsymbol{b}, \boldsymbol{\theta}) + \boldsymbol{e} = \boldsymbol{G}(\boldsymbol{t}, \boldsymbol{\theta})\, \boldsymbol{b} + \boldsymbol{e} \quad \text{where} \quad \boldsymbol{e} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}). \tag{4}$$

Here $\boldsymbol{d} = \{d_1 \cdots d_N\}$ is a vector holding the dataset of N points sampled at times $\boldsymbol{t} = \{t_1 \cdots t_N\}$, $\boldsymbol{f}$ is the model function, $\boldsymbol{G}$ is an $N \times m$ matrix holding the m basis functions which are function of $\boldsymbol{t}$ and the r "nonlinear" parameters $\boldsymbol{\theta}$, and $\boldsymbol{b} = \{b_1 \cdots b_m\}$ is a vector of m "linear" amplitudes. (The variables just listed describe the standard linear model in general terms; we connect these variables to our specific problem below in Section 2.1) The probabilistic aspect enters with our pdf for the vector of N errors $\boldsymbol{e} = \{e_1 \cdots e_N\}$ which is the classical separable multivariate Gaussian characterized by a single parameter, the noise amplitude σ (For more on the rationale behind assigning this pdf, see References [39,40] or, more concisely, Reference [41]). The noise power σ^2 may be also be expressed as the signal-to-noise ratio $\text{SNR} = 10 \log_{10} \boldsymbol{f}^T \boldsymbol{f} / N\sigma^2$.

It is well-known that for certain priors the simple form in Equation (4) allows for the marginalization over the amplitudes $\boldsymbol{b}$ and noise amplitude σ in the posterior distribution $p(\boldsymbol{b}, \boldsymbol{\theta}, \sigma | \boldsymbol{d}, I)$ (where $I \equiv$ our prior information), such that the posterior for the r nonlinear parameters $p(\boldsymbol{\theta} | \boldsymbol{d}, I)$ can be written in closed form.

The variation on the standard linear model just mentioned consists of promoting the dataset to a set of n dataset vectors, one for each pitch period in the steady-state vowel,

$$\boldsymbol{d} \to \{\boldsymbol{d_1} \cdots \boldsymbol{d_n}\}, \qquad \boldsymbol{t} \to \{\boldsymbol{t_1} \cdots \boldsymbol{t_n}\},$$

and we fit the model to each $\boldsymbol{d_i}$ simultaneously while keeping $\boldsymbol{\theta}$ and σ fixed but allowing each pitch period its own set of m amplitudes $\boldsymbol{b_i}$ and errors $\boldsymbol{e_i}$. Thus Equation (4) becomes the set of equations:

$$\boldsymbol{d_i} = \boldsymbol{f}(\boldsymbol{t_i}, \boldsymbol{b_i}, \boldsymbol{\theta}) + \boldsymbol{e_i} = \boldsymbol{G}(\boldsymbol{t_i}, \boldsymbol{\theta})\, \boldsymbol{b_i} + \boldsymbol{e_i} \quad \text{where} \quad \begin{cases} i = 1 \cdots n & \text{is the pitch period index} \\ \boldsymbol{e_i} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}) \end{cases} \tag{5}$$

The form of Equation (5) and our choice of priors below ensure that the marginalization over the $\{\boldsymbol{b_i}\}$ and σ is still possible. What remains is to specify the model function $\boldsymbol{f}$ and the priors for the $\{\boldsymbol{b_i}\}$, $\boldsymbol{\theta}$ and σ [42].

2.1. The Model Function

Given a steady-state vowel $\{\boldsymbol{d_1} \cdots \boldsymbol{d_n}\}$ segmented into n pitch periods, typically with the help of an automatic GCI detector. We assume that the data have been normalized and sampled at regular intervals and choose dimensionless units, such that the sampling times are $\boldsymbol{t_i} = \{0, 1, \ldots, N_i - 1\}$ where N_i is the length of the ith pitch period.

If we assume that the steady-state vowel is characterized by Q formants, the $r = 2Q$ nonlinear parameters of the model are the Q formant bandwidths $\boldsymbol{\alpha} = \{\alpha_1 \cdots \alpha_Q\}$ and Q formant frequencies $\boldsymbol{\omega} = \{\omega_1 \cdots \omega_Q\}$, i.e.,

$$\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\omega}) \equiv (\alpha_1, \ldots, \alpha_Q, \omega_1, \ldots, \omega_Q).$$

Further assuming that the trend in each pitch period can be modeled by a polynomial of degree $P - 1$ (which may differ in shape in each pitch period but always has that same degree), the model function for the ith pitch period has $m = P + 2Q$ basis functions and the same amount of amplitudes $\boldsymbol{b_i}$:

$$\boldsymbol{f}(\boldsymbol{t_i}, \boldsymbol{b_i}, \boldsymbol{\theta}) = \boldsymbol{G}(\boldsymbol{t_i}, \boldsymbol{\theta})\, \boldsymbol{b_i} \equiv \boldsymbol{G_i}\, \boldsymbol{b_i}, \tag{6}$$

where $\boldsymbol{G_i}$ is an $N_i \times m$ matrix holding the P polynomials and $2Q$ damped sinusoids:

$$[\boldsymbol{G_i}]_{jk} = \begin{cases} j^{k-1} & (1 \le k \le P) \\ \cos(j\omega_l) \exp\{-j\alpha_l\} & (P < k \le P + Q \quad \text{while} \quad l = 1 \cdots Q) \\ \sin(j\omega_l) \exp\{-j\alpha_l\} & (P + Q < k \le P + 2Q \quad \text{while} \quad l = 1 \cdots Q) \end{cases} \tag{7}$$

It is easy to verify that Equations (6) and (7) together are equivalent to the model for a single pitch period in Equation (3) we have motivated in Section 1.3. Thus, to summarize, our pitch-synchronous linear model for a steady-state vowel is essentially a chain of n linear single pitch period models, all constrained by (or rather, frustrated into) sharing the same parameters $\boldsymbol{\theta}$, which embodies our steady-state assumption that the formants do not change appreciably. The trend functions are only constrained by their degree, such that their shape may vary from period to period. In addition, the amplitudes of the damped sines may vary as well, which means the intensity and phase of the damped sines can vary from period to period.

2.2. The Priors

The prior pdfs for the model parameters are

$$\begin{aligned} p(\{\boldsymbol{b_1}\cdots\boldsymbol{b_n}\}|I) &= \prod_{i=1}^{n}(2\pi\delta^2)^{-m/2}\exp\left\{-\frac{\boldsymbol{b_i}^T\boldsymbol{b_i}}{2\delta^2}\right\} \\ &= (2\pi\delta^2)^{-nm/2}\exp\left\{-\frac{\sum_{i=1}^{n}\boldsymbol{b_i}^T\boldsymbol{b_i}}{2\delta^2}\right\} && (\boldsymbol{b_i}\in\mathbb{R}^m) \\ p(\boldsymbol{\theta}|I) &= \prod_{k=1}^{r}\left[\log\theta_k^{\mathrm{hi}}\Big/\theta_k^{\mathrm{lo}}\right]^{-1}\frac{1}{\theta_k} && (\theta_j^{\mathrm{lo}}<\theta_j<\theta_j^{\mathrm{hi}}) \\ p(\sigma|I) &= \left[\log\sigma^{\mathrm{hi}}\Big/\sigma^{\mathrm{lo}}\right]^{-1}\frac{1}{\sigma} && (\sigma^{\mathrm{lo}}<\sigma<\sigma^{\mathrm{hi}}) \end{aligned}$$

The pdfs are zero outside of the ranges indicated between the parentheses on the right.

In our implementation of the model we set $\delta = 1$ because (a) we normalize the data before fitting the model and (b) we use normalized Legendre functions instead of the computationally inconvenient polynomial functions in Equation (7) [43]. Thus this assignment for $p(\{\boldsymbol{b_i}\}|I)$ essentially expresses that we expect all amplitudes to be of order one and acts like a regularizer such that the amplitudes take on "reasonable" values [44].

The Jeffreys priors for the formant bandwidths and frequencies $\boldsymbol{\theta}$ stem from representation invariance arguments reflecting our a priori ignorance about their true values—see in App. A in [39]. The ranges $[\theta_j^{\mathrm{lo}}, \theta_j^{\mathrm{hi}}]$ in which the true values are supposed to lie must be decided by the user. Typically, the user can estimate the $\boldsymbol{\theta}$ directly with LPC. Additionally, based on the phoneme at hand, one can look up plausible ranges for the formant frequencies and bandwidths in the literature, e.g., References [4,45]. In our experiments in Section 4 we used both approaches to set the ranges for $\boldsymbol{\theta}$ (see Table 1 below) [46].

Likewise, the Jeffreys prior for the noise amplitude σ is bounded by σ^{lo} and σ^{hi}. For the lower bound we may take $\sigma^{\mathrm{lo}} \ll 1$ (e.g., on the order of the quantization noise amplitude) and for the upper bound we would conceivably have $\sigma^{\mathrm{hi}} \approx 1$ because of the normalization imposed on the data. As we will discuss below, the precise values of these bounds have a negligible influence on our final conclusions (i.e., model averaging over values of (P, Q) and obtaining weighted samples from the posterior of $\boldsymbol{\theta}$) on the condition that they may be taken sufficiently wide such that most of the mass of the integral below in Equation (20) is contained within them.

2.3. The Likelihood Function

The likelihood function is

$$\begin{aligned} L(\{\boldsymbol{b_1}\cdots\boldsymbol{b_n}\},\boldsymbol{\theta},\sigma) &= p(\{\boldsymbol{d_1}\cdots\boldsymbol{d_n}\}|\{\boldsymbol{b_1}\cdots\boldsymbol{b_n}\},\boldsymbol{\theta},\sigma,I) \\ &= (2\pi\sigma^2)^{-N/2}\exp\left\{-\frac{Q_F}{2\sigma^2}\right\}, \end{aligned} \tag{8}$$

where $N = \sum_{i=1}^{n} N_i$ and the scalar "least-squares" quadratic form

$$Q_F = \sum_{i=1}^{n}\boldsymbol{e_i}^T\boldsymbol{e_i} = \sum_{i=1}^{n}(\boldsymbol{d_i} - \boldsymbol{G_i}\,\boldsymbol{b_i})^T(\boldsymbol{d_i} - \boldsymbol{G_i}\,\boldsymbol{b_i}).$$

2.4. The Origin of the Trend

In this Section we present a heuristic derivation of the trend from source-filter theory.

According to the source-filter model of speech, the speech wave $y(t)$ is in general the output of a linear time-invariant system which models the VT with impulse response (transfer function) $h(t)$

($H(s)$) and the radiation characteristic with impulse response (transfer function) $r(t)$ ($R(s)$), and is driven by the glottal flow $u(t)$:

$$y(t) = u(t) * h(t) * r(t),$$

where $*$ denotes convolution.

We now proceed in the canonical way. As $R(s) \propto s$ for up to 4 kHz [4] (p. 128), and we work modulo rescaling, we can take [47]

$$y(t) \simeq u'(t) * h(t) \tag{9}$$

in this frequency range; accordingly, the glottal source may be taken to be the derivative of the glottal flow $u'(t)$. In this paper we will always work with signals resampled to a bandwidth of 4 kHz, which essentially limits us to the first three or four formants [48] (p. 20).

If we decompose the glottal cycle into a sharp delta-like excitation at GCI and a weak excitation at GOI, the glottal source during a pitch period may be written as

$$u'(t) \approx a_1 \delta(t) + l(t), \tag{10}$$

where a_1 is a constant, and $l(t)$ represents a slowly changing and broad function in which the spectral power $|L(\omega)|$ decreases quickly with increasing frequency ω. Substituting Equation (10) in Equation (9) gives

$$y(t) \simeq u'(t) * h(t) \approx (a_1 \delta(t) + l(t)) * h(t) = a_1 h(t) + l(t) * h(t),$$

where, as before, $y(t)$ represents the speech signal during one pitch period.

As $l(t)$ is a smooth and broad function, the magnitude of its Fourier transform $|L(\omega)|$ will be quite narrowly concentrated around $\omega \approx 0$. In contrast, $|H(i\omega)|$ will have its first peak only at $F_1 \gg 0$, the first formant, and therefore generally only a slowly rising slope near frequencies $\omega \approx 0$ if $H(s)$ can be represented as an all-pole transfer function (which is the assumption behind LPC [49]). Since

$$l(t) * h(t) = \mathcal{L}^{-1}[L(\omega)\, H(i\omega)]\,(t),$$

where $\mathcal{L}$ denotes the Fourier transform, if $|L(\omega)|$ falls off sufficiently fast, the slope of $|H(i\omega)|$ near $\omega \approx 0$ will be near constant, and we may write

$$L(\omega)\, H(i\omega) \approx a_2 L(\omega),$$

where $a_2 = H(0)$ is a real constant.

Thus we may write, very roughly, that the speech signal during one pitch period

$$y(t) \simeq u'(t) * h(t) \approx (a_1 \delta(t) + l(t)) * h(t) = a_1 h(t) + a_2 l(t).$$

As we have assumed that $l(t)$ is a smooth and broad function, it is reasonable to assume that it can be modeled as a polynomial. Thus $l(t)$ is our trend function, which modulo scaling essentially passes unscathed through convolution with $h(t)$ because of the absolute mismatch in their frequency content.

3. Inferring the Formant Bandwidths and Frequencies: Theory

The posterior distribution for the $(nm + r + 1)$ model parameters is

$$p(\{\boldsymbol{b_i}\}, \boldsymbol{\theta}, \sigma | \{\boldsymbol{d_i}\}, I) = L(\{\boldsymbol{b_i}\}, \boldsymbol{\theta}, \sigma)\; p(\{\boldsymbol{b_i}\}|I)\; p(\boldsymbol{\theta}|I)\; p(\sigma|I). \tag{11}$$

We use the nested sampling algorithm [37] to infer the parameters of interest, the formant bandwidths and frequencies $\boldsymbol{\theta}$ in the following way:

Once the data $\{\boldsymbol{d_i}\}$ are gathered, a "grid" of plausible model order values and prior ranges for the $\boldsymbol{\theta}$ are proposed. Each point (P, Q) on that grid parametrizes a particular model for the $\{\boldsymbol{d_i}\}$. The *evidence for a* (P, Q) *model*

$$\begin{aligned} Z(P,Q) &= p(\{\boldsymbol{d_i}\}|P,Q,I) \\ &= \int \mathrm{d}\boldsymbol{b_1} \cdots \mathrm{d}\boldsymbol{b_n} \int \mathrm{d}\boldsymbol{\theta} \int \mathrm{d}\sigma \, L(\{\boldsymbol{b_i}\},\boldsymbol{\theta},\sigma)\, p(\{\boldsymbol{b_i}\}|I)\, p(\boldsymbol{\theta}|I)\, p(\sigma|I), \end{aligned} \tag{12}$$

where the integrand is a function of the model orders implicitly, can be estimated with the nested sampling algorithm. A (highly desirable) byproduct of this estimation is the acquisition of a set of weighted samples from the posterior in Equation (11), from which the estimates and error bars of the formant bandwidths and frequencies can be calculated, as well as any other function of them (such as the VT transfer function). However, we show in Section 3.1 that by performing the integrals over $\{\boldsymbol{b_i}\}$ and σ Equation (12) can be written as

$$Z(P,Q) = \int \mathrm{d}\boldsymbol{\theta}\, L_I(P,Q,\boldsymbol{\theta})\, p(\boldsymbol{\theta}|I), \tag{13}$$

where L_I is the *integrated likelihood*. Thus we can sample our parameters of interest, the $\boldsymbol{\theta}$, directly from Equation (13) instead of Equation (12).

The uncertainty quantification for the formants is then accomplished through Bayesian model averaging. Using Equation (13), the evidence $Z(P, Q)$ is calculated and samples $\boldsymbol{\theta}_{P,Q}^{(l)}$ with weights $w_{P,Q}^{(l)}$ are gathered for all allowed (P, Q) values (all values on the grid). Then the formant bandwidth and frequency estimates are calculated from the first $(M = 1)$ and second $(M = 2)$ moments of the samples through model averaging over (P, Q) (though in practice only one to two values of (P, Q) dominate). Assuming uniform priors for the model orders (i.e., $p(P, Q|I) \propto 1$),

$$p(\boldsymbol{\theta}|\{\boldsymbol{d_i}\},I) = \frac{\sum_{P,Q} Z(P,Q)\, p(\boldsymbol{\theta}|\{\boldsymbol{d_i}\},P,Q,I)}{\sum_{P,Q} Z(P,Q)} \quad \text{so that} \quad \left\langle \boldsymbol{\theta}^M \right\rangle \approx \frac{\sum_{P,Q,l} Z(P,Q)\, w_{P,Q}^{(l)} \left[\boldsymbol{\theta}_{P,Q}^{(l)}\right]^M}{\sum_{P,Q,l} Z(P,Q)\, w_{P,Q}^{(l)}}. \tag{14}$$

Likewise, the posterior probabilities for the model orders considered jointly and separately are

$$p(P,Q|\{\boldsymbol{d_i}\},I) = \frac{Z(P,Q)}{\sum_{P,Q} Z(P,Q)}; \quad p(P|\{\boldsymbol{d_i}\},I) = \frac{\sum_{Q} Z(P,Q)}{\sum_{P,Q} Z(P,Q)}; \quad p(Q|\{\boldsymbol{d_i}\},I) = \frac{\sum_{P} Z(P,Q)}{\sum_{P,Q} Z(P,Q)}. \tag{15}$$

Finally, we note that sampling $\boldsymbol{\theta}$ from Equation (13) instead of Equation (12) reduces the dimensionality of the parameter space from

$$nm + r + 1 = n(P + 2Q) + 2Q + 1 \rightarrow r = 2Q,$$

which compares favorably to the increased cost of evaluating L_I compared to L. In a typical application, $n = 3$, $P = 5$ and $Q = 3$, such that the dimensionality is reduced from 40 to a mere 6 dimensions. The dimensionality of the problem does not depend on the number of pitch periods n (but its complexity does).

3.1. The Integrated Likelihood

We begin by writing Equation (12) as

$$\begin{aligned} Z(P,Q) = \int \mathrm{d}\boldsymbol{\theta} \int \mathrm{d}\sigma\, p(\boldsymbol{\theta}|I)\, p(\sigma|I) \\ \times \prod_{i=1}^{n} \int \mathrm{d}\boldsymbol{b_i}\, (2\pi\sigma^2)^{-N_i/2}\, (2\pi\delta^2)^{-m/2} \exp\left\{-\frac{Q_F^{(i)}}{2}\right\}, \end{aligned} \tag{16}$$

where the quadratic form for the ith pitch period

$$Q_F^{(i)} = (\boldsymbol{d_i} - \boldsymbol{G_i}\,\boldsymbol{b_i})^T\,\frac{\boldsymbol{I}}{\sigma^2}\,(\boldsymbol{d_i} - \boldsymbol{G_i}\,\boldsymbol{b_i}) + \frac{1}{\delta^2}\boldsymbol{b_i}^T\,\frac{\boldsymbol{I}}{\delta^2}\,\boldsymbol{b_i}.$$

The expression for the integrated likelihood L_I, defined implicitly by Equation (13), can be found from Equation (16) by marginalizing over the amplitudes $\{\boldsymbol{b_i}\}$ and the standard deviation σ.

We begin with the $\{\boldsymbol{b_i}\}$. Defining (see, e.g., App. A in [38])

$$\begin{aligned}
\boldsymbol{g_i} &= \boldsymbol{G_i}^T\boldsymbol{G_i} \\
\hat{\boldsymbol{b}}_{\boldsymbol{i}} &= \text{solution of } \left\{\frac{\partial}{\partial \boldsymbol{b_i}} Q_F^{(i)} = \boldsymbol{0}\right\} = \boldsymbol{g_i}^{-1}\,\boldsymbol{G_i}^T\,\boldsymbol{d_i} \\
\hat{\boldsymbol{f}}_{\boldsymbol{i}} &= \boldsymbol{f}(\boldsymbol{t_i}, \hat{\boldsymbol{b}}_{\boldsymbol{i}}, \boldsymbol{\theta}) = \boldsymbol{G_i}\,\hat{\boldsymbol{b}}_{\boldsymbol{i}}
\end{aligned}$$

it can be shown after some effort that (using, e.g., [50])

$$Q_F^{(i)} \simeq \left(\boldsymbol{b_i} - \hat{\boldsymbol{b}}_{\boldsymbol{i}}\right)^T\,\frac{\boldsymbol{g_i}}{\sigma^2}\left(\boldsymbol{b_i} - \hat{\boldsymbol{b}}_{\boldsymbol{i}}\right) + \frac{\boldsymbol{d_i}^T\boldsymbol{d_i}}{\sigma^2} - \frac{\hat{\boldsymbol{f}}_{\boldsymbol{i}}^{\,T}\hat{\boldsymbol{f}}_{\boldsymbol{i}}}{\sigma^2} + \frac{\hat{\boldsymbol{b}}_{\boldsymbol{i}}^{\,T}\hat{\boldsymbol{b}}_{\boldsymbol{i}}}{\delta^2}. \tag{17}$$

This is a good approximation if $\left(\boldsymbol{g_i}/\sigma^2 + \boldsymbol{I}/\delta^2 \approx \boldsymbol{g_i}/\sigma^2\right)$, i.e., if for all $j = 1 \cdots m$ it holds that

$$\left[\frac{\boldsymbol{g_i}}{\sigma^2}\right]_{jj} \gg \frac{1}{\delta^2} \quad\Leftrightarrow\quad \sum_{k=1}^{N_i}\left[\frac{\boldsymbol{G_i}}{\sigma^2}\right]_{kj}^2 \gg \frac{1}{\delta^2} \quad\Leftrightarrow\quad \frac{\text{integrated power of the } j\text{th basis function}}{\text{noise power}} \gg \frac{1}{\delta^2}.$$

In our implementation, where $\delta = 1$, we found this to be an acceptable approximation for all states $(\{\boldsymbol{b_i}\}, \boldsymbol{\theta}, \sigma)$ with an appreciable likelihood L.

It is interesting to note that when the least-squares measure

$$\chi_i^2(\boldsymbol{b_i}, \boldsymbol{\theta}) = \boldsymbol{e_i}^T\boldsymbol{e_i} = (\boldsymbol{d_i} - \boldsymbol{G_i}\,\boldsymbol{b_i})^T(\boldsymbol{d_i} - \boldsymbol{G_i}\,\boldsymbol{b_i})$$

is evaluated at the optimal amplitudes $\hat{\boldsymbol{b}}_{\boldsymbol{i}}$, it reduces to

$$\chi_i^2(\hat{\boldsymbol{b}}_{\boldsymbol{i}}, \boldsymbol{\theta}) \equiv \hat{\chi}_i^2(\boldsymbol{\theta}) = \boldsymbol{d_i}^T\boldsymbol{d_i} - \hat{\boldsymbol{f}}_{\boldsymbol{i}}^{\,T}\hat{\boldsymbol{f}}_{\boldsymbol{i}}.$$

Thus, Equation (17) can be written from left to right as the sum of (a) a density term, (b) a term quantifying the goodness-of-fit and (c) a regularization term:

$$Q_F^{(i)} \simeq \left(\boldsymbol{b_i} - \hat{\boldsymbol{b}}_{\boldsymbol{i}}\right)^T\,\frac{\boldsymbol{g_i}}{\sigma^2}\left(\boldsymbol{b_i} - \hat{\boldsymbol{b}}_{\boldsymbol{i}}\right) + \frac{\hat{\chi}_i^2}{\sigma^2} + \frac{\hat{\boldsymbol{b}}_{\boldsymbol{i}}^{\,T}\hat{\boldsymbol{b}}_{\boldsymbol{i}}}{\delta^2}. \tag{18}$$

Having completed the square in $Q_F^{(i)}$, the integral over the amplitudes $\boldsymbol{b_i}$ is elementary, and we arrive at

$$\begin{aligned}
Z(P,Q) \simeq \int \mathrm{d}\boldsymbol{\theta}\, p(\boldsymbol{\theta}|I)\left[\int \mathrm{d}\sigma\, p(\sigma|I)\,(2\pi\sigma^2)^{\frac{nm-N}{2}}\,\exp\left\{-\frac{\sum_{i=1}^n \hat{\chi}_i^2}{2\sigma^2}\right\}\right] \\
\times\,(2\pi\delta^2)^{-nm/2}\prod_{i=1}^n |\det \boldsymbol{g_i}|^{-1/2}\exp\left\{-\frac{\hat{\boldsymbol{b}}_{\boldsymbol{i}}^{\,T}\hat{\boldsymbol{b}}_{\boldsymbol{i}}}{2\delta^2}\right\}.
\end{aligned} \tag{19}$$

When it comes to the polynomial amplitudes, marginalizing over them can be seen as detrending (also called background removal [41]). Likewise, marginalization over the damped sinusoid amplitudes corresponds to removing their amplitudes and phases, i.e., we are only interested in the poles.

The next step is to marginalize over the standard deviation by performing the integral in the large square brackets in Equation (19), i.e.,

$$\frac{1}{\log\left(\sigma^{\mathrm{hi}}/\sigma^{\mathrm{lo}}\right)} \int_{\sigma^{\mathrm{lo}}}^{\sigma^{\mathrm{hi}}} \mathrm{d}\sigma \, \frac{1}{\sigma} \left(2\pi\sigma^2\right)^{\frac{nm-N}{2}} \exp\left\{-\frac{\sum_{i=1}^{n} \hat{\chi}_i^2}{2\sigma^2}\right\}. \tag{20}$$

We assume a reasonable amount of model functions m in relation to the number of data points N such that $N > nm$. For states $(\boldsymbol{\theta}, \sigma)$ with appreciable likelihood practically all of the mass of Equation (20) is concentrated near the peak of its integrand at

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} \hat{\chi}_i^2}{N - nm + 1}}, \tag{21}$$

which we may assume to be within the bounds $[\sigma^{\mathrm{lo}}, \sigma^{\mathrm{hi}}]$ if these are sufficiently wide and a reasonable fit to the data is possible (see App. A in [39] for more details). Assuming this is the case, Equation (20) can be safely converted into an elementary gamma integral by letting $\sigma^{\mathrm{lo}} \to 0$ and $\sigma^{\mathrm{hi}} \to \infty$ and the marginalization can be performed analytically.

Doing so we finally obtain the expression for the integrated likelihood L_I:

$$Z(P, Q) = \int \mathrm{d}\boldsymbol{\theta} \, L_I(P, Q, \boldsymbol{\theta}) \, p(\boldsymbol{\theta}|I)$$

$$\text{with} \quad L_I(P, Q, \boldsymbol{\theta}) \simeq C(P, Q) \times \left[\sum_{i=1}^{n} \hat{\chi}_i^2\right]^{\frac{nm-N}{2}} \times \prod_{i=1}^{n} \left|\det \boldsymbol{g_i}\right|^{-1/2} \exp\left\{-\frac{\hat{\boldsymbol{b}_i}^T \hat{\boldsymbol{b}_i}}{2\delta^2}\right\},$$

where

$$C(P, Q) = \frac{1}{2} \frac{1}{\log\left(\sigma^{\mathrm{hi}}/\sigma^{\mathrm{lo}}\right)} \pi^{\frac{nm-N}{2}} \, \Gamma\left(\frac{N - nm}{2}\right) \left(2\pi\delta^2\right)^{-nm/2}$$

is a pure model order regularization term, being function only of P and Q. The factor $[\log\left(\sigma^{\mathrm{hi}}/\sigma^{\mathrm{lo}}\right)]^{-1}$ due to the normalization of the Jeffreys prior for σ is a constant independent of $(P, Q, \boldsymbol{\theta})$ and subsequently cancels out in model averaging and the weighting of posterior samples of $\boldsymbol{\theta}$.

3.2. Optimization Approaches

Though the nested sampling approach proposed here is different from the optimization approach we used in the conference paper, it is still possible to formulate a straightforward iterative optimization scheme for this problem. Indeed, a least-squares search in $\boldsymbol{\theta}$ can still be used—the marginalization over the amplitudes would then be called "variable projection" [51], and the amplitude regularization in Equation (18) can still be incorporated by treating the $\{\boldsymbol{b_i}\}$ as model predictions for nm additional datapoints measured to be zero with errorbar δ. To fix the scale for the actual N datapoints, these could be assigned fictional errorbars as well with magnitude $\hat{\sigma}$ as defined in Equation (21). Fast Fourier transformations, initially on the data and after on fit residuals, could be used for the initial guesses for the formant frequencies [39].

However, we have refrained from developing this approach mainly because of two reasons. First, due to the low dimensionality of the problem the nested sampling runs we ran to calculate $Z(P, Q)$ for the next section were not unbearably slow (even in a Python/`NumPy` context) as most runs finished under two minutes. Second, nested sampling allowed us to calculate the evidence for a model order $Z(P, Q)$ with confidence, while an optimization approach based on a Laplace approximation can easily give poor results. We did not, however, consider variational approaches.

4. Application to Data

In our experiments we used Praat [52] interfaced with the `parselmouth` Python library [53] with the default recommended settings to segment steady-state vowels into n pitch periods and get initial estimates for the formant bandwidths and frequencies to determine plausible ranges for their true values (Table 1). For the nested sampling we used the static sampler of the excellent `dynesty` Python library [54], again with default settings.

Table 1. The prior ranges $[\theta_j^{\text{lo}}, \theta_j^{\text{hi}}]$ for the Jeffreys priors for $\boldsymbol{\theta}$ used in the two applications of the model to data. The ranges for the formant bandwidths (α_j) and frequencies (ω_j) are given in Hz; for example, the prior range for the first formant ω_1 is 200–700 Hz. The data consists of a synthetic steady-state /ɤ/ (Section 4.1) and a real steady-state /æ/ (Section 4.2).

	$\boldsymbol{\alpha_1}$		$\boldsymbol{\alpha_2}$		$\boldsymbol{\alpha_3}$		$\boldsymbol{\alpha_4}$		$\boldsymbol{\omega_1}$		$\boldsymbol{\omega_2}$		$\boldsymbol{\omega_3}$		$\boldsymbol{\omega_4}$	
/ɤ/	10	180	10	250	10	420	/		200	700	700	1500	1500	3000	/	
/æ/	40	180	40	250	60	420	60	420	300	900	1000	2000	2000	3000	2500	4000

The range of the model order P was generally set to $P = (1, 2, \ldots, 10)$ while the allowed values of Q were more specific to the application (given that the signal bandwidth was limited to 4 kHz). In our experiments we found that high-degree polynomials tend to become too wiggly, thereby competing against the damped sinusoids for spectral power in awkwardly high frequency regions. It appeared that a good rule of thumb to prevent this behavior was to limit $P \leq 10$ (and thus set the maximum polynomial degree to 9). We also note that the case $P = 1$ together with $n = 1$ would correspond to the Pinson model of Section 1.2, if we would disregard the fact that we model the entire pitch period (from GCI to GCI) as opposed to only the portion between GCI and GOI.

4.1. Synthesized Steady-State /ɤ/

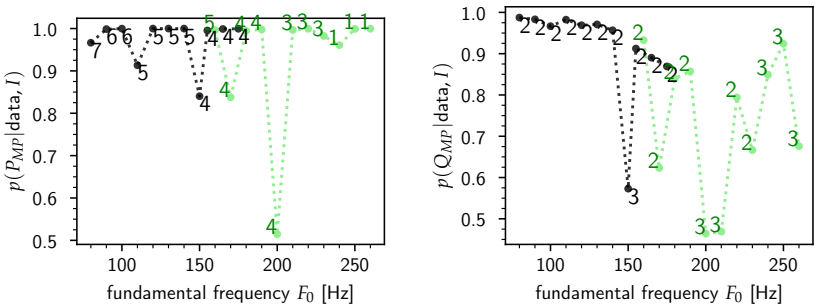
We apply the model first to a synthetic steady-state vowel /ɤ/ to verify the model's prediction accuracy and to see whether the inferred polynomial correlates with $u'(t)$, which we would expect based on the arguments of Section 2.4. The vowel was generated with different parameter settings [32] (p. 121) which emulate female and male speakers at different fundamental frequencies F_0 spanning the entire range of normal (non-pathological) speakers [4].

The vowel /ɤ/ was synthesized by first generating an artificial glottal source signal—the glottal flow derivative $u'(t)$—at a sampling rate of 16 kHz, which was then filtered by an all-pole VT transfer function consisting of $Q_{\text{true}} = 3$ poles with realistic formant values (which we will refer to as "the true values" from now on) based on Reference [55] (p. 163), and then downsampled to 8 kHz. Now yielding to the usual notation of acoustic phonetics, the true bandwidths $\boldsymbol{B}_{\text{true}}$ (our $\boldsymbol{\alpha}$s) and frequencies $\boldsymbol{F}_{\text{true}}$ (our $\boldsymbol{\omega}$s) used for the VT transfer function are $\boldsymbol{B}_{\text{true}} = (54, 22, 19)$ Hz and $\boldsymbol{F}_{\text{true}} = (430, 1088, 2142)$ Hz.

The glottal flow derivative $u'(t)$ was generated using the LF model [56]. For the male speakers, the 11 values of $F_0 = (80, 90, \ldots, 180)$ Hz, increasing in steps of 10 Hz. Likewise, for female speakers, the 11 values of $F_0 = (160, 170, \ldots, 260)$ Hz. We applied tiny but realistic values of jitter (0.5%) and shimmer (2%) to the generated pitch periods, which greatly improved the perceived naturalness of the steady-state vowel's sound. Finally, we selected the $n = 3$ central pitch periods for analysis.

As mentioned before, the range of P was set to $P = (1, 2, \ldots, 10)$. The allowed range of Q was set to $Q = (1, 2, 3)$. The ranges for the formant bandwidths and frequencies $[\theta_j^{\text{lo}}, \theta_j^{\text{hi}}]$ are given in Table 1.

For each synthesized steady-state /ɤ/, the posterior probability of P and Q was calculated—see Figure 2. It is clear that the majority of the most probable values P_{MP} and Q_{MP} are close to unity, which indicates that typically the model has an outspoken preference for a particular value of P and Q. In this case we see that $Q = 2 \neq Q_{\text{true}} = 3$ is heavily preferred, which means that in this particular experiment the model did not pick up the third formant [57]. Contrary to the number of formants Q, the preferred polynomial order $P - 1$ is more dependent on variations in F_0, with P about 5 to 6 for male speakers and smaller for female speakers.



(a) (b)

Figure 2. The most probable model orders P_{MP} and Q_{MP} and their posterior probability as calculated according to Equation (15). Each point in the graphs represents a synthesized steady-state /ɤ/ according to a speaker sex and fundamental frequency F_0. The sex is indicated by black (male) or lightgreen (female). The values of P_{MP} and Q_{MP} are indicated by text. **(a)** $p(P_{\text{MP}}|\text{ɤ}, I)$. **(b)** $p(Q_{\text{MP}}|\text{ɤ}, I)$.

In Figure 3, the results of a test of the model's prediction accuracy according to the model-averaged estimates in Equation (14) are shown for the frequency and bandwidth of the first (B_1, F_1) and second (B_2, F_2) formants. In accordance with Figure 2, we did not show the estimates for the third (B_3, F_3) formant as the most probable value of the number of formants in the data is $Q_{\text{MP}} = 2$—indeed, the errorbars for the estimates for (B_3, F_3) were huge, rendering those estimates practically useless.

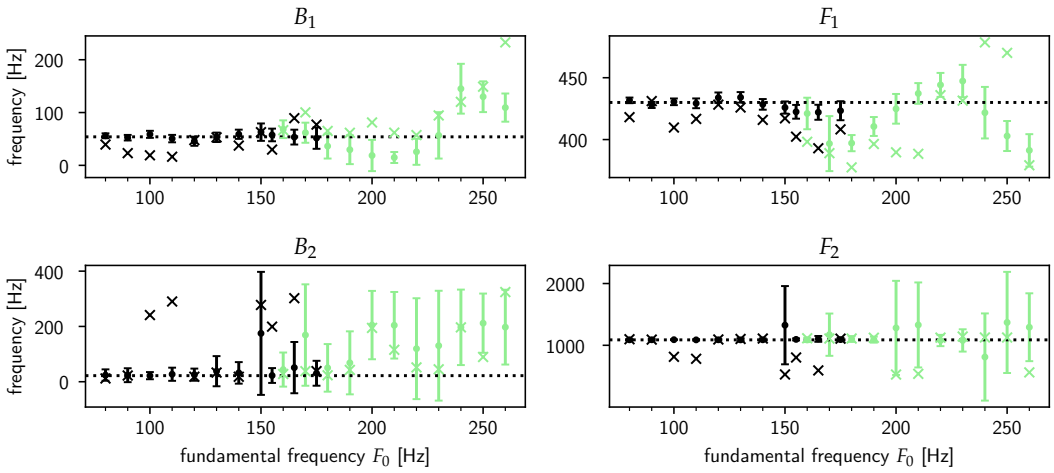


Figure 3. The model's prediction accuracy for the bandwidth and frequency of the first (B_1, F_1) and second (B_2, F_2) formants. Each point in the graphs represents an estimate either by our model or by Praat for a synthesized steady-state /ɤ/ according to a speaker sex and fundamental frequency F_0. The sex is indicated by black (male) or lightgreen (female). The model's estimates are averaged over all allowed model orders (i.e., values of (P, Q)) according to Equation (14), though in practice only one or two values of (P, Q) dominate (as suggested by Figure 2). The model's estimates are the dots with the errorbars at three standard deviations. The linear predictive coding (LPC) estimates acquired with Praat are plotted as crosses. The true values $\boldsymbol{B}_{\text{true}}$ and $\boldsymbol{F}_{\text{true}}$ are drawn as dotted horizontal lines.

Figure 3 shows a striking dichotomy between the results for male and female speakers. For the male speakers, the model's estimates seem to perform equally well or better than the LPC estimates. In contrast, the performance dramatically decreases for female speakers with the true values often outside of the already exceedingly large error bars. For B_2 and F_2, the model does communicate its uncertainty about the true value of the bandwidths and frequencies by returning estimates with huge error bars, but unfortunately for B_1 and F_1 its estimates can be quite misleading.

The reason for this significant change in performance is mainly due to the change in the fundamental frequency F_0. As F_0 rises, the near-impulse response waveforms triggered by the GCIs

tend to "spill over" into the next pitch period, i.e., the damped sinusoids caused by a given GCI are still ringing out appreciably when the next GCI happens, thus contaminating the pack of newly triggered decaying sinusoids. These nearest-neighbor effects increasingly wreck the assumption that a pitch period is only made up of a trend plus the VT impulse response. This is also evident from Figure 2a, where the most probable degree of the trend polynomial $(P_{\mathrm{MP}} - 1)$ drops to zero for very high fundamental frequencies, suggesting that the low-frequency content is picked up by F_1, which Figure 3 confirms. From this experiment it appears that the threshold of F_0 is around 150 Hz. This essentially means that *the model is limited to male speakers only*.

The difficulties we encounter here reflect a known phenomenon in the literature: formant analysis for female voices is in general harder compared to male voices, regardless of the method used (e.g., [15], see also [16] (pp. 124–126) for an excellent discussion). Indeed, the negative correlation between F_0 and the estimates' accuracy in Figure 3 is exhibited for both our model and Praat's LPC algorithm. Next to the higher F_0 we already mentioned, another cause of this phenomenon is the fact that the coupling between the glottal source and VT is generally stronger in females [58], which violates the assumption of source and filter separability which underlies source-filter theory ("the female VT is not merely a small-scale version of the male VT" [59]).

Next, in Figure 4, we look at a typical case for male speakers, $F_0 = 120$ Hz [32], for which the model performed quite well. The third formant F_3 can be seen clearly in the spectrum of the residuals, though the model concluded it was noise as can be seen by the rather large error bars on the fit residuals. The bottom panels also show that the inferred polynomial trend correlates well with the true $u'(t)$.

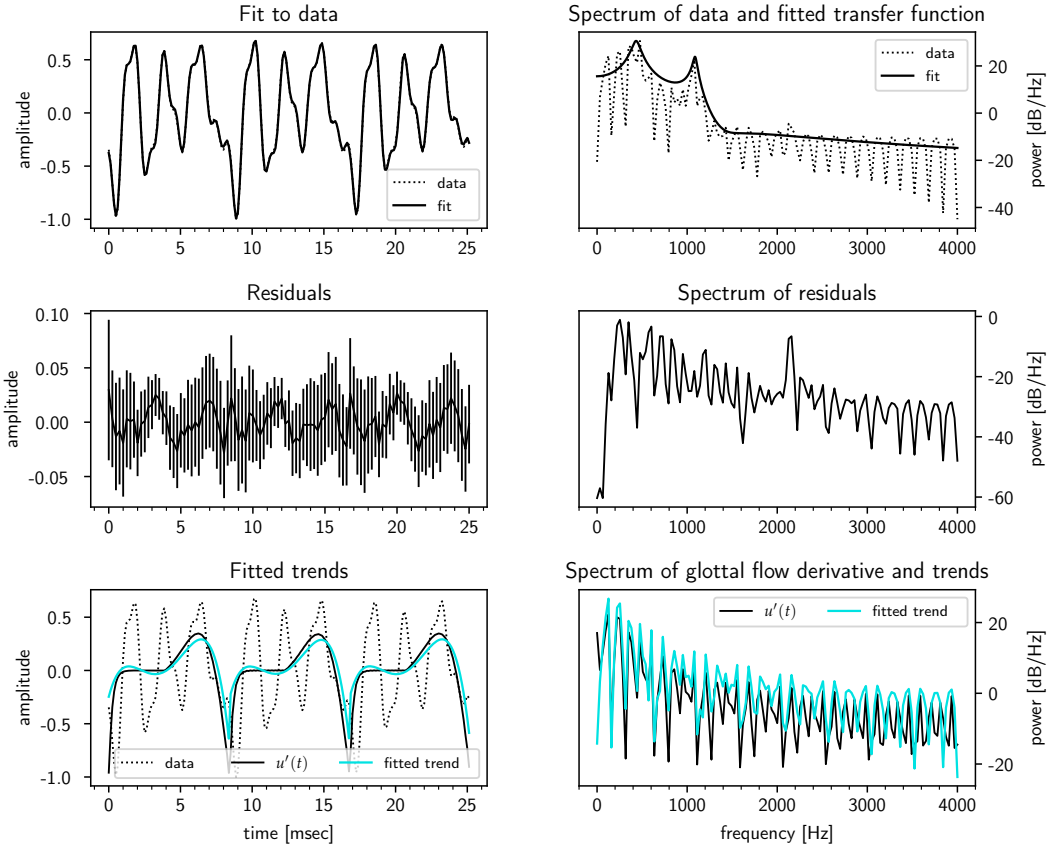


Figure 4. Fit results for a synthetic /ɤ/ in the case $F_0 = 120$ Hz for a male speaker. The fitted transfer function (solid line) in the top right panel is averaged over the $n = 3$ pitch periods as the inferred vocal tract (VT) transfer functions can in general have different zeros and gain constants (but must share the same poles $\boldsymbol{\theta}$). The errorbars on the residuals in the center left panel are at three standard deviations.

Finally, using the same synthesized steady-state vowel as in Figure 4, we gauge how inaccuracies in the segmentation of the steady-state vowel into n pitch periods affect the model's estimates. In Figure 5, we simulate errors in this preprocessing step by parametrizing the relative error in estimating the pitch periods $\{T_i = \tau_{i+1} - \tau_i\}$ as ϵ $(0 \leq \epsilon \leq 1)$ and perturb the $n+1$ known GCIs at $\{\tau_i\}$ according to

$$\tau_i \to \tau_i^{(\epsilon)} = \tau_i + \left[T_i - T_i^{(\epsilon)}\right] \quad \text{where} \quad \log T_i^{(\epsilon)} \sim N(\log T_i, \epsilon) \qquad (1 \leq i \leq n+1).$$

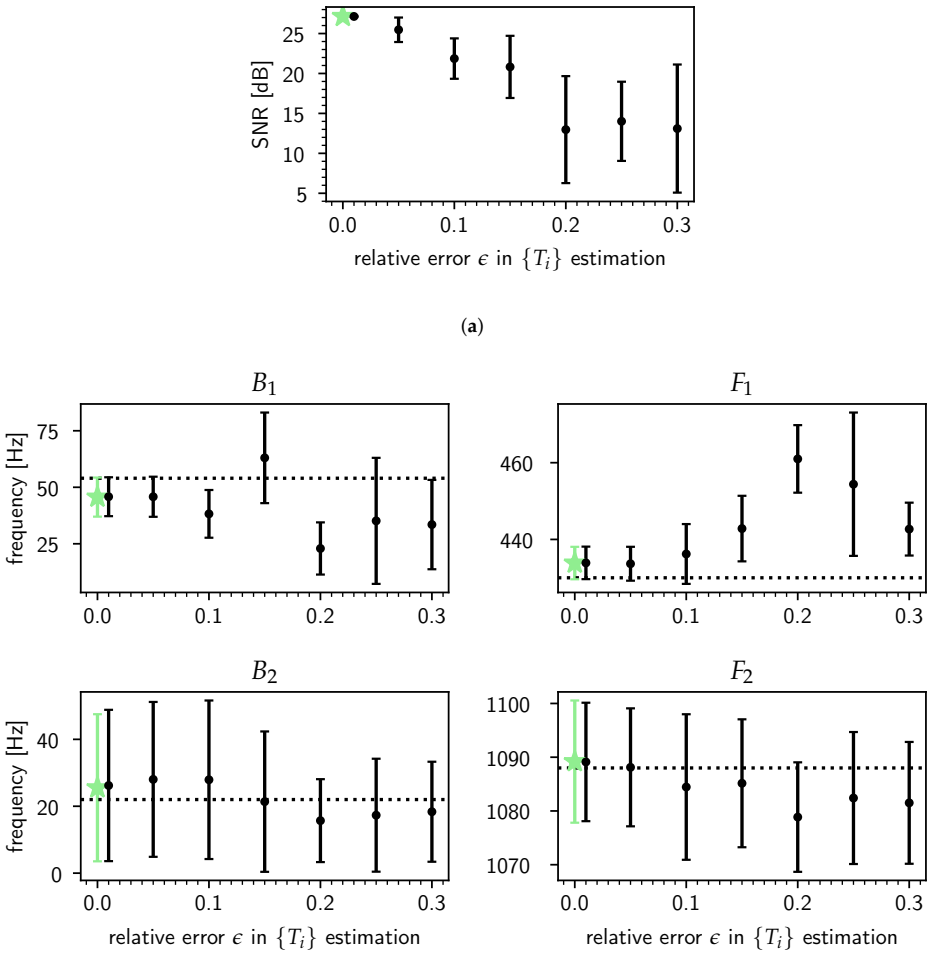


Figure 5. Testing the robustness of the bandwidth and frequency estimates of the first (B_1, F_1) and second (B_2, F_2) formants against increasing relative error ϵ in pitch period $\{T_i\}$ estimation for a synthetic /ɤ/ in the case $F_0 = 120$ Hz for a male speaker. Errors in $\{T_i\}$ estimation induce errors in the pitch period segmentation according to the GCIs $\{\tau_i\}$ and thus transfer to the formant estimates, which are acquired through model averaging as defined in Equation (14). The method is explained in detail in the main text. In each panel the green star indicates the estimates for the unperturbed $\{\tau_i\}$ (i.e. $\epsilon = 0\%$), for which no averaging has been done. (**a**) The fit quality as gauged by the SNR (defined in Section 2) as a function of ϵ. Each point and its errorbar are the empirical mean and standard deviation at three σ, respectively, over 6 draws. For reference, the *prediction gain* of adaptive LPC for stationary voiced speech sounds is typically about 20 dB [60] (p. 70). (**b**) Comparison of the formant estimates as a function of ϵ to their true values $\boldsymbol{B}_{\text{true}}$ and $\boldsymbol{F}_{\text{true}}$ (dotted horizontal lines). Each point is the empirical mean of the point estimates over 6 draws, and each errorbar is the empirical mean of the point estimates' *errorbars* at three standard deviations over the same 6 draws.

The steady-state vowel is then segmented into n pitch periods $\{\boldsymbol{d}_1^{(\epsilon)} \cdots \boldsymbol{d}_n^{(\epsilon)}\}$ using the set of perturbed GCIs $\{\tau_i^{(\epsilon)}\}$ and estimates of the noise amplitude $\hat{\sigma}^{(\epsilon)}$ and the $\hat{\boldsymbol{\theta}}^{(\epsilon)}$ are obtained. We repeated this procedure 6 times by drawing 6 sets of the $\{T_i^{(\epsilon)}\}$ for each value of $\epsilon \in \{1\%, 5\%, 10\%, 15\%, \ldots, 30\%\}$. The results averaged over the draws are shown in Figure 5a,b. The conclusion from this particular experiment is that while the fit quality deteriorates strongly as the relative error in estimating the pitch periods grows (a), the formant estimates degrade relatively gracefully (b). One contributing factor for this is a feature of our model: it performs a kind of generalized averaging [39] (Sec. 7.5) over the pitch periods to arrive at "robust" estimates of the $\boldsymbol{\theta}$.

4.2. Real Steady-State /æ/

For real data, we do not know the underlying glottal source $u'(t)$ as this is very hard to measure reliably. An alternative to measuring the glottal flow directly is the electroglottograph (EGG). The EGG signal can be used as a probe for the glottal source, as we will explain below.

The CMU ARCTIC database [29] consists of utterances which are recorded simultaneously with an EGG signal. The source of the steady-state /æ/ used in this section is a male speaker called BDL, file `bdl/arctic_a0017.wav`, 0.51–0.54 s, resampled to 8000 Hz. The fundamental frequency F_0 is about 138 Hz.

Once again the range of P was set to $P = (1, 2, \ldots, 10)$. The allowed range of Q was set to $Q = (1, 2, 3, 4)$. The ranges for the formant bandwidths and frequencies $[\theta_j^{\text{lo}}, \theta_j^{\text{hi}}]$ are given in Table 1.

In Figure 6, the posterior probability for the individual model orders is shown, from which a clear preference for $P = 5$ and $Q = 4$ arises.

In Figure 7, we show the model-averaged posterior distributions in Equation (14) for the formant bandwidths and frequencies. The estimates of the frequencies are reasonably sharp and agree quite well with the LPC estimates obtained with Praat. The error bars on the bandwidths increase gradually until the uncertainty on B_4 has become so large that it is essentially unresolved. This increase in the uncertainty on the bandwidths mirrors the fact that measuring bandwidths becomes increasingly difficult for higher formants [13].

Finally, we correlate the inferred trend together with the EGG signal in Figure 8. The EGG signal is the electrical conductance between two electrodes placed on the neck. When the glottis is closed and the glottal flow $u(t)$ is zero, the measured conductance is high, and vice versa. The EGG signal rises sharply at the GCI, i.e., when the glottal flow drops abruptly to zero. From the discussion in Section 2.4 and in particular Equation (10), the glottal flow $u(t)$ modulo a bias constant can be estimated very roughly from the inferred polynomial by integrating it over time. It is seen in the plot that the expected anticorrelation is borne out: when the $u(t)$ estimate hits a through (GCI), the EGG signal rises sharply. Conversely, when the $u(t)$ estimate hits a peak (GOI), the EGG signal hits a through, as the electrical conductance across the two electrodes drops due to the opening of the glotts.

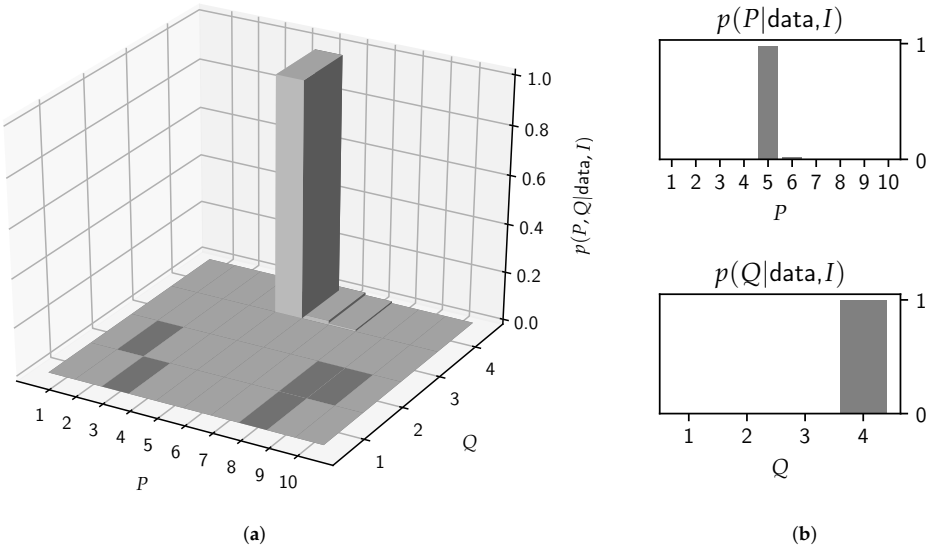


Figure 6. The posterior probabilities of the joint (**a**) and separate (**b**) model orders for the steady-state /æ/ according to Equation (15). In this case, model averaging is for all practical purposes equivalent to model selection as the model $(P = 5, Q = 4)$ occupies 98% of the posterior mass.

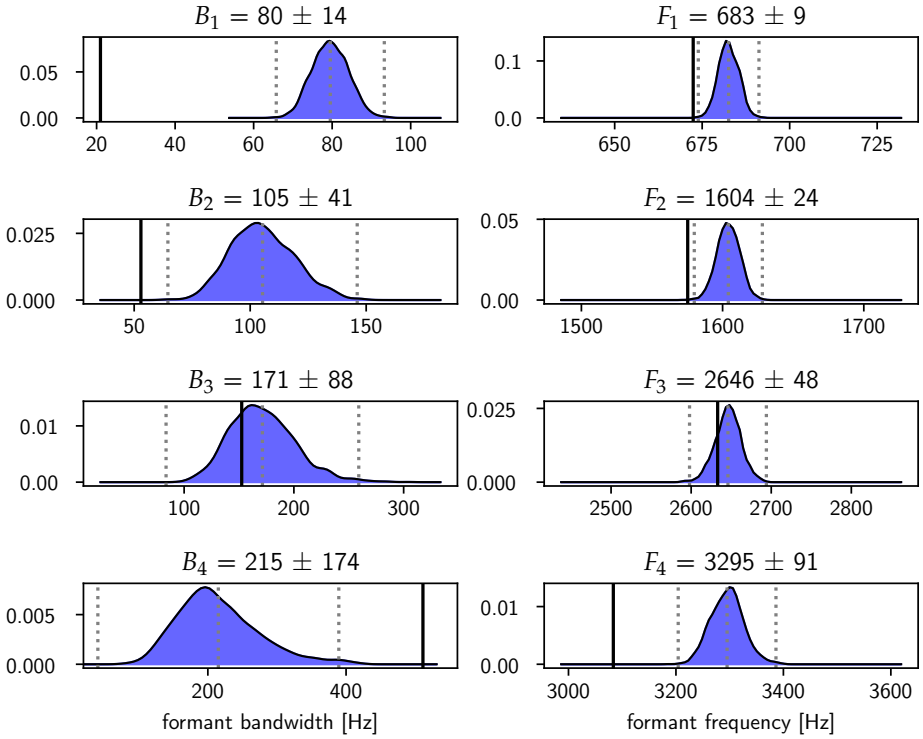


Figure 7. Posterior distributions $p(\boldsymbol{\theta}|æ, I)$ of the formant bandwidths $\boldsymbol{B}$ and frequencies $\boldsymbol{F}$. The distributions are estimated using Gaussian kernel density estimation for the combined samples $\boldsymbol{\theta}_{P,Q}^{(l)}$ which are reweighted according to $w_{P,Q}^{(l)} \to w_{P,Q}^{(l)} \times Z(P,Q)/\sum_{P,Q} Z(P,Q)$. The dotted vertical lines indicate a distance of three standard deviations from the mean, which is also stated in the panel titles together with the point estimate. The solid vertical lines indicate the LPC estimates obtained with Praat.

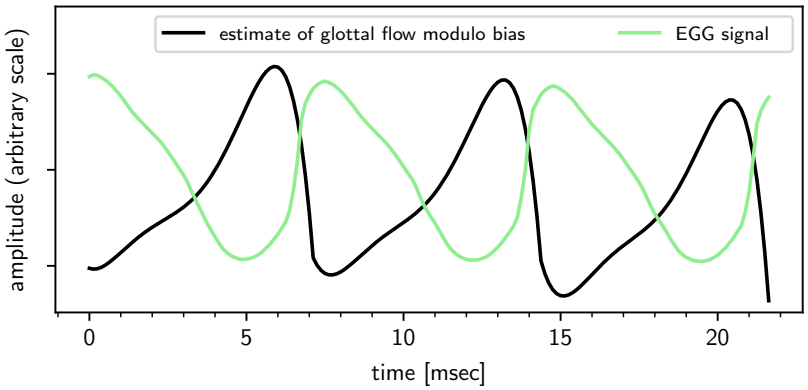


Figure 8. Comparison of the estimate of $u(t)$ modulo a bias constant and the measured electroglottograph (EGG) signal. The speech signal, and therefore the $u(t)$ estimate, lags behind the EGG signal by approximately 1 ms due to the distance between the glottis and the microphone.

5. Conclusions

The proposed model is a modest step towards formant estimation with reliable uncertainty quantification in the case of steady-state vowels. In our approach, the uncertainty quantification is implemented through Bayesian model averaging. The validity of our approach depends on the assumption that pitch periods can be modeled accurately as being composed of a slowly changing trend superimposed on a set of decaying sinusoids that represent the impulse response of the VT. It appears that this assumption likely holds only for fundamental frequencies F_0 below about 150 Hz, which poses a grave restriction on its use as this excludes most female speakers.

Author Contributions: Conceptualization and writing: M.V.S. and B.d.B.; methodology and analysis: M.V.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Flemish AI plan and by the Research Foundation Flanders (FWO) under grant number G015617N.

Acknowledgments: The authors thank the three anonymous reviewers for their valuable comments which have greatly improved the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References and Notes

1. Van Soom, M.; de Boer, B. A New Approach to the Formant Measuring Problem. *Proceedings* **2019**, *33*, 29. [CrossRef]
2. Fulop, S.A. *Speech Spectrum Analysis*; OCLC: 746243279; Signals and Communication Technology; Springer: Berlin, Germany, 2011.
3. Fant, G. *Acoustic Theory of Speech Production*; Mouton: Den Haag, The Netherlands, 1960.
4. Stevens, K.N. *Acoustic Phonetics*; MIT Press: Cambridge, CA, USA, 2000.
5. Rabiner, L.R.; Schafer, R.W. *Introduction to Digital Speech Processing*; Foundations and Trends in Signal Processing: Hanover, MA, USA, 2007. [CrossRef]
6. Rose, P. *Forensic Speaker Identification*; CRC Press: Boca Raton, FL, USA, 2002.
7. Ng, A.K.; Koh, T.S.; Baey, E.; Lee, T.H.; Abeyratne, U.R.; Puvanendran, K. Could Formant Frequencies of Snore Signals Be an Alternative Means for the Diagnosis of Obstructive Sleep Apnea? *Sleep Med.* **2008**, *9*, 894–898. [CrossRef] [PubMed]
8. Singh, R.; Raj, B.; Gencaga, D. Forensic Anthropometry from Voice: An Articulatory-Phonetic Approach. In Proceedings of the 2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 30 May–3 June 2016; pp. 1375–1380. [CrossRef]
9. Jaynes, E.T. *Probability Theory: The Logic of Science*; Bretthorst, G.L., Ed.; Cambridge University Press: Cambridge, UK; New York, NY, USA, 2003.

10. Bonastre, J.F.; Kahn, J.; Rossato, S.; Ajili, M. Forensic Speaker Recognition: Mirages and Reality. Available online: https://www.oapen.org/download?type=document&docid=1002748#page=257 (accessed on 12 March 2020).
11. Hughes, N.; Karabiyik, U. Towards reliable digital forensics investigations through measurement science. *WIREs Forensic Sci.* **2020**, e1367, [https://onlinelibrary.wiley.com/doi/pdf/10.1002/wfs2.1367]. [CrossRef]
12. De Witte, W. A Forensic Speaker Identification Study: An Auditory-Acoustic Analysis of Phonetic Features and an Exploration of the "Telephone Effect". Ph.D. Thesis, Universitat Autònoma de Barcelona, Bellaterra, Spain, **2017**.
13. Kent, R.D.; Vorperian, H.K. Static Measurements of Vowel Formant Frequencies and Bandwidths: A Review. *J. Commun. Disord.* **2018**, *74*, 74–97. [CrossRef]
14. Mehta, D.D.; Wolfe, P.J. Statistical Properties of Linear Prediction Analysis Underlying the Challenge of Formant Bandwidth Estimation. *J. Acoust. Soc. Am.* **2015**, *137*, 944–950. [CrossRef]
15. Harrison, P. Making Accurate Formant Measurements: An Empirical Investigation of the Influence of the Measurement Tool, Analysis Settings and Speaker on Formant Measurements. Ph.D. Thesis, University of York, York, UK, 2013.
16. Maurer, D. *Acoustics of the Vowel*; Peter Lang: Bern, Switzerland, 2016.
17. Shadle, C.H.; Nam, H.; Whalen, D.H. Comparing Measurement Errors for Formants in Synthetic and Natural Vowels. *J. Acoust. Soc. Am.* **2016**, *139*, 713–727. [CrossRef]
18. Knuth, K.H.; Skilling, J. Foundations of Inference. *Axioms* **2012**, *1*, 38–73. [CrossRef]
19. Pinson, E.N. Pitch-Synchronous Time-Domain Estimation of Formant Frequencies and Bandwidths. *J. Acoust. Soc. Am.* **1963**, *35*, 1264–1273. [CrossRef]
20. Fitzgerald, W.J.; Niranjan, M. Speech Processing Using Bayesian Inference. In *Maximum Entropy and Bayesian Methods: Paris, France, 1992*; Mohammad-Djafari, A., Demoment, G., Eds.; Fundamental Theories of Physics; Springer: Dordrecht, The Netherlands, 1993; pp. 215–223. [CrossRef]
21. Nielsen, J.K.; Christensen, M.G.; Jensen, S.H. Default Bayesian Estimation of the Fundamental Frequency. *IEEE Trans. Audio Speech Lang. Process.* **2013**, *21*, 598–610. [CrossRef]
22. Peterson, G.E.; Shoup, J.E. The Elements of an Acoustic Phonetic Theory. *J. Speech Hear. Res.* **1966**, *9*, 68–99. [CrossRef] [PubMed]
23. Little, M.A.; McSharry, P.E.; Roberts, S.J.; Costello, D.A.; Moroz, I.M. Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection. *Biomed. Eng. Online* **2007**, *6*, 23. [CrossRef] [PubMed]
24. Titze, I.R. *Workshop on Acoustic Voice Analysis: Summary Statement*; National Center for Voice and Speech: Salt Lake City, UT, USA, 1995.
25. While our (and others' [27]) everyday experience of looking at speech waveforms confirms this, we would be very interested in a formal study on this subject.
26. Hermann, L. Phonophotographische Untersuchungen. *Pflüg. Arch.* **1889**, *45*, 582–592. [CrossRef]
27. Chen, C.J. *Elements of Human Voice*; World Scientific: Singapore, 2016. [CrossRef]
28. Scripture, E.W. *The Elements of Experimental Phonetics*; C. Scribner's Sons: New York, NY, USA, 1904.
29. Kominek, J.; Black, A.W. The CMU Arctic Speech Databases. 2004. Available online: http://festvox.org/cmu_arctic/cmu_arctic_report.pdf (accessed on 12 March 2020).
30. Ladefoged, P. *Elements of Acoustic Phonetics*; University of Chicago Press: Chicago, IL, USA, 1996.
31. Drugman, T.; Thomas, M.; Gudnason, J.; Naylor, P.; Dutoit, T. Detection of Glottal Closure Instants From Speech Signals: A Quantitative Review. *IEEE Trans. Audio Speech Lang. Process.* **2012**, *20*, 994–1006. [CrossRef]
32. Fant, G. The LF-model revisited. Transformations and frequency domain analysis. *Speech Trans. Lab. Q. Rep. R. Inst. Tech. Stockh.* **1995**, *2*, 40.
33. Thus, when a vowel is perceived with a clear and constant pitch, it is reasonable to assume that the vowel attained steady-state at some point (though perceptual effects forbid a one-to-one correspondence).
34. Fant, G. Formant Bandwidth Data. *STL-QPSR* **1962**, *3*, 1–3.
35. House, A.S.; Stevens, K.N. Estimation of Formant Band Widths from Measurements of Transient Response of the Vocal Tract. *J. Speech Hear. Res.* **1958**, *1*, 309–315. [CrossRef]
36. Sanchez, J. Application of Classical, Bayesian and Maximum Entropy Spectrum Analysis to Nonstationary Time Series Data. In *Maximum Entropy and Bayesian Methods*; Springer: Berlin/Heidelberger, Germany, 1989; pp. 309–319.

37. Skilling, J. Nested Sampling for General Bayesian Computation. *Bayesian Anal.* **2006**, *1*, 833–859. [CrossRef]
38. Ó Ruanaidh, J.J.K.; Fitzgerald, W.J. *Numerical Bayesian Methods Applied to Signal Processing*; Statistics and Computing; Springer: New York, NY, USA, 1996. [CrossRef]
39. Bretthorst, G.L. *Bayesian Spectrum Analysis and Parameter Estimation*; Springer Science & Business Media: Berlin/Heidelberger, Germany, 1988.
40. Jaynes, E.T. Bayesian Spectrum and Chirp Analysis. In *Maximum-Entropy and Bayesian Spectral Analysis and Estimation Problems*; Reidel: Dordrecht, The Netherlands, 1987; pp. 1–29.
41. Sivia, D.; Skilling, J. *Data Analysis: A Bayesian Tutorial*; OUP: Oxford, UK, 2006.
42. These are improved versions of the ones used in the conference paper.
43. This makes no difference in the posterior distribution for $\boldsymbol{\theta}$ because the Legendre functions are linear combinations of the polynomials.
44. In our experiments, this regularizer prevented a situation that can best be described as polynomial and sinusoidal basis functions with huge amplitudes conspiring together into creating beats that, added together, fitted the data quite well but would yield nonphysical values for the $\boldsymbol{\theta}$.
45. Peterson, G.E.; Barney, H.L. Control Methods Used in a Study of the Vowels. *J. Acoust. Soc. Am.* **1952**, *24*, 175–184. [CrossRef]
46. Another approach which avoids setting these prior ranges explicitly uses the perceived reliability of the LPC estimates to assign an expected *relative* accuracy ρ (e.g. 10%) to the LPC estimates $\hat{\boldsymbol{\theta}}_{\mathrm{LPC}}$. It is then possible to assign lognormal prior pdfs for the $\boldsymbol{\theta}$ which are parametrized by setting the mean and standard deviation of the underlying normal distributions to $\log \hat{\boldsymbol{\theta}}_{\mathrm{LPC}}$ and ρ, respectively [41]. This works well as long as $\rho \leq 0.40$, regardless of the value of $\hat{\boldsymbol{\theta}}_{\mathrm{LPC}}$. We use the same technique in Section 4.1 to simulate errors in pitch period segmentation.
47. Fant, G. The Voice Source-Acoustic Modeling. *STL-QPSR* **1982**, *4*, 28–48.
48. Rabiner, L.R.; Juang, B.H.; Rutledge, J.C. *Fundamentals of Speech Recognition*; PTR Prentice Hall Englewood Cliffs: Upper Saddle River, NJ, USA, 1993; Volume 14.
49. Deller, J.R. On the Time Domain Properties of the Two-Pole Model of the Glottal Waveform and Implications for LPC. *Speech Commun.* **1983**, *2*, 57–63. [CrossRef]
50. Petersen, K.; Pedersen, M. *The Matrix Cookbook* Technical University of Denmark: Copenhagen, Denmark, 2008; Volume 7.
51. Golub, G.; Pereyra, V. Separable nonlinear least squares: The variable projection method and its applications. *Inverse Prob.* **2003**, *19*, R1. [CrossRef]
52. Boersma, P. Praat, a system for doing phonetics by computer. *Glot. Int.* **2001**, *5*, 341–345.
53. Jadoul, Y.; Thompson, B.; de Boer, B. Introducing Parselmouth: A Python interface to Praat. *J. Phonetics* **2018**, *71*, 1–15. [CrossRef]
54. Speagle, J.S. dynesty: A dynamic nested sampling package for estimating Bayesian posteriors and evidences. *arXiv* **2019**, arXiv:1904.02180.
55. Vallée, N. Systèmes Vocaliques: De La Typologie Aux Prédictions. Ph.D. Thesis, l'Université Stendhal, Leeuwarden, The Netherlands, 1994.
56. Fant, G.; Liljencrants, J.; Lin, Q.G. A Four-Parameter Model of Glottal Flow. *STL-QPSR* **1985**, *4*, 1–13.
57. Praat's default formant estimation algorithm preprocesses the speech data by applying a +6 dB/octave *pre-emphasis* filter to boost the amplitudes of the higher formants. The goal of this common technique [2] is to facilitate the measurement of the higher formants' bandwidths and frequencies. We have not applied pre-emphasis in our experiments, so it remains to be seen whether the third formant could have been picked up in this case. On that note, it might be of interest that this *plus* 6 dB/oct pre-emphasis filter can be expressed in our Bayesian approach as prior covariance matrices $\{\boldsymbol{\Sigma}_i\}$ for the noise vectors $\{\boldsymbol{e}_i\}$ in Equation (5) specifying approximately a *minus* 6 dB/oct slope for the prior noise spectral density. To see this, write the pre-emphasis operation on the data *and model functions* as: $\boldsymbol{d_i} \to \boldsymbol{E_i}\,\boldsymbol{d_i}, \quad \boldsymbol{G_i} \to \boldsymbol{E_i}\,\boldsymbol{G_i}, \quad (i = 1 \cdots n)$, where $\boldsymbol{E}_i$ is a real and invertible $N_i \times N_i$ matrix representing the pre-emphasis filter (for example, Praat by default uses $y_t \approx x_t - 0.98\,x_{t-1}$ which corresponds to $\boldsymbol{E_i}$ having ones on the principal diagonal and -0.98 on the subdiagonal). Then, the likelihood function in Equation (8) is proportional to $\exp\{-(1/2)\sum_{i=1}^{n}(\boldsymbol{d_i} - \boldsymbol{G_i}\,\boldsymbol{b_i})^T\,\boldsymbol{\Sigma_i}^{-1}\,(\boldsymbol{d_i} - \boldsymbol{G_i}\,\boldsymbol{b_i})\}$ where the $\boldsymbol{\Sigma_i} \propto (\boldsymbol{E_i}^T\boldsymbol{E_i})^{-1}$ are positive definite covariance matrices specifying the prior for the spectral density of the noise, which turns out to be approximately -6 dB/oct. Thus, pre-emphasis preprocessing can be interpreted as a more informative noise prior.

58. Titze, I.R. *Principles of Voice Production (Second Printing)*; National Center for Voice and Speech: Iowa City, IA, USA, 2000.
59. Sundberg, J. Synthesis of singing. *Swed. J. Musicol.* **1978**, pp. 107–112.
60. Schroeder, M.R. *Computer Speech: Recognition, Compression, Synthesis*; Springer Series in Information Sciences; Springer-Verlag: Berlin/Heidelberg, Germany, 1999. [CrossRef]

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).