Time-Varying Modeling of Glottal Source and Vocal Tract and Sequential Bayesian Estimation of Model Parameters for Speech Synthesis

by

Adarsh Akkshai Venkataramani

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved November 2018 by the
Graduate Supervisory Committee:

Antonia Papandreou-Suppappola, Chair
Daniel W. Bliss
Pavan Turaga

ARIZONA STATE UNIVERSITY

December 2018

ABSTRACT

Speech is generated by articulators acting on a phonatory source. Identification of this phonatory source and articulatory geometry are individually challenging and ill-posed problems, called speech separation and articulatory inversion, respectively. There exists a trade-off between decomposition and recovered articulatory geometry due to multiple possible mappings between an articulatory configuration and the speech produced. However, if measurements are obtained only from a microphone sensor, they lack any invasive insight and add additional challenge to an already difficult problem. A joint non-invasive estimation strategy that couples articulatory and phonatory knowledge would lead to better articulatory speech synthesis. In this thesis, a joint estimation strategy for speech separation and articulatory geometry recovery is studied. Unlike previous periodic/aperiodic decomposition methods that use stationary speech models within a frame, the proposed model presents a non-stationary speech decomposition method. A parametric glottal source model and an articulatory vocal tract response are represented in a dynamic state space formulation. The unknown parameters of the speech generation components are estimated using sequential Monte Carlo methods under some specific assumptions. The proposed approach is compared with other glottal inverse filtering methods, including iterative adaptive inverse filtering, state-space inverse filtering, and the quasi-closed phase method.

DEDICATION

To my family, friends and mentors

ACKNOWLEDGMENTS

Firstly, I would like to thank my advisor Dr. Antonia Papandreou-Suppappola for her continued guidance throughout my M.Sc. program. I'm really grateful for her patience, encouragement, guidance, motivation and trust in my abilities. She gave me a chance to work with many people and helped me grow. Additionally, my committee members Dr. Pavan Turaga and Dr. Daniel W. Bliss for their feedback and their time in attending my defense. Secondly, my family, my mom for being a role model in my life. Her enthusiasm and tenacity of life has always amazed me, and I strive to learn from her every day. My dad for his advice on "way of life" and lessons to enjoy the journey of life. My sister Reshma for all the sisterly love and care. My aunts, uncles and cousins, Ashwin, Anisha, Mihir, and Sumir, for their support, love, and all the wonderful memories. Finally, my grandparents for sticking through all my tough phases of life and nurturing me to believe in myself and to never give up.

I would like to thank my friend Bahman Moraffah for his unyielding support as a, friend, counselor and understanding lab mate. Cesar Brito and Bindiya Venkatesh for being great friends and accompanying me in my lab adventures. This thesis wouldn't be complete without my besties; Sharath Jayan, Anik Jha, Madhura Bhat and Utkarsh Pandey. I'm grateful for their trust and love in me during hard times. Special thanks to Gauri Jagatap for being a loyal friend and perpetually supportive of my future, I'm thankful for her love and support in my life. My roomates Nischal Naik, Ram Kiran and Jayanth Kumar for bearing my erratic schedules and excusing me for switching off the central air conditioning! My extended roomates Namrata Gorantla, Deepika Reddy, Ashwini Jambagi, Reshma Naladala, Simarpreet Kaur for all the fun conversations, especially during cooking! Lastly, I would like to thank Dinesh Kunte, Dinesh Mylsamy and Rohith Undralla for saving my life.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Background and Motivation

Speech is widely considered as a confluence of two fundamentally disjoint events, perception and acoustic radiation. Language is the semantic content perceived by the brain from an often discontinuous set of acoustic waves. On the other hand, acoustic radiation is the physical phenomenon of wave propagation due to changes in pressure inside a medium. In humans, acoustic radiation is produced as a consequence of exciting the air in the lungs through a thin opening in the larynx called glottis (or phonatory source). The air flows downstream of the glottis, and as it passes through a series of biological resonators (or vocal tract), the frequency content of the radiated acoustic waves is altered. In particular, a particular bandwidth of frequencies are either dampened or intensified. For these frequencies, there exist spectral peaks that form the basis for all acoustic radiation emitted and perceived as speech. Since, every human has a unique resonator, speech varies in quality through timbre, granularity and color. The study of speech is a challenging task and is broadly classified based on these two events, perception and generation.

Speech recognition, which is the digital conversion of acoustic radiation to language, connects the observed spectral content to human understanding of acoustic radiation. This is essential for generating transcriptions (speech-to-text) that help in deducing deeper meaning and understanding of language in the absence of human supervision [4, 5]. Similarly, speech synthesis targets to reproduce speech for a given context using our understanding of language. The synthesis process is segregated into two sub domains: a *statistical* framework, that exploits data obtained through mathematical transformations, and an *articulatory* framework, that follows the physics and biological constructions of the human vocal apparatus [2, 6–9]. The statistical framework includes, for example, Mel-frequency ceptral coefficents, logarithmic features resident in observed speech, linear prediction coefficients, coefficients of an auto regressive filter that approximate peaks in frequency [10], and deep neural network generators. In articulatory speech synthesis, the resonator is influenced by biological aids or articulators such as lips, incisors, tongue tip, tongue blade, tongue dorsum and jaws. An envelope over positions of these articulatory units represents a structural configuration inside the vocal apparatus. The shape of this envelope forms an individual's personality, identity, and expression in voice [11]. Articulatory speech[1] sounds more natural due to its resemblance to human biology [12]. In their respective interpretations and usage, both frameworks have unique advantages.

Biologically, the glottal source and the vocal tract (VT) are the main two components in speech generation. Contributing new quantitative results for each of these components can provide information to help in speech learning studies, speech analysis, speech coding, speech synthesis and speaker recognition [13]. It is hence, important to study each source separately. This is similar to how understanding the anatomy during speech production can help in identifying symptoms for major disorders like dysarthia, ALS, and many more [14]. A long-standing issue in articulatory speech research is the acoustic-articulatory inversion problem. This is the problem of estimating a unique articulatory envelope and its mapping to acoustic parameters [15]. In articulatory speech production, the synthesis process is time-varying. A rapid transition between different articulatory states generates speech. Hence, the extraction of the dynamic information of the articulatory envelope is crucial for iden-

[1] In this thesis, all future use of the word "speech" is synonymous to observed acoustic radiation.

tifying and synthesizing speech. A non-invasive process in obtaining this anatomy paves the way to many avenues in speech research [16].

1.2 Current Work on Glottal Source and VT Response Modeling

Numerous methods have been considered in the literature that jointly model the glottal source and the VT response for speech generation. These methods mainly consider speech as the output of a composition of linear and time-invariant (LTI) systems using auto-regressive (AR) or auto-regressive moving-average (ARMA) models [1, 17–21]. Using these LTI system models of speech, the estimation of the VT response is tantamount to obtaining a glottal signal. In particular, glottal inverse filtering and VT response recovery are treated as interchangeable problems [22]. In contrast, clinically observed glottal signals include jitter and shimmer, which are time-varying phenomena in that their frequency content can change with time [23]. LTI models do not incorporate any time-varying speech components within a locally analyzed frame; as a result, the harmonic component of glottis is considered periodic. Furthermore, the coupling of articulators, that are assumed absent in LTI models, least resemble biological phenomena [24]. In [25], linear time-varying (LTV) systems were introduced to speech using ARMA models. Also, in [26], articulatory analysis-by-synthesis methods were combined with the Maeda model for magnetic resonance imaging (MRI) data. Note, however, that not much work has been done to study the glottal source and the VT response jointly. Recently, in [27], a concatenated tube model has been used to understand the coupling between the VT and glottis for glottal inverse filtering. In [18, 19], an LTV system was used to model the VT response and a parametric glottal source was estimated, assuming an ARMA model. In [13], a glottal inverse filtering estimate was obtained using a sum of sinusoids model that better matched empirical data obtained through Electromagnetic Midsaggital Articulography (EMA) and Electroglottalography(EGG).

The relationship between articulatory vectors and speech signals has been well-studied using MRI, for the problem of speech-to-articulatory inversion [14, 20, 26, 28, 29]. Recently, in [30], a ResNet model was trained using MRI and speech data to identify articulatory envelopes. However, a main concern in such formulations is the requirement of sufficient MRI sampling. MRI signals are sampled at 200 Hz, as compared to the much larger sampling frequencies of 16 to 48 kHz used by speech signals. Recently, blind approaches were used in [28, 29] to attempt recovery based on the sensitivity to acoustic features instead of the actual waveforms. To the best of our knowledge, a non-invasive speech-to-articulatory inversion using time-domain signal matching has not been attempted.

1.3 Thesis Contribution

In this work, we propose a time-varying model that separately decomposes the glottal source and the VT response speech generation components. The model indirectly obtains the VT configuration in the form of a two-dimensional area function that follows acoustic theory [31]. In particular, we select an articulatory model that translates the VT area function to impedances, and it produces a VT transfer function that is biologically coupled for a glottal source [32]. We also assume a Liljencrant-Fant parametric glottal source model that is coupled, due to the decomposition time and complexity [33]. The resulting glottal source and VT response is formulated in a dynamic state space formulation. The resulting glottal source, VT response, and unknown state components are jointly estimated using the bootstrap particle filter sequential Monte Carlo method. Note, however, that the estimated speech generation components are highly dependent on the assumptions we made, and clearly state, in the complexity of the equations in the dynamic state space formulation.

Our proposed method is compared against other glottal inverse filtering methods, including the iterative adaptive inverse filtering, the state-space inverse filtering, and the quasi-closed phase method. In order to assess the error in glottal source estimation, we use various metrics including the normalized amplitude quotient, formant estimation error, and estimation mean-squared error. The estimated VT configuration is compared using extracted VT areas from MRI data of selected speech samples.

1.4 Thesis Organization

The remainder of the thesis is organized as follows. Chapter 2 provides background information on models for the glottal source and the VT response. Chapter 3 reviews sequential Bayesian estimation methods. The dynamic state space formulation for the speech generation components is described in Chapter 4, and comparative results are provided in Chapter 5. Chapter 6 concludes and provides possible future work extensions.

Chapter 2

PHYSICAL MODELS FOR SPEECH SYNTHESIS

Speech synthesis has received immense interest since the discovery of electricity. It has come a long way from nascent strategies to replicate voice by synthesizing robotic sounds [31], to human-like voice synthesis [34]. Studying speech synthesis in the context of articulators is important for applications in communication, health, and automation. A robust method for speech inversion can help one understand the physics governing articulatory trajectories and thus improve speech coding methods [24]. Articulatory knowledge can lead to new studies on the adaptation of conversational behavior between two speakers to their interlocutors [35]. This adaptation between speakers, called speech entrainment, could help reveal learning habits of second (L2) or third (L3) languages in individuals [36]. Articulatory knowledge is also a vital key for speech therapy in patients suffering from dysarthia, stuttering, or other neurological speech disorders [16]. One efficient method to improve both synthesis and analysis of speech is to use physical models of speech and compare synthesized signals to the original speech signals

2.1 Speech Production

Different models have been considered for speech production. Figure 2.1 depicts some selected elements of various voice production models. The color codes depict various elements in each model: glottis is shown in red, passive structures in grey, and articulators in blue. In a microphone recording, acoustic waves radiated from the nostrils and mouth are recorded as data. During an utterance of a voiced phoneme, air pushed from the lungs travels along the vocal folds (glottis). At the intersection between the glottis and the vocal tract, air is modulated by the vocal folds creating a modulated acoustic source. This traveling air source is re-modulated by resonances and anti-resonances of the oral, nasal and vocal cavities. As the air radiates from the vocal tract, it is affected by impedances of the lip and nostrils until it is captured by a microphone. A recording played through the speakers needs noise compensation to remove any channel induced noise that may have altered the radiated acoustic wave. For a physiological model, as the one shown in Figure 2.1(a), voice is synthesized based on the physical model of the system and the mechanical properties of the vocal fold/tract. Generally, the physics and mechanical properties are represented as ordinary differential equations [37].

Figure 2.1(b) depicts an acoustic model of the vocal tract. This model simplifies the physiology of the vocal apparatus (including the vocal folds) using approximate solutions of ordinary differential equations in the physiological model. The larynx and other articulators are simplified into impedance functions that are dependent on the area of opening (usually as sections). In [32], the glottal flow (the air travelling through the glottal area) is derived as a function of the glottal area and subglottal pressure. Since it corresponds to the air flowing through a designated glottal area, the glottal flow is an implicit function of the glottal area [8]. On the contrary, some models consider the glottal area to be an implicit function of the glottal flow [9]. An implicit formation of speech that strongly adheres to the mechanical properties of the vocal folds may be represented by a single variable, a set of area sections called vocal tract area (VTA) function.

Figure 2.1(c) depicts a well known source-filter model. This model assumes the absence of any coupling between the vocal tract and the driving glottal source. Due to the no coupling assumption, voice is considered to be due to a periodic pulse train applied to a vocal tract filter (VTF). Mathematically, the voiced phonemes are then represented by

$$S(\omega) = G(\omega)V(\omega)L(\omega)\,, \tag{2.1}$$

where $G(\omega)$ is the spectrum of acoustic excitation at the glottis; $V(\omega)$ is the spectrum of the vocal tract that merges all the vocal tract physiology, which is generally an all pole filter with peaks (formants); and $L(\omega)$ is a single filter that merges the mouth and nostril radiations. Together, these three responses form speech. This model is simple and stable to use; it also has tractable error measurements and spectral estimation properties. Hence, it is widely used in a range of applications.

It is important to differentiate between a source-filter model and an acoustic model. A source-filter model aims to manipulate the perceived elements of the voice by an all-pole estimate. This manipulation provides an easy approach to analyze speech with relatively high accuracy. An acoustic model, On the other hand, may be able to reproduce, both analytically and numerically, the voice production as closely as possible to the physical measurement; this can be useful when studying the mechanical behavior of vocal folds, and when studying the different levels of coupling between the acoustic source and the vocal tract impedance. As this study matches with the latter objective, that of obtaining the articulatory parameters, we consider an intermediate model. This is called the chain matrix method or hybrid time-frequency synthesizer and consists of both articulatory parameters and perceptual based analysis [32].

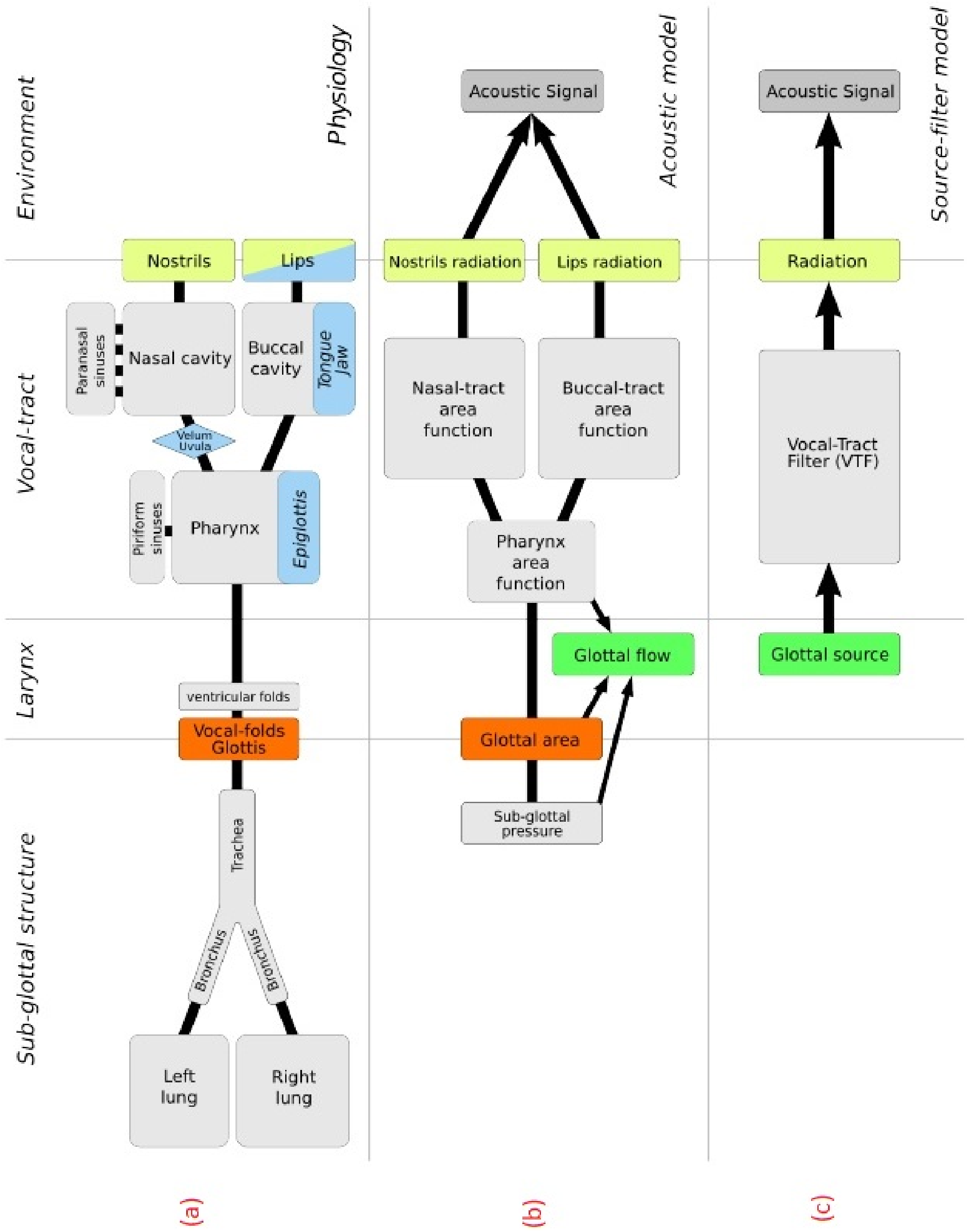


Figure 2.1: Different voice production models: (a) physiological model, (b) acoustic model; (c) source-filter model [1].

2.2 Physiology of Glottis

The physiology of glottis enables a periodic/aperiodic assumption of speech. When a voiced phoneme is sustained, it is observed to have periodic components in time. During this duration, the glottis has three phases. A resting time interval, called the closed phase, an opened time duration called the open phase, and a recovery period, called the return phase. The vocal folds are at rest during the closed phase. The open phase is the duration when vocal folds expand and allow the passage of air. As the thyroarytenoid (TA) and cricothyroid (CT) muscles lose energy, they contract to their original position and create the return phase. This completes one period of the glottal signal during a voiced phoneme. A typical speech waveform is shown in the time domain in Figure 2.2(a); its spectrogram time-frequency representation is shown in Figure 2.2(b) to have high peak frequencies in the periodic section. Figure 2.2(c) shows a time domain close up representation of the of the vowel a.

Figure 2.2(a) reveals an inside view of the glottis. Ideally, the vibration of the vocal folds generates an acoustic source. However, in reality, this vibration may be caused by numerous agents. Similarly, the vocal tract harmonics result from articulatory configurations that have a many-to-one mapping.

A first step towards articulatory speech synthesis is choosing an appropriate glottal source model. There exist many physiological glottal models [38]. However, to ascertain stability and reduce computational load, we consider parametric glottal sources. Various methods have been proposed in the literature to define analytically one period of glottal flow [33, 39–42]. The glottal flow, however, is well defined from deterministic components [33]. In particular, the glottal flow $g(t)$ is formed when an analytical curve passes conditions of open and closed phase intervals. The set of parameters that define these phases are listed below.

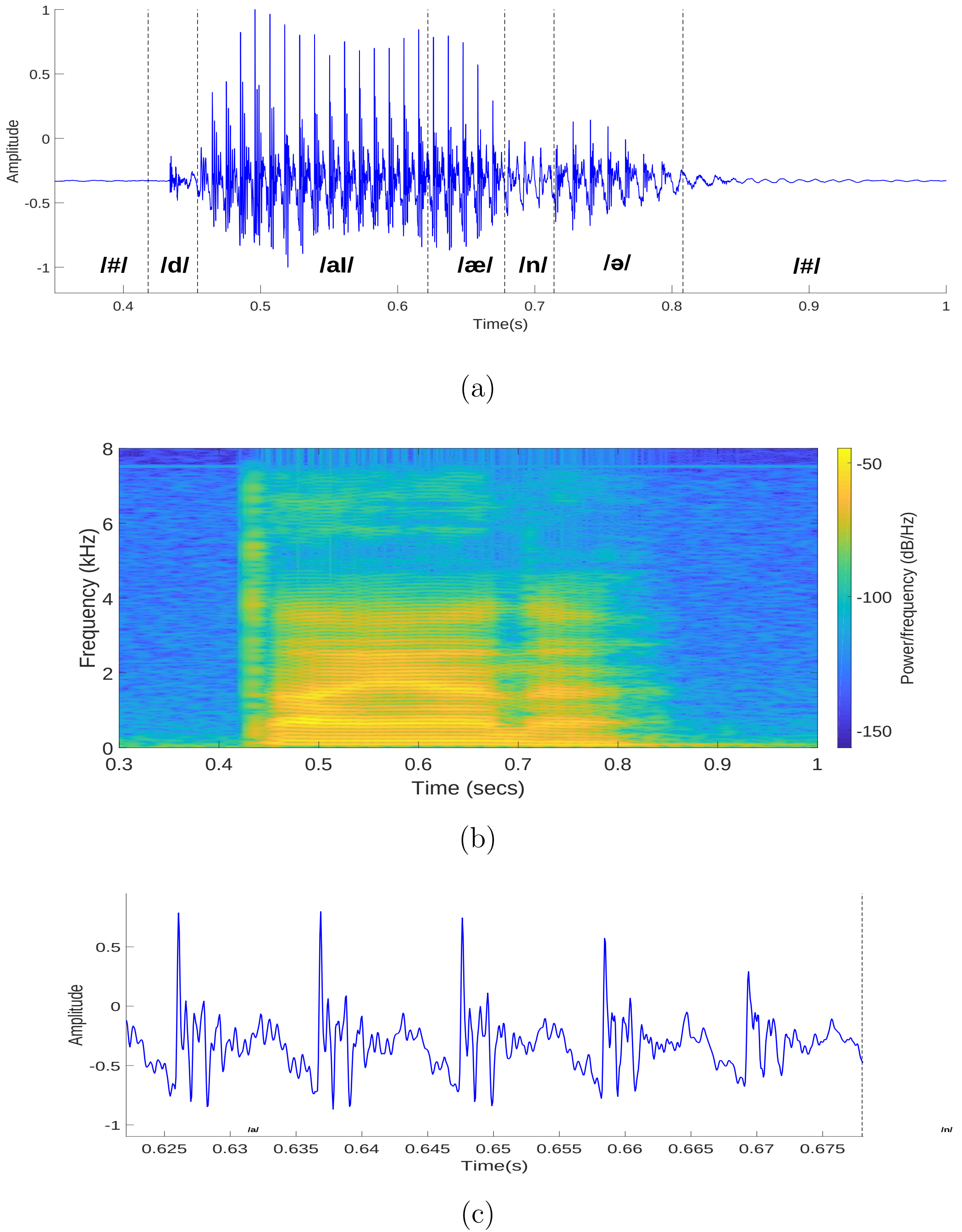
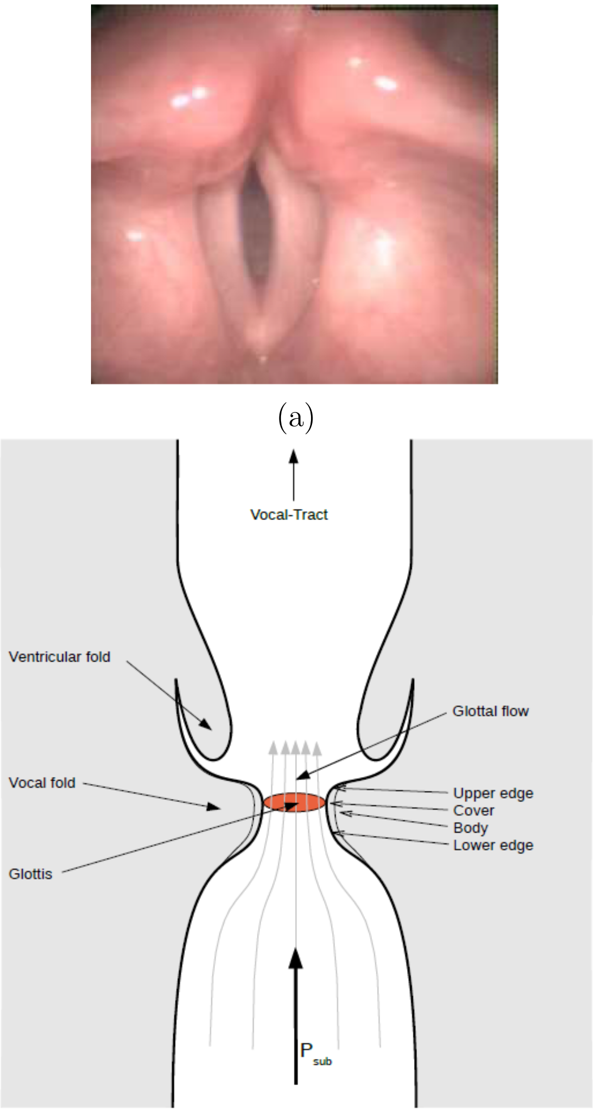


Figure 2.2: Phonetically annotated speech waveform representations [2]: (a) time domain; (b) spectrogram, using a 512 length window with 480 sample overlap; (c) time domain close up of a vowel *a*.


(a)

(b)

Figure 2.3: (a) Diagram of a vertical cut of the vocal folds; (b) High-speed videoendoscopic image of the larynx, taken from the oropharynx in the direction to the larynx. The top of the image corresponds to the back (posterior) of the larynx, and the glottis is the dark area in the center, which is delimited by the vocal folds [1].

- t_0: Time that corresponds to the start of a pulse in a voiced phoneme; this relates to an integer multiple of the pitch period F_0. In this work, it is assumed that $t_0 = 0$.

- t_p: Time that corresponds to the start of the closing phase; the time of the first zero crossing after reaching the maximum amplitude of the voicing E_0 under an acceleration of α and ω.

- t_e: Time instance of the maximum glottal flow derivative E_e.

- t_a: Time that corresponds to the start of the closed phase, assuming that the rest time for the glottal folds has a recovery rate ϵ.

- t_c: Time elapsed for one pulse radiation.

- N_0: The period of one pulse, called pitch, where the fundamental period is $T_0 = 1/N_0$.

A detailed explanation of these parameters and their properties follows.

2.2.1 Closed Phase, Open Phase, Return Phase

The glottal source may be subdivided into three main phases. The air from the lungs moves towards the vocal folds, as the pressure changes between both sides of the vocal folds. The air pushes the vocal folds to open and release into the vocal tract. The period when the vocal folds open and release air into the vocal tract causing a rise in the air pressure is called the open phase.

Towards the end of the open phase, pressure across the vocal tract and the subpraglottal (behind the vocal folds) equalizes. This collapses the vocal folds into a brief period of closure, called the return phase. The resulting closure of the vocal folds allows the vocal folds to rest, as the pressure for the next pulse builds up. This brief period of resting is called the closed phase. The pressure expelled from inside the lungs is the period of the non-nasalized vowels. Figure 2.4 depicts the three phases in a glottal pulse cycle, together with its timing parameters.

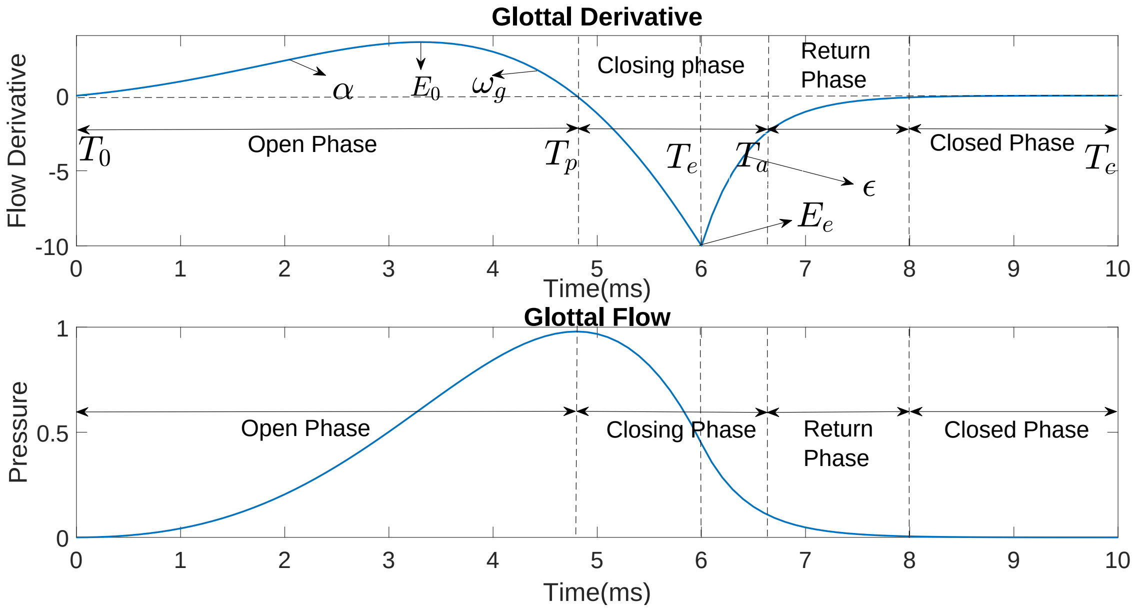


Figure 2.4: Phases of glottal flow and its derivative.

2.2.2 Excitation Amplitude, Shimmer, Pitch Period, Duration and Jitter

The maximum amplitude of the time-derivative of the glottal pulse at time t_e is denoted by E_e. In our study, we prefer to characterize the amplitude excitation of the glottal model by this value instead of the voicing amplitude E_0 (the maximum amplitude of the glottal pulse at time t_p). In any natural voice, this pulse amplitude is never perfectly constant. The inherent variations, termed shimmer, reveal voice quality and provide uniqueness to individuals. Consequently, an amplitude modulation of the glottal source always exists. In this study, we assume that this modulation is negligible inside a short window of observing speech ($\approx$ 3 periods). However, a variation would only increase the variance of the noise that would otherwise describe a perfect glottal pulse.

As per empirical evidence, a voiced speech signal has two main quasi-periodic pulses, each with duration T_0. This fundamental period of the glottal pulse is called pitch and is denoted by N_0. A periodic source is necessary in many contexts, such as singing, voicing and nasals. However, these pulses can be irregular when the pressure in the lungs varies or the atmospheric temperature changes or due to vocal fold fatigue. Variations in pitch or jitter across multiple analysis windows exist in natural voice. These irregularities add to a natural and healthy voice within an acceptable transiency. The analysis window used has to be short enough to model fast variations of the fundamental frequency.

2.2.3 Shape Parameters: Glottal Closure Instants, Glottal Opening Instant

The glottal opening instant (GOI) corresponds to the start of the open phase. The glottal pulse starts to increase when compared to its minimal value, which is generally taken to be zero. The glottal closure Instant (GCI) corresponds to the minimum of the time derivative of the pulse. This instant is not symmetrical to the GOI. Therefore, the instant when the glottal pulse reaches the minimum value of the pulse (t_c) is referred to as the effective closure instant.

2.2.4 Effective Duration of GOI and GCI

Additional parameters are used to control the shape of the pulse, and in particular to normalize the pulse's duration, amplitude and excitation amplitude. These parameters are as follows.

- Open Quotient (OQ): this is the duration from the GOI to the GCI, normalized by the pulse period, $\mathrm{OQ} = t_e/T_0$. Even though the glottal pulse can be larger than zero during the return phase, this phase is not considered in OQ. The sum of the return phase and the open phase is called the effective open quotient.

- Asymmetry: this is the skewness of the pulse and it is given by $\alpha = t_p/t_e$. The closer this value is to 0.5, the more symmetric is the pulse.

- Return phase: this is the duration of the return phase, normalized by the pulse duration, $Q_a = t_a/T_0$. It is used to represent how abrupt the closure is; the smaller this duration, the more abrupt the closure.

2.2.5 Spectral Properties: Glottal Formant and Spectral Tilt

The glottal pulse has a peak in its amplitude spectrum, called the glottal formant because of its similarity to the shape of the vocal-tract formants. This glottal formant is characterized by frequency F_m, which is the frequency that corresponds to the maximum of the amplitude spectrum of the time derivative of the pulse. This frequency is not easy to determine, and it depends on the analytical form of the selected glottal model. The glottal formant is also characterized by the frequency that corresponds to the maximum value of stylization of the amplitude spectrum.

2.3 Glottal Parametric Models

Various glottal parametric models are presented next, that translate timing parameters into pulses.

2.3.1 Rosenberg Model

Rosenberg initially proposed six models to fit a pulse estimated by inverse filtering [39]. The model found to best fit the glottal source consists of two polynomial parts and is given by

$$g(t) = \begin{cases} t^2(t_e - t), & 0 < t < t_e \\ 0, & t_a < t < T_0 \end{cases}$$

where, $t_e = t_a$. This model has only one shape parameter, t_e, the instant of closure; the instant of maximum flow is proportional to $t_p = \frac{2}{3}\, t_e$.

2.3.2 Klatt Model

The Klatt glottal pulse model is similar to the Rosenberg; it has only two shape parameters, an open quotient (OQ) and a spectral tilt parameter. The model is given by

$$g(t) = a\, t^3 - b\, t^2,\ 0 \leq t \leq T_0$$

where a/b is a ratio of the time of the opening to spectral tilt. The spectral tilt is not explicitly modeled here, unlike in KLGLOTT88 [43]. The model is mainly used in the KLSYN88 synthesizer [40].

2.3.3 Fant Model

This is the first version of the model proposed by Fant, and it consists of two sinusoidal parts, [41]

$$g(t) = \begin{cases} \frac{1}{2}(1 - \cos(\Omega\, t)), & 0 < t < t_p \\ K\, \cos(\Omega(t - t_p)) - K + 1, & t_p < t < t_c = t_p + \Big(\arccos(\frac{K+1}{K})\Big)/\Omega \\ 0, & t_c < t < T_0 \end{cases}$$

where $\Omega = \pi/t_p$. This model has two shape parameters, t_p and K, that control the slope of the descending branch. When $K = 0.5$, the pulse is symmetric. When $K \geq 1$, then $t_e = t_a$.

2.3.4 Liljencrant-Fant Model

The Liljencrants-Fant (LF) model is an acoustic model of the glottal source derivative [33]. The LF model is an extension of the Fant model, with curvature and acceleration parameters. It is given by

$$g(t) = \begin{cases} E_0\, e^{\alpha\, t} \cos(\Omega t), & t_0 \leq t \leq t_e \\ -\dfrac{E_e}{\epsilon\, t_a}\Big(\exp\big(-\epsilon\,(t - t_e)\big) - \exp\big(-\epsilon\,(t_c - t_e)\big)\Big), & t_e < t \leq t_c \\ 0, & t_c < t \leq T_0 \end{cases} \tag{2.2}$$

where α and ϵ are acceleration parameters, and $\Omega = \pi/t_p$ is the curvature before reaching the amplitude of voicing. Also,

$$\begin{aligned} \epsilon\, t_a &= 1 - \exp\big(-\epsilon\,(t_c - t_e)\big) \\ E_0 &= \frac{E_e}{e^{\alpha\, t_e}\, \sin(\Omega\, t_e)} \end{aligned}$$

Note that the following constraint must be satisfied

$$\int_0^{T_0} g'(t)\, dt = 0$$

where $g'(t) = \frac{d}{dt} g(t)$ is the derivative of the glottal flow. Also, E_e can be estimated instead of E_0 due to the strong dependence on t_e. This continuous time representation can be easily discretized using $g[n] = g(nT_s)$, where $f_s = 1/T_s$ is the sampling frequency.

This model has been extensively studied for both its time and spectral properties [33, 44] and serves as an appropriate model of glottis to couple with an articulatory synthesizer.

2.4 Vocal Tract Response

The interaction between the vocal tract (VT) and vocal folds can be described in innumerable ways. There exist many models for the vocal tract [9, 45–47]. Broadly, these can be classified into three categories: (i) ARMA models, (ii) physiological models, and (iii) acoustic models. The linear prediction (LP) model of speech considers the vocal tract as an all pole filter model, as shown in Figure 2.5. In our study, we use the physiological model. One prime consideration in the physiological models are the types of synthesis methods used. These include: (a) the reflection model, that considers propagation in a reflective environment such that the lips are closed; (b) the transmission model, that approximates the vocal tract into RLC circuits; and (c) the vortex model, that considers the air propagating as volumnar flow. Transmission models are simplified into circuit impedances that are clumped together, hence are also known as clumped circuit models, as shown in Figure 2.6.

2.5 Chain Matrix Vocal Tract Model

The chain-matrix model is a preferred approach for computing the spectral response of the VT, given an area function [28]. The equations used to reduce ordinary differential equations into transfer functions can be found in [32]. In this model, the pressure P and volume velocity U are coupled at the input and output of a concatenated acoustic tube. Under strict assumptions that the propagated wave has a planar wavefront, a frequency relation exists and the resulting pressure for the jth section of this tube is

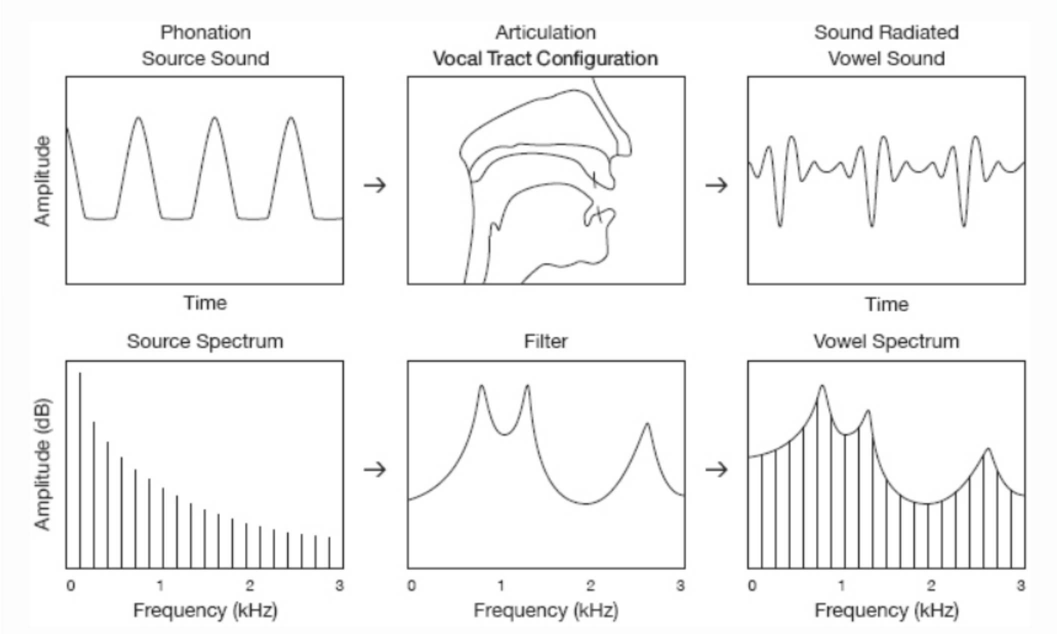


Figure 2.5: The source-filter model for vowel production [3].

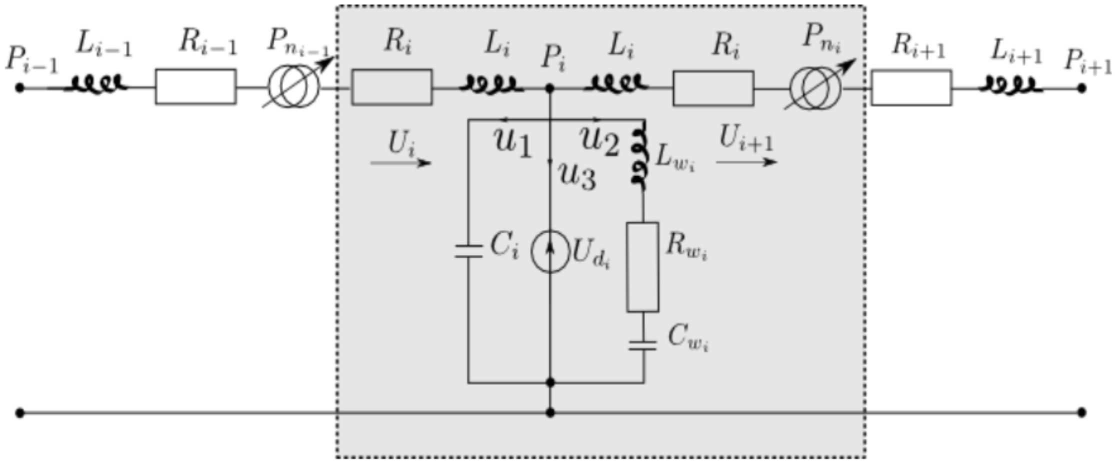


Figure 2.6: The electro-acoustic lumped circuit model of synthesis.

given by

$$\begin{bmatrix} P_{out} \\ U_{out} \end{bmatrix} = \boldsymbol{\psi}^j(\omega; a^j) \begin{bmatrix} P_{in} \\ U_{in} \end{bmatrix} = \underbrace{\begin{bmatrix} \mathcal{A}^j(\omega; a^j) & \mathcal{B}^j(\omega; a^j) \\ \mathcal{C}^j(\omega; a^j) & \mathcal{D}^j(\omega; a^j) \end{bmatrix}}_{\boldsymbol{\psi}^j(\omega; a^j)} \begin{bmatrix} P_{in} \\ U_{in} \end{bmatrix} \tag{2.3}$$

where $\mathcal{A}^j(\omega; a^j), \mathcal{B}^j(\omega; a^j), \mathcal{C}^j(\omega; a^j), \mathcal{D}^j(\omega; a^j)$ are chain matrix (CM) parameters of the tube, *in* and *out* denote the input and output of the tube, a^j is the jth sectional area of the jth articulatory unit inside an S segment tube. The articulatory vector $\mathbf{a} = \begin{bmatrix} a^1 & \ldots & a^S \end{bmatrix}$ represents the articulatory envelope (geometry or configuration). In our simulations, we consider $S = 44$.

The matrix $\boldsymbol{\psi}(\omega; \mathbf{a})$ formed as a result of S uniform tubes (starting at the glottis and ending at the lips) is a product of S individual CMs

$$\begin{aligned} \boldsymbol{\psi}(\omega; \mathbf{a}) &= \boldsymbol{\psi}^S(\omega; a^S)\, \boldsymbol{\psi}^{S-1}(\omega; a^{S-1})\, \boldsymbol{\psi}^{S-2}(\omega; a^{S-2}) \ldots \boldsymbol{\psi}^1(\omega; a^1) \\ &= \begin{bmatrix} \mathcal{A}(\omega; \mathbf{a}) & \mathcal{B}(\omega; \mathbf{a}) \\ \mathcal{C}(\omega; \mathbf{a}) & \mathcal{D}(\omega; \mathbf{a}) \end{bmatrix} \end{aligned}$$

The transfer function of the VT for a non-nasalized vowel can then be shown to be

$$V(\omega; \mathbf{a}) = \frac{U_L}{U_G} = \frac{\mathcal{D}(\omega; \mathbf{a}) Z_L - \mathcal{B}(\omega; \mathbf{a})}{\mathcal{A}(\omega; \mathbf{a}) - \mathcal{C}(\omega; \mathbf{a}) Z_L} \tag{2.4}$$

where U_G and U_L are volume velocities at the glottis and lips, and Z_L is the radiation impedance at the lips, often approximated by that of a pulsating disk of air at the mouth opening [28]. The CM model can also be extended to compute the VT transfer functions of other speech sounds, such as nasals, nasalized vowels, fricatives, and laterals.

In this work, the chain matrix model is assumed to account for all losses due to air viscosity, heat conduction, and yielding tract walls. The CM parameters of a uniform lossy cylindrical tube of area a^j, $j=1,\ldots,44$ and length $l^j=0.37cm^2$, $j=1,\ldots,44$, at frequency ω, is given by:

$$\begin{aligned}
\mathcal{A}^j(\omega; a^j, l^j) &= \cosh\left(\frac{\sigma(\omega)\, l^j}{c}\right) \\
\mathcal{B}^j(\omega; a^j, l^j) &= -\frac{\rho\, c\, \gamma(\omega)}{a^j} \sinh\left(\frac{\sigma(\omega) l^j}{c}\right) \\
\mathcal{C}^j(\omega; a^j, l^j) &= -\frac{a^j}{\rho\, c\, \gamma(\omega)} \sinh\left(\frac{\sigma(\omega)\, l^j}{c}\right) \\
\mathcal{D}^j(\omega; a^j, l^j) &= \cosh\left(\frac{\sigma(\omega)\, l^j}{c}\right)
\end{aligned}$$

where ρ and c are the density of air and speed of sound in air, respectively, as described in [32]. Note that $\gamma(\omega)$ and $\sigma(\omega)$ are independent of the area and the length of the tube. Lastly, to obtain a time domain version of the VT frequency response, we take the inverse discrete-time Fourier transform (DTFT), $\mathcal{F}^{-1}$ of (2.4)

$$\mathbf{v}(\mathbf{a}) = \mathcal{F}^{-1}\left(V(\omega; \mathbf{a})\right)$$

where $\mathbf{v} \in \mathbb{R}^M$ and M is the length of the DTFT. In this study, as we only consider periodic/vowel signals, we do not consider the modeling of fricatives and stop consonants, as shown in [9].

Chapter 3

REVIEW ON SEQUENTIAL BAYESIAN ESTIMATION METHODS

Parameter prediction is a well established statistical problem that allows one to approximate a set of hidden variables that transform into observed data [48]. Typically, observed data or measurements $\mathbf{y_k} \in \mathbb{R}^n$ at time k, depend on implicit variables $\mathbf{x_k} \in \mathbb{R}^m$ either in linear or non-linear manner. A task of estimating the distribution of implicit variables for time k, $\mathbf{x}_k$, forms the basis of a prediction framework. Formally, it is identifying distribution of $\mathbf{x}_k$ in relation to observation $\mathbf{y}_k$: $p(\mathbf{x}_k|\mathbf{y}_k)$, or *posterior* density [49]. Once the density is available, any number of useful estimates can be taken. The prediction of state distribution is found from the Chapman-Kolmogorov equation [49].

$$p(\mathbf{x}_k|\mathbf{y}_{0:k-1}) = \int p(\mathbf{x}_k|\mathbf{x}_{k-1})p(\mathbf{x}_{k-1}|\mathbf{y}_{0:k-1})d\mathbf{x}_{k-1} \tag{3.1}$$

using Bayes' rule, an update of the state distribution may be formed

$$p(\mathbf{x}_k|\mathbf{y}_{0:k}) = \frac{p(\mathbf{y}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{y}_{0:k-1})}{\int p(\mathbf{y}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{y}_{0:k-1})d\mathbf{x}_k} \tag{3.2}$$

In this thesis, problem formulation is based on sequential estimation by Monte Carlo methods. A review of generic state space formulation and solution methodology is provided for completeness.

3.1 Kalman Filter

The Kalman Filter is one of the best analytically tractable linear estimators, under restricted conditions of having a state space model perturbed by Gaussian noise. The filter's origin can be historically traced back to R. E. Kalman (1960), who described that solving a discrete data filtering problem is in essence solving a recursive error relationship between observed data and predictions. A solution may be formed by predicting posterior relationship from a priori states that take a random walk in a single or multi-dimensional space. He described this process in control theory, that paved way to many applications in warfare, stock markets, communications, GNSS, satellite attitude corrections, remote sensing and ballistic tracking [48].

Mathematically, distribution of a normally distributed latent state variable, $\mathbf{X}_k$ which forms a Markov chain, from *a priori* states is given as $p(\mathbf{x}_k|\mathbf{x}_{0:N-1}) = p(\mathbf{x}_k|\mathbf{x}_{k-1})$ such that $\mathbf{X}_k = \mathbf{x}_{0:k-1} = \{\mathbf{x}_0, \ldots, \mathbf{x}_{k-1}\}$. Also given a normally distributed observation values $\mathbf{Y}_k = \mathbf{y}_{0:k-1} = \{\mathbf{y}_0, \ldots, \mathbf{y}_{k-1}\}$ that depend only on $\mathbf{X}_k$, together form a state space model. A joint distribution of $p(\mathbf{x}_0, \ldots, \mathbf{x}_k, \mathbf{y}_0, \ldots, \mathbf{y}_k)$ and marginal distributions $p(\mathbf{x_k}|\mathbf{y}_0, \ldots, \mathbf{y_k})$ are used to predict the latent observation variable. The Kalman filter intelligently combines the observations and predictions based on system dynamic and state models to produce an estimate by reducing mean square error between predicted density and true posterior density [50]. Each time step of the Kalman filter will output a current state estimate $\mathbf{x}_{k|k}$ that is ideal in measure squared sense. A general form of linear state space model is assumed for a Kalman filter:

$$\mathbf{x_k} = F_{k-1}\mathbf{x}_{k-1} + \mathbf{w_k} \tag{3.3}$$

$$\mathbf{y_k} = H_k\mathbf{x_k} + \boldsymbol{\nu}_{\boldsymbol{k}} \tag{3.4}$$

where, $\boldsymbol{\nu}_{\boldsymbol{k}} \sim \mathcal{N}(0, \sigma_\nu^2)$, $\mathbf{w_k} \sim \mathcal{N}(0, \sigma_w^2)$ are observation noise and state noise respectively.

In principle, this state-space may be solved using standard results obtained either through MLE (Maximum Likelihood Estimation), LS (Least Squares) and their respective variations [48]. However, the Kalman Filter approach embraces on an idea of non-extant perpetual priori data, i.e. data is never stored and only available from prior time $k-1$. In a time series analysis this is highly attractive since it may not be feasible to store data at all times [51]. The Kalman filter uses data available only at previous time step to predict data at next time step by propagating the prior density based on a gain, the Kalman gain. The *a priori* estimates $\mathbf{x}_{k-1}, \Sigma_{w_{k|k-1}}$ update *a posterior* estimate $\mathbf{x}_k, \Sigma_{w_{k|k}}$ by minimizing the likelihood $p(\mathbf{y_k}|\mathbf{x_k})$. The operation of the Kalman filter has two recursive steps:

- Predict: The prediction process projects forward in time and obtains a priori estimate at next time step with error covariance P.

$$\hat{\mathbf{x}}_{k|k-1} = F_{k-1}\hat{\mathbf{x}}_{k-1|k-1} \tag{3.5}$$

$$P_{k|k-1} = \Sigma_{w_{k-1}} + F_{k-1}P_{k-1|k-1}F_{k-1}^T \tag{3.6}$$

$$S_k = H_k P_{k|k-1} H_k^T + \Sigma_{\nu_k} \tag{3.7}$$

$$K_k = P_{k|k-1} H_k^T S_k^{-1} \tag{3.8}$$

- Update: Incorporates the new observation $\mathbf{y}_k$, into the *a priori* estimate to obtain an improved *a posteriori* estimate.

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + K_k(\mathbf{y_k} - H_k\hat{\mathbf{x}}_{k|k-1}) \tag{3.9}$$

$$P_{k|k} = P_{k|k-1} - K_k H P_{k|k-1} \tag{3.10}$$

A detailed derivation along with an algorithm to implement the Kalman Filter for time series data may be found in [49].

3.2 Extended Kalman Filter

The Kalman filter fails miserably in non-linear transitions, for ex: predicting the flight path of a bee. This task is challenging due to non-linear dependence with turn rate, acceleration, Earth's gravitational laws and Coriolis effect. These non-linear dependencies break assumptions made using the Kalman filter and restrict its general use. In such situations, where models are non-linear an alternative method is devised with sub-optimal performance, called the *Extended* Kalman Filter (EKF). The idea is to naively linearize the model using Taylor expansion [51]. It may be sufficient to describe non-linearity in a Jacobian matrix, allowing us to use a Kalman filter. Consider the general state space model:

$$\mathbf{x}_k = \mathbf{f}_{k-1}(\mathbf{x}_{k-1}, k) + \mathbf{w}_k \tag{3.11}$$

$$\mathbf{y}_k = \mathbf{h}_k(\mathbf{x}_k, k) + \boldsymbol{\nu}_{\boldsymbol{k}} \tag{3.12}$$

To obtain a linear approximation of $\mathbf{f}_k(\mathbf{x}_k, k)$ about the value $\mathbf{x}_k$ we drop all but constant and linear terms in the Taylor expansion:

$$\mathbf{f}_k(\mathbf{x}_k, k) \approx \mathbf{f}_k(\mathbf{x}_k^R, k) + (\mathbf{x}_k - \mathbf{x}_k^R)\frac{\partial \mathbf{f}_k(\mathbf{x}_k, k)}{\partial \mathbf{x}_k}\bigg|_{\mathbf{x}_k = \mathbf{x}_{k|k}^R} + \mathcal{O}(2) + \dots \tag{3.13}$$

where $\mathbf{x}_k^R$ is some reference trajectory and expansion maybe written as first order terms in a Jacobian of $\mathbf{f}_k(.)$, defined as:

$$F_k = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix} \tag{3.14}$$

If we assume the functions to be time invariant we have $\dot{F} = F_{k-1} = F_k$ and $\dot{H} = H_{k-1} = H_k$. This simplifies the prediction and update steps to those obtained through Kalman filtering:

- Predict: The prediction process projects forward in time and obtains a priori estimate at next time step with error covariance P

$$\hat{\mathbf{x}}_{k|k-1} = \dot{F}\mathbf{x}_{k-1|k-1} \tag{3.15}$$

$$P_{k|k-1} = \Sigma_w + \dot{F}P_{k-1|k-1}\dot{F}^T \tag{3.16}$$

$$S_k = \dot{H}P_{k|k-1}\dot{H}^T + \Sigma_{\nu_k} \tag{3.17}$$

$$K_k = P_{k|k-1}\dot{H}^T S_k^{-1} \tag{3.18}$$

- Update: Incorporates the new observation $\mathbf{y}_k$, into *a priori* estimate to obtain an improved *a posteriori* estimate

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + K_k(\mathbf{y_k} - \dot{H}\hat{\mathbf{x}}_{k|k-1}) \tag{3.19}$$

$$P_{k|k} = P_{k|k-1} - K_k\dot{H}P_{k|k-1} \tag{3.20}$$

The sub-optimality of EKF is evident when a) functions are not analytical and hard to form a Jacobian b) when non-linear transformation severely alters the statistical randomness into non-Gaussian distributions. The EKF still performs resonably well and is still widely used in many physical applications such as biological networks, chemistry, stock markets, navigation systems etc. [51].

3.3 Unscented Kalman Filter

The EKF captures mean and covariance upto first order term and propagates it through the non-linear dynamics. This approximation can be improved if a minimal set of sample points can be carefully chosen to capture the true mean and covariance of the Gaussian random vectors [49]. The unscented transformation (UT) is a statistical method to calculate the statistics of a random variable which undergoes a non-linear transformation through a set of sigma points [52]. Since UT no longer imposes a requirement to compute Jacobian(s) for $\mathbf{f}_k(\mathbf{x}_k)$ and $\mathbf{h}_k(\mathbf{x}_k)$ in the dynamic state-space. This method proves advantageous with reduced restrictions for signals to be analytic albeit the noise to be still Gaussian.

Unscented Transform

Assume $\mathbf{x}_k \in \mathbb{R}^n$ that undergoes a nonlinear transformation $\mathbf{y}_k = \mathbf{h}_k(\mathbf{x}_k)$, with mean $\bar{\mathbf{x}}$ and covariance $\mathbf{P}_{\mathbf{x}}$. Initialize a sigma vector $\chi \in \mathcal{R}^{2N+1}$ with N sigma points, χ_i is the sigma weight associated with Nth random state [52].

$$\chi_0 = \mathbb{E}[x] \tag{3.21}$$

$$\chi_i = \mathbb{E}[x] + \left(\sqrt{(N+\lambda)\mathbf{P}_x}\right)_i, \qquad i = 1,\ldots,N \tag{3.22}$$

$$\chi_i = \mathbb{E}[x] - \left(\sqrt{(N+\lambda)\mathbf{P}_x}\right)_{i-N}, \qquad i = N+1,\ldots,2N \tag{3.23}$$

$$W_0^{(m)} = \frac{\lambda}{N+\lambda} \tag{3.24}$$

$$W_0^{(c)} = \frac{\lambda}{N+\lambda} + (1-\alpha^2+\beta) \tag{3.25}$$

$$W_i^{(m)} = W_i^{(c)} = \frac{1}{2(N+\lambda)}, \qquad i = 1,\ldots,2N \tag{3.26}$$

where $\lambda = \alpha^2(N+\kappa) - N$ is a scaling parameter, α determines the spread of the sigma points around $\mathbb{E}[\mathbf{x}]$ and is usually set to a small positive value, κ is a secondary scaling parameter which is usually set to 0, β is used to incorporate prior knowledge of the distribution of $\mathbb{E}[\mathbf{x}]$ for Gaussian distributions, $\beta = 2$ is optimal, and $\left(\sqrt{(N+\lambda)P_{\mathbf{x}}}\right)_i$ is the i^{th} row of the matrix square root. These sigma vectors are propagated through the nonlinear function [49].

$$\mathcal{Y}_i = \mathbf{h}_k(\chi_{\mathbf{i}}), \qquad i = 0,\ldots,2N \tag{3.27}$$

and the mean and covariance of $\mathbf{x_k}$ are approximated using a weighted sample mean and covariance of the posterior sigma points,

$$\bar{\mathcal{Y}} = \mathbb{E}[\mathbf{y}] = \sum_{i=0}^{2N} W_i^{(m)} \mathcal{Y}_i \tag{3.28}$$

$$P_{\mathbf{y}} = \sum_{i=0}^{2N} W_i (\mathcal{Y}_i - \bar{\mathcal{Y}})(\mathcal{Y}_i - \bar{\mathcal{Y}})^T \tag{3.29}$$

$$\bar{\mathcal{X}} = \mathbb{E}[\mathbf{x}] = \sum_{i=0}^{2N} W_i^{(m)} \mathcal{X}_i \tag{3.30}$$

$$P_{\mathbf{x}} = \sum_{i=0}^{2N} W_i (\mathcal{X}_i - \bar{\mathcal{X}})(\mathcal{X}_i - \bar{\mathcal{X}})^T \tag{3.31}$$

These estimates of mean and covariance are accurate upto third order for Gaussian priors for any non-linear function expanded using Taylor series. Errors introduced may be scaled by the parameter κ. UKF is widely used in many mechanical problems and is reported to be successful [52].

3.4 Overview of Monte Carlo Methods

Monte Carlo(MC) methods were invented in the late 1940s to evaluate complex and often intractable integrals [49]. Integrals like:

$$I = \int_{x_0}^{x_1} f(x)dx = \int_{x_0}^{x_1} h(x)p(x)dx = \mathbb{E}[h(x)] \tag{3.32}$$

for $f : \mathbb{R}^n \mapsto \mathbb{R}^n$. It was suggested that to evaluate such integrals a set of pseudo-random number generators could be used in such a way that one could decompose $f(x) = h(x)p(x)$ where $p(x)$ is a valid PDF with domain $p(x) = \{x : x_0 \leq x_0 \leq x_1\}$ that we can draw samples from. To obtain sample mean, we need $p(x)$ or we can sample N_p independent random variables from $\mathbf{x}$ such that $\{\mathbf{x}^{(i)}\}_{i=1}^{N_p}$, then by central limit theorem, first moment of $p(x)$ approaches empirical measure as:

$$\frac{1}{N_p}\sum_{i=1}^{N_p} x^{(i)} \xrightarrow{N_p>>1} \mathbb{E}[\mathbf{x}] \tag{3.33}$$

A general assumption that could be thought of is sampling the data would converge the estimated mean to it's true mean, such that:

$$\frac{1}{N_p}\sum_{i=1}^{N_p} h(x^{(i)}) \xrightarrow{N_p>>1} \mathbb{E}[h(x)] \tag{3.34}$$

Hence, we may approximate the integral from all N_p independent samples.

$$I \approx \frac{1}{N_p}\sum_{i=1}^{N_p} h(x^{(i)}) \tag{3.35}$$

Behind this general principle of Monte Carlo we can establish a ground basis for a class of distributions. Often a limitation of standard MC integration technique arises when sampling from a very complex $p(x)$. In such cases where it may not be possible to directly sample from $p(x)$ and in these cases we resort to utilizing Markov chain properties.

A Markov process/model is one which directly depends on the previous value(s) of the random variable $\mathbf{x}$. The temporal dependence is called order of Markov process and determines maximum time dependence of a random variable, such as $k|k-1$ is said to be first order. Let $p(\mathbf{x}_k|\mathbf{x}_{k-1})$ be prior distribution to a first order Markov process such that all current values of $\mathbf{x}$ depend only on previous time instance. We may then represent this Markov chain in its traditional stochastic matrix $\mathbf{K}$.

Given an initial distribution vector $\pi_0 \in \mathbb{R}^n$ with $\pi_0\mathbf{1}^T = 1$. We can determine the probability distribution at time step k as:

$$\pi_k = \pi_{k-1}\mathbf{K} = \pi_0\mathbf{K}^k \tag{3.36}$$

this is a discrete version of the Chapman-Kolmogorov equation. The equilibrium distribution of a Markov chain is the distribution vector π_e, such that:

$$\pi_e = \pi_e\mathbf{K} \tag{3.37}$$

and π_e can be found from solving the eigenvalue problem

$$\mathbf{y}_m^T(\lambda_m\mathbf{I} - \mathbf{K}) = 0 \tag{3.38}$$

$$\underset{i}{max} \quad \lambda_i(\mathbf{K})_{i=1}^n = \lambda_m = 1 \tag{3.39}$$

where for a stochastic matrix, the maximum eigenvalue is 1. Therefore, the left eigenvector of $\mathbf{K}$ with maximum eigenvalue of $\lambda = 1$ is the equilibrium distribution π_e. Such a vector is scale invariant even when being multiplied with a transition kernel matrix $\mathbf{K}$, as shown in (3.38).

3.5 Particle Filters

The basis of particle filtering methods lies in sequentially updating a distribution using importance sampling techniques. One particle filtering method is the sequential importance sampling (SIS) method, introduced in [49, 53]. SIS involves using importance sampling to solve the recursion equation.

To begin, let us describe our model as

$$\mathbf{x}_k = \mathbf{f}_k(\mathbf{x}_{k-1}, \mathbf{w}_k) \tag{3.40}$$

$$y_k = \mathbf{h}_k(\mathbf{x}_k, \eta_k) \tag{3.41}$$

where we choose state noise $\mathbf{w}_k$ to be white Gaussian $\mathbf{w}_k \sim \mathcal{N}(0, \sigma_v^2)$, similarly we choose dynamic noise η_k to be $\eta_k \sim \mathcal{N}(0, \sigma_\eta^2)$. Our aim is to find a posterior distribution $p(\mathbf{x}_k|\mathbf{y}_{0:k})$ from prior distribution $p(\mathbf{x}_k|\mathbf{x}_{k-1})$ and observation density $p(\mathbf{y}_k|\mathbf{x}_k)$. An estimate of the state can be determined for any performance criterion and filtering distribution. The distribution of interest is the marginal or joint distribution of the latent variables at time k, given all observations up to that point.

$$p(\mathbf{x_k}|\mathbf{y}_{0:k}) = \frac{p(\mathbf{x_k}|\mathbf{y}_{0:k-1})p(\mathbf{y_k}|\mathbf{x_k})}{\int p(\mathbf{x_k}|\mathbf{y}_{0:k-1})p(\mathbf{y}_k|\mathbf{x_k})d\mathbf{x_k}} \tag{3.42}$$

Furthermore, the predictive distribution can be expressed as:

$$p(\mathbf{x}_k|\mathbf{y}_{0:k-1}) = \int p(\mathbf{x}_k|\mathbf{x}_{k-1})p(\mathbf{x}_{k-1}|\mathbf{y}_{0:k-1})d\mathbf{x}_{k-1} \tag{3.43}$$

The basis of a particle filter is to draw a sufficient number of particles, such that the pdf of the likelihood $p(.)$ is approximated by the probability mass function (PMF).

$$p(\mathbf{x}) = \sum_{i=0}^{N_p} \mathbf{w}_k^{(i)} \delta(\mathbf{x}_k - \mathbf{x}_k^{(i)}) \tag{3.44}$$

where the particle weight $\mathbf{w}_k^{(i)} \propto \pi(.)/q(.)$. In case of a state space modeling the recursive weight update equation approximates the PDF and defined as:

$$\mathbf{w}_k^{(i)} = \mathbf{w}_{k-1}^{(i)} \frac{p(\mathbf{y}_k|\mathbf{x}_k^{(i)})p(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)})}{q(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)}, \mathbf{y}_k)} \tag{3.45}$$

This leave the update equation as

$$\mathbf{w}_k^{(i)} = \mathbf{w}_{k-1}^{(i)} p(\mathbf{y}_k|\mathbf{x}_k^{(i)}) \tag{3.46}$$

It is proved in [49], that the variance weights $\mathbf{w}_k^{(i)}$ will increase with time k. However, after a few iterations almost all of the normalized weights will be very small and causes loss of convergence. This problem is solved using a technique known as resampling as described in [49]. A sampling importance sampling particle filter is described in algorithm 1.[1]

[1] For sake of simplicity as well as tractability we assume all models to have AWGN (additive white Gaussian noise) and the prior knowledge about $\mathbf{x}_0$ given by $p(\mathbf{x}_0)$.

Algorithm 1: Sequential Importance Resampling

```
begin
    // Initialize
    forall *particles* $p = 1, \dots N_p$ do
        Draw $\mathbf{x}_{p,0} \sim \pi(\mathbf{x_k})$ from an initial prior distribution
    end
    for $k \leftarrow 1, \dots N - 1$ do
        forall *particles* $p = 1, \dots N_p$ do
            // Correct
            $w_{p;k} = w_{p;k-1} \frac{p(\mathbf{y}_k|\mathbf{x}_k^{(p)})p(\mathbf{x}_k^{(p)}|\mathbf{x}_{k-1}^{(p)})}{q(\mathbf{x}_k^{(p)}|\mathbf{x}_{k-1}^{(p)},\mathbf{y}_k)}$
        end
        $w_p \leftarrow w_p\{\Sigma_p w_p\}^{-1}$;        // Normalize
        $\hat{\mathbf{x}}_k \leftarrow \Sigma_p w_p \mathbf{x}_p$;        // Estimate
        $\mathbf{x}_p \leftarrow R(w_p, \mathbf{x}_p)$;        // Resample
        forall *particles* $p = 1, \dots N_p$ do
            //Predict
            $\mathbf{x}_{p,k} \sim \pi(\mathbf{x}_k)$
            Propagate $\mathbf{x}_p = f(\mathbf{x}_p, \mathbf{e}_k)$
        end
    end
end
```

Chapter 4

ESTIMATION OF GLOTTAL SOURCE AND VOCAL TRACT DYNAMIC MODEL PARAMETERS

4.1 State Space Formulation of Glottal Source and Vocal Tract Model

4.1.1 Glottal Source and Vocal Tract Model State Parameters

In Chapter 2, we presented various parametric glottal source models and an articulatory model that results in a vocal tract transfer function that is biologically coupled to the glottal source. Considering the problem of speech decomposition of a non-nasalized vowel, we devise a dynamic state space formulation for the models of the two speech generation components. The formulation is highly nonlinear, as it is based on an acoustic parametric model of the glottal source and a physiological based model for the vocal tract response. In addition, the unknown time-varying state parameters to be estimated have high dimensionality. Note that solving problems in dynamic state-space formulations can provide estimates of the model parameters at each time step [54, 55]. Such estimation formulations have been applied in functional magnetic-resonance imaging (fMRI) applications [56] and in biological networks [57].

The dynamic state-space formulation is given by

$$\begin{aligned} \mathbf{x}_k &= \mathbf{f}_{k-1}\big(\mathbf{x}_{k-1}, \mathbf{w}_{k-1}\big) && (4.1) \\ y_k &= \mathbf{h}_k(\mathbf{x}_k) + \eta_k\,. && (4.2) \end{aligned}$$

In our formulation, the unknown parameter state vector $\mathbf{x}_k$ at time step k consists of all the unknown glottal source model parameters and vocal tract response model parameters. In particular, the state (row) vector is defined as

$$\mathbf{x}_k = \Big[\boldsymbol{\theta}_k \;\; \mathbf{g}_k(\boldsymbol{\theta}_k) \;\; \mathbf{a}_k \;\; \mathbf{v}_k(\mathbf{a}_k) \;\; \mathbf{C}_k\Big], \tag{4.3}$$

where $\boldsymbol{\theta}_k$ and $\mathbf{g}_k(\boldsymbol{\theta}_k)$ are parameters of the glottal source model, $\mathbf{a}_k$ and $\mathbf{v}_k(\mathbf{a}_k)$ are parameters of the vocal tract response model, and $\mathbf{C}_k$ is a covariance matrix for both models.

In more detail, using the Liljencrants-Fant (LF) glottal source parametric model described in Section 2.3.4, the (1×4) row vector $\boldsymbol{\theta}_k$ is defined in terms in of acceleration and voicing amplitudes as

$$\boldsymbol{\theta}_k = \Big[\alpha_k \;\; \Omega_k \;\; E_k^0 \;\; E_k^e\Big]$$

in (4.3). Using the LF model in Equation (2.2), we can obtain the $(1\times N)$ row vector $\mathbf{g}_k(\boldsymbol{\theta}_k)$, that corresponds to a glottal waveform whose nth sample, $\Big[\mathbf{g}_k(\boldsymbol{\theta}_k)\Big]_n$, for $n = n_0, \ldots, N + n_0 - 1$, is given by

$$\Big[\mathbf{g}_k(\boldsymbol{\theta}_k)\Big]_n = \begin{cases} E_k^0\, e^{\alpha_k\, n} \cos(\Omega_k\, n), & n_0 \leq n \leq n_e \\ -\dfrac{E_k^e}{\epsilon\, n_a}\Big(\exp\big(-\epsilon\,(n - n_e)\big) - \exp\big(-\epsilon\,(n_c - n_e)\big)\Big), & n_e < n \leq n_c \\ 0, & n_c < n \leq N - 1 \end{cases}$$

Here, N is the fundamental pitch of the speech waveform, and n_0, n_e, n_c, and n_a are timing parameters that are evaluated offline based on a codebook.

Using the chain-matrix (CM) vocal tract (VT) model in Section 2.5, $S = 44$ uniform tubes are formed from a concatenated acoustic tube, starting at the glottis and ending at the lips. The $(1 \times S)$ row vector of sectional articulatory areas inside the segmented tube is given by

$$\mathbf{a}_k = \begin{bmatrix} a_k^{(1)} & a_k^{(2)} & \dots & a_k^{(S)} \end{bmatrix}$$

in (4.3). The CM model provides the VT impulse response function $\mathbf{v}_k(n; \mathbf{a}_k)$ obtained as the inverse discrete-time Fourier transform (DTFT) of the transfer function $\mathbf{V}(\omega; \mathbf{a}_k)$. Specifically, if the DTFT relationship is given by

$$\mathbf{v}_k(n; \mathbf{a}_k) \quad \xleftrightarrow{\text{DTFT}} \mathbf{V}(\omega; \mathbf{a}_k)\,, \tag{4.4}$$

to obtain a length M impulse response sequence, then we obrain the $(1 \times M)$ row vector $\mathbf{v}_k(\mathbf{a}_k)$ in (4.3). The transfer function $\mathbf{V}(\omega; \mathbf{a}_k)$ in (4.4) is obtained from the CM model as

$$\mathbf{V}(\omega; \mathbf{a}_k) = \frac{\mathcal{A}_k(\omega; \mathbf{a}_k)\, Z_L - \mathcal{B}_k(\omega; \mathbf{a}_k)}{\mathcal{A}_k(\omega; \mathbf{a}_k) - \mathcal{C}_k(\omega; \mathbf{a}_k) Z_L}\,, \tag{4.5}$$

where Z_L is the radiation impedance at the lips. The parameters in (4.5) are obtained from matrix

$$\boldsymbol{\psi}_k(\omega; \mathbf{a}_k) = \begin{bmatrix} \mathcal{A}_k(\omega; \mathbf{a}_k) & \mathcal{B}_k(\omega; \mathbf{a}_k) \\ \mathcal{C}_k(\omega; \mathbf{a}_k) & \mathcal{A}_k(\omega; \mathbf{a}_k) \end{bmatrix}$$

which is given as the final (or chain) matrix formed as the result of multiplying S segment matrices according to

$$\boldsymbol{\psi}_k(\omega; \mathbf{a}_k) = \boldsymbol{\psi}_{k-1}^{(S)}\Big(\omega; a_k^{(S)}\Big)\, \boldsymbol{\psi}_k^{(S-1)}\Big(\omega; a_k^{S-1}\Big)\, \boldsymbol{\psi}_k^{(S-2)}\Big(\omega; a_k^{(S-2)}\Big) \;\; \dots \;\; \boldsymbol{\psi}_k^{(1)}\Big(\omega; a_k^{(1)}\Big) \tag{4.6}$$

where

$$\boldsymbol{\psi}_k^{(j)}(\omega;\mathbf{a}_k^{(j)}) = \begin{bmatrix} \mathcal{A}_{k-1}^{(j)}(\omega;\mathbf{a}_k^{(j)}) & \mathcal{B}_{k-1}^{(j)}(\omega;\mathbf{a}_k^{(j)}) \\ \mathcal{C}_{k-1}^{(j)}(\omega;\mathbf{a}_k^{(j)}) & \mathcal{A}_{k-1}^{(j)}(\omega;\mathbf{a}_k^{(j)}) \end{bmatrix}, \quad j = 1,\ldots,S\,. \tag{4.7}$$

The matrix elements in (4.7) are given by

$$\begin{aligned} \mathcal{A}_k^{(j)}(\omega;a_k^{(j)},l_k^j) &= \cosh\left(\frac{\sigma(\omega)\,l_k^{(j)}}{c}\right) \\ \mathcal{B}_{k-1}^{(j)}(\omega;a_k^{(j)},l_k^j) &= -\frac{\rho\,c\,\gamma(\omega)}{a_k^{(j)}}\sinh\left(\frac{\sigma(\omega)\,l_k^{j)}}{c}\right) \\ \mathcal{C}_k^{(j)}(\omega;a_k^{(j)},l_k^{(j)}) &= -\frac{a_k^{(j)}}{\rho\,c\,\gamma(\omega)}\sinh\left(\frac{\sigma(\omega)l_k^{(j)}}{c}\right) \end{aligned} \tag{4.8}$$

where ρ and c are the density of air and the speed of sound in the air, respectively, and the frequency parameters $\gamma(\omega)$, and $\sigma(\omega)$ are evaluated based on [32], and Z_L is a load impedance as calculated in [45].

Lastly, the $(Q{\times}Q)$ covariance matrix in (4.3), where $Q = (4 + N + S + M)$ is given by

$$\mathbf{C}_k = \mathrm{diag}\Big(\Sigma_{\theta_k},\ \Sigma_{g_k},\ \Sigma_{a_k},\ \Sigma_{v_k}\Big)$$

where

$$\begin{aligned}
\Sigma_{\theta_k} &= \left\{\sigma^2_{k;\theta_1},\ \ldots,\ \sigma^2_{k;\theta_4}\right\} \\
\Sigma_{g_k} &= \left\{\sigma^2_{k;g_1},\ \ldots,\ \sigma^2_{k;g_N}\right\} \\
\Sigma_{a_k} &= \left\{\sigma^2_{k;a_1},\ \ldots,\ \sigma^2_{k;a_S}\right\} \\
\Sigma_{v_k} &= \left\{\sigma^2_{k;v_1},\ \ldots,\ \sigma^2_{k;v_M}\right\}.
\end{aligned}$$

Note that the state parameter $\mathbf{x}_k$ in (4.3) is an $(1 \times Q)$ row vector.

4.1.2 State Transition Equation

The state transition equation in (4.1) must provide a relationship between the unknown state parameter vector $\mathbf{x}_k$ at time step k and its value $\mathbf{x}_{k-1}$ at the previous time step $(k-1)$. This equation is needed in order to predict the unknown state vector $\mathbf{x}_k$ using its previously estimated value, before using the given measurement at time k to update the estimated $\mathbf{x}_k$. The random process $\mathbf{w}_k$ in (4.1) models a transition modeling error; it becomes important when the transition model used is empirically based and not based on any available physical models.

For the estimation of the glottal source and VT parameters, the transition equation depends on the unknown function $f_k(\mathbf{x}_k)$ in Equation (4.1). We can make certain assumptions based on the models used. For example, we can use the fact that for voiced sounds, it has been shown that formants vary slowly with time [19]. So, vocal tract behavior in the vector $\mathbf{a}_k$ can be modeled as a first order Markov chain

$$\mathbf{a}_k = \mathbf{a}_{k-1} + \mathbf{w}^{(\mathbf{a})}_{k-1}.$$

However, this slow variation in $\mathbf{a}_k$ does not necessarily imply a slow variation in the VT impulse response or VT transfer function in Equation (4.5). The state transition equation for the VT impulse response,

$$\mathbf{v}_k(\mathbf{a}_k) = f^{\mathbf{v}}_{k-1}(\mathbf{v}_{k-1}, \mathbf{a}_{k-1}) + \mathbf{w}^{(\mathbf{v})}_{k-1}\,,$$

could affect the estimation results based on the choice of the transition function; possible choices include $f^{\mathbf{v}}_{k-1}(\mathbf{v}_{k-1}, \mathbf{a}_{k-1}) = \mathbf{v}_{k-1}(\mathbf{a}_k)$ or $f^{\mathbf{v}}_{k-1}(\mathbf{v}_{k-1}, \mathbf{v}_{k-1}) = \mathbf{v}_{k-1}(\mathbf{v}_{k-1})$. Other possibilities may affect, for example, how the chain matrix is formed in (4.6) when transitioning from time step $(k-1)$ to time step k. Similar problems could arise for the LF glottal source model. For this thesis, and without testing for accuracy, we assumed the following transition equation

$$\begin{bmatrix}\boldsymbol{\theta}_k & \mathbf{g}_k(\boldsymbol{\theta}_k) & \mathbf{a}_k & \mathbf{v}_k(\mathbf{a}_k) & \mathbf{C}_k\end{bmatrix} = \begin{bmatrix}\boldsymbol{\theta}_{k-1} & \mathbf{g}_{k-1}(\boldsymbol{\theta}_{k-1}) & \mathbf{a}_{k-1} & \mathbf{v}_{k-1}(\mathbf{a}_{k-1}) & \mathbf{C}_{k-1}\end{bmatrix}$$
$$+\begin{bmatrix}\mathbf{w}^{(\boldsymbol{\theta})}_{k-1} & \mathbf{w}^{(\mathbf{a})}_{k-1} & \mathbf{w}^{(\mathbf{g})}_{k-1} & \mathbf{w}^{(\mathbf{v})}_{k-1} & \mathbf{w}^{(\mathbf{C})}_{k-1}\end{bmatrix}.$$

Note that the glottal source undergoes variations due to the changing physical surroundings such as temperature, pressure, humidity etc. These variations along with physiological variations from muscle fatigue and perceptual language modifications affect an ideal glottal source [14]. It was seen in [19] that when a glottal source is considered stochastic, it improves the estimation of inverse filtering. Considering $\mathbf{g}_k$ to be stochastic is realistic as glottis is not always ideal.

In this thesis, the VT model articulatory geometry length is considered constant, $l^{(j)}_k = 0.37$ cm, in (4.8). It may be of interest to increase dimension of vector $\mathbf{a}_k$ to obtain higher resolution geometrical description. However, doing so results in additional complexity and cascade estimation errors. It is of further interest to note that the vocal tract is a contiguous tube and any biological shrinking/elongation influences another length/section of the VT. It is possible to obtain perceptually similar speech even if we consider length and areas to be uncorrelated and ignore any coupling between them [45]. Hence, under this assumption of independence, we design the articulatory vector $\mathbf{a}_k$.

4.2 Time Varying Observation Model

As described in Chapter 2, a vowel is produced upon convolving the response of the VT and glottal input. This can be viewed as a blind decomposition/deconvolution problem [58]. There are numerous developed methods to separate these signals based on a stationary concept [58], [25], [19]. However, it is advantageous to express speech as a time-varying signal. This time-varying nature resembles speech production, where VT and glottis are coupled temporally [14]. A speech utterance can be written as follows:

$$y_k = \sum_{m=0}^{N-1} v[n;k]g[n-m;k] \tag{4.9}$$

where a shortened VT impulse response $\mathbf{v}_k(\mathbf{a}_{k-1}) \in \mathbb{R}^M$ is chosen at time k equal to $\mathbf{g}_k$, given by:

$$\mathbf{v}_k(\mathbf{a}_{k-1}) = [v[0;k] \quad v[1;k] \quad \ldots \quad v[M;k]]^T \tag{4.10}$$

The LF glottal input $\mathbf{g}_k(\boldsymbol{\theta}_{k-1}) \in \mathbb{R}^M$ is mirrored after being obtained from $\mathbf{x}_k$ and is given as

$$\mathbf{g}_k(\boldsymbol{\theta}_{k-1}) = [g[M-1;k], \ldots, g[0;k]]^T \tag{4.11}$$

where M is the fundamental pitch period which may be calculated using any pitch calculation technique [59, 60]. We use RAPT for due to its time domain pitch calculations [61].

We denote (4.9) as $h_k(\mathbf{x_k})$, a function of states $\mathbf{x_k}$. In a state-space representation this is simply given by:

$$y_k = h_k(\mathbf{x_k}) + \eta_k \tag{4.12}$$

that is perturbed by AWGN noise source $\eta_k \sim \mathcal{N}(0, \sigma_y^2)$ with variance σ_y^2, and $\mathbf{x_k}$ is parameter-state vector of glottal input and VT. Under this state-space framework, one may use numerous state-estimation methods to solve for posterior states.

4.3 Boostrap Particle Filter

Sampling Importance Resampling particle filter requires the ability to evaluate and draw particles from a proposal distribution $p(\mathbf{x}_k)$. An optimal choice optimal choice that minimizes particle weight variance is $p(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{y}_k)$. However, this is difficult to obtain and instead an alternative is to set importance density as prior density $p(\mathbf{x}_k|\mathbf{x}_{k-1})$. The importance weight then reduces to

$$w_k^{(i)} = w_{k-1}^{(i)} p(\mathbf{y}_k^{(i)}|\mathbf{x}_k^{(i)}) \tag{4.13}$$

By resampling the particles at every iteration, the particle weights are forced to be equal.

$$w_k^{(i)} = p(\mathbf{y}_k^{(i)}|\mathbf{x}_k^{(i)}) \tag{4.14}$$

which can be derived explicitly from the observation likelihood $p(\mathbf{y}_k|\mathbf{x}_k^{(i)})$. Drawing from the prior distribution is then a matter of propagating the previous estimate $\mathbf{x}_{p,k-1}$ from each particle through the state evolution model (**??**). In this process, the propagation of each particle state includes a sampled realization of the process noise for all of the random variables in the model equations. The resulting distribution of particle states is then a discrete approximation to $p(\mathbf{x}_k|\mathbf{x}_{k-1})$. The complete bootstrap particle filter algorithm is summarized in algorithm 2.

The bootstrap particle filter provides an estimator for the system and works well provided sufficient particles are used. However, the required number of particles grows exponentially with the number of dimensions in the estimated state.

4.4 Biomechanical constraints

The recovered data is unconstrained and hence may obtain unrealistic estimates of area function and glottal voicing thresholds. Acoustic theory by Fant [31], describes temporal change in area function to be minimized, since muscles move very slowly. Similarly, a rapid change in geometry of a concatenated tube is unrealistic so the sectional change should be smooth. To accommodate this we impose

$$\boldsymbol{\kappa}_a \leq |\mathbf{a}_k^i - \mathbf{a}_k^{i-1}| \leq \boldsymbol{\kappa}_b \tag{4.15}$$

where $2 \leq i \leq S$, are S VT sections. A similar expression may be formed for VT length $\mathbf{l}$. Empirical observation through MRI suggests a maximum threshold for area physically possible to achieve [62]. We hence impose this physical limit as a biological constraint (typically maximum articulatory area $\approx 14cm^2$).

Algorithm 2: Bootstrap Particle Filtering

```
begin
    // Initialize
    forall particles $p = 1, \dots N_p$ do
        Draw $\mathbf{x}_{p,0} \sim \pi(\mathbf{x_k})$ from an initial prior distribution
    end
    for $k \leftarrow 1, \dots N-1$ do
        forall particles $p = 1, \dots N_p$ do
            // Correct
            $w_p \leftarrow \mathcal{N}(h_k(\mathbf{x}_p), \sigma^2_{s,p})$;
        end
        $w_p \leftarrow w_p \{\Sigma_p w_p\}^{-1}$;        // Normalize
        $\hat{\mathbf{x}}_k \leftarrow \Sigma_p w_p \mathbf{x}_p$;        // Estimate
        $\mathbf{x}_p \leftarrow R(w_p, \mathbf{x}_p)$;        // Resample
        forall particles $p = 1, \dots N_p$ do
            //Predict
            //Sample $\mathbf{x}_{p,k} \sim p(\mathbf{x}_k | \mathbf{x}_{p,k-1})$
            Propagate $\mathbf{x}_p = f(\mathbf{x}_p, \mathbf{e}_k)$
        end
    end
end
```

Evidence of potential and kinetic energy change during onset and offset of vowels also suggest minimization of temporal energy, consequently temporal area in VT [29].

The resulting constraint $\mathcal{C}_{\nu_k}$ is:

$$\mathcal{C}_{\nu_k} = \|\mathbf{a}_k - \mathbf{a}_{k-1}\|_2^2 \tag{4.16}$$

If $\mathbf{a}_{k-1} = \mathbf{a}_{rest}$, rest configuration of the vocal tract we instead minimize potential energy [31]. This is applicable when VT dynamics are constant or negligible. We define constraint $\mathcal{C}_{\mathcal{T}_k}$ as:

$$\mathcal{C}_{\mathcal{T}_k} = \frac{\delta \mathcal{C}_{\nu_k}}{\delta \mathbf{a}_k} \mathcal{C}_{\nu_k} \tag{4.17}$$

where,

$$\frac{\delta \mathcal{C}_{\nu_k}}{\delta \mathbf{a}_k} = \begin{cases} 2\Delta \mathbf{a}_k, & k = 0 \\ 2\left[\Delta \mathbf{a}_k - \Delta \mathbf{a}_{k-1}\right], & 2 \leq k \leq N-2 \\ 2\Delta \mathbf{a}_k, & k = N-1 \end{cases} \tag{4.18}$$

Equations (4.15), (4.16), and (4.17) impose constraints on articulatory geometry. We may consider combined contribution from energy constraints as:

$$\boldsymbol{\kappa}_c \leq c_{pot}\mathcal{C}_{\mathcal{T}_k} + c_{kin}\mathcal{C}_{\nu_k} + \leq \boldsymbol{\kappa}_d \tag{4.19}$$

c_{in} and c_{pot} are empirically chosen parameters such that $c_{pot} + c_{kin} < 1$. A final constraint is imposed on glottal parameters as described Equations (**??**) - (**??**). Together these form a set of constraints

$$\boldsymbol{\varepsilon}^{LB} \leq \Upsilon(\mathbf{x}_k) \leq \boldsymbol{\varepsilon}^{UB} \tag{4.20}$$

where ε determines upper and lower bound of constraints. One possible way to impose these constraints is projection of the unconstrained density onto a constraint set. A widely used alternative is constrained sequential Monte Carlo by acceptance/rejection approach [63]. The acceptance/rejection process does not make any assumption on distributions and therefor maintains generic properties of the particle filter. However, due rejection the number of samples will be reduced, the resulting conditional mean distribution comes from a truncated set of particles, this effectively lowers accuracy if there are insufficient number of particles. An extreme example is when all particles violate the constraints and algorithm fails. One way to overcome this issue is by initiating high number of particles N_p. This is a brute force approach and results in high complexity. The rejection also reduces support on proposal distribution generates a truncated distribution, a truncated Gaussian in our case. However, an elaborate proof of convergence is shown in [64]. Algorithm 3 shows a bootstrap particle filter with constraints and acceptance/rejection approach.

Algorithm 3: Bootstrap Particle Filtering with Constraints

```
begin
    // Initialize
    forall particles $p = 1, \dots N_p$ do
        Draw $\mathbf{x}_{p,0} \sim \pi(\mathbf{x}_k)$ from an initial prior distribution
        if violates $\boldsymbol{\varepsilon}^{LB} \leq \Upsilon(\mathbf{x}_k) \leq \boldsymbol{\varepsilon}^{UB}$ then
            // Reject particle
            // Resample rejected particle from $\pi(\mathbf{x}_0)$
        else
            // Continue
        end
    end
    for $k \leftarrow 1, \dots N - 1$ do
        forall particles $p = 1, \dots N_p$ do
            // Correct
            $w_p \leftarrow \mathcal{N}(h_k(\mathbf{x}_p), \sigma^2_{s,p})$;
        end
        $w_p \leftarrow w_p \{\Sigma_p w_p\}^{-1}$;        // Normalize
        $\hat{\mathbf{x}}_k \leftarrow \Sigma_p w_p \mathbf{x}_p$;        // Estimate
        $\mathbf{x}_p \leftarrow R(w_p, \mathbf{x}_p)$;        // Resample
        if violates $\boldsymbol{\varepsilon}^{LB} \leq \Upsilon(\mathbf{x}_{p,k}) \leq \boldsymbol{\varepsilon}^{UB}$ then
            // Reject particle
        else
            Propagate $\mathbf{x}_p = f(\mathbf{x}_p, \mathbf{e}_{k+1})$
        end
        forall particles $p = 1, \dots N_p$ do
            //Predict
            //Sample $\mathbf{x}_{p,k+1} \sim p(\mathbf{x}_{k+1} | \mathbf{x}_{p,k})$
            // Resample rejected particles from $\pi(\mathbf{x}_{k|k-1})$
        end
    end
end
```

4.5 Computational Complexity

The asymptotic computational complexity of the proposed estimator is dominated by the following factors. First is calculation of weight updates in bootstrap particle filter for each particle. This involves estimating posterior from $2M + 88$-dimensional multivariate Gaussian. To obtain frequency response $V_k(\omega; \mathbf{a}_k)$ given size of CM $\boldsymbol{\psi}(\omega; \mathbf{a}_k)$ as 2×2 and $S = 44$, is $\mathcal{O}(N_\omega \times 2^2 \times (S-1))$. The complexity of inverse Fourier of $V_k(\omega; \mathbf{a}_k)$ is $\mathcal{O}(N_\omega \log(N_\omega))$, for this thesis $N_\omega = F_s$. The estimation of glottal input is $\mathcal{O}(M)$. The calculations above must be completed for every particle, hence total complexity is:

$$\mathcal{O}(N_p(M + N_w \times (S-1) \times 2^2 + N_\omega \log N_\omega))) \tag{4.21}$$

The number of particles in the preceding equation can be approximately expressed in terms of the other dimensions of the problem, sampling rate and pitch ($M = N_0$). Depending on target constraints this could require parallel processing for N_p particles. In this study, a GPU NVIDIA GTX 1070 with Max-Q design is used.

Chapter 5

RESULTS

For synthesized glottal flows, the Normalized Amplitude Quotient(NAQ) was estimated for each cycle. In order to compare the NAQ values of the original and the estimated glottal flows.

$$\mathrm{NAQ}_k = \frac{E_{0_k}}{E_{e_k} \cdot T_0} \tag{5.1}$$

NAQ scored obtained through mngu0 corpus [2] are considered to be true values and an error metric is calculated using

$$NAQ_{error;k} = \frac{1}{k} \sum_{1}^{k} \frac{\|NAQ_{ref;k} - NAQ_{estimated;k}\|_2^2}{NAQ_{ref;k}} \tag{5.2}$$

these are shown in Table 5.1 as percentages and compared against [17], IAIF method [22], conventional LP method [3], SSIF [19] and QCP [27].

After obtaining a VT response, we looking at the conventional LP method to validate peaks for our estimates as depicted in Table 5.3

$$e_{F_{i;k}} = 100 \sqrt{\frac{1}{N} \sum_{n=1}^{N} \left(\frac{\hat{F}_{i;k} - F_{i;k}}{F_{i;k}} \right)^2} \tag{5.3}$$

where $F_{i;k}$ is the i^{th} formant at time k.

Lastly, we have $H_1 H_2$ index which are the difference between the first harmonic

Table 5.1: NAQ Error for glottal input

NAQ	100 Hz	200 Hz	300 Hz
IAIF	60.2	76.9	81.2
SSIF	59.3	70.2	70.3
QCP	42.2	55.2	80.2
BSSAR	30.8	24.4	30.7

Table 5.2: H1H2 Error for glottal input

H1-H2	100 Hz	200 Hz	300 Hz
IAIF	1.3	43.4	50.9
SSIF	0.3	35.4	14.9
QCP	0.15	25.6	10.8
BSSAR	0.1	8.2	2.9

and second harmonics of glottis

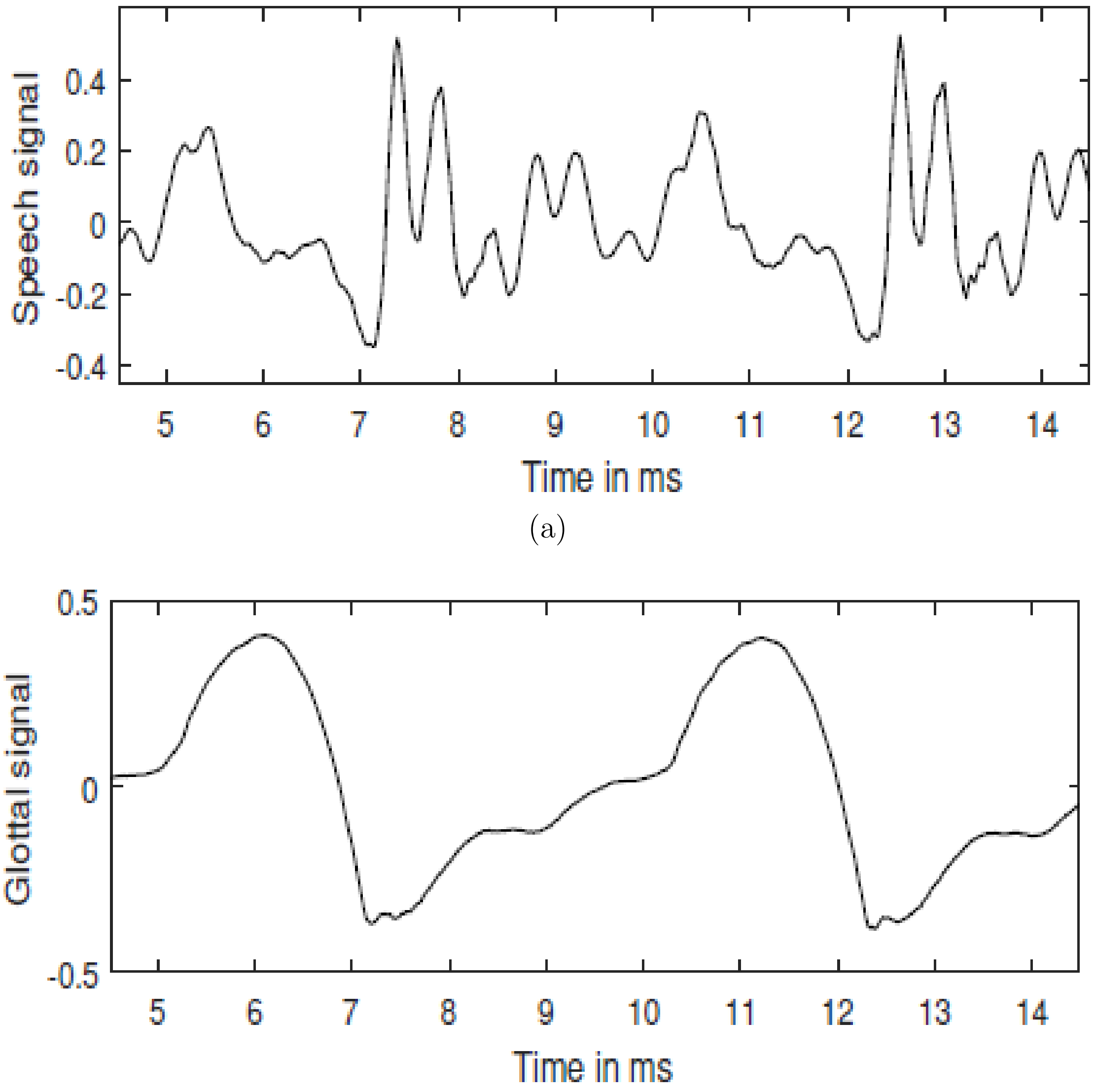
$$H_1 - H_2 = -6 + 0.27 \cdot exp(5.5OQ) \tag{5.4}$$

with OQ as open quotient of recovered glottal pulse. An error metric is computed as

$$e_{H_1-H_2;k} = \sum_{p=1}^{k} |H_1H_{2_{ref;k}} - H_1H_{2_{estimate;k}}| \tag{5.5}$$

Table 5.3: Vocal tract formant (root-mean square error) RMSE error

	/i/			/a/			/u/		
	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3
SSIF	0.39	0.5	0.72	0.42	0.61	0.77	0.67	0.75	0.99
IAIF	0.31	1.54	0.68	0.39	0.95	0.45	1.23	0.97	0.88
QCP	5.62	2.0	0.48	3.55	2.42	0.78	2.58	1.51	0.91
BSSAR	0.22	0.15	0.41	0.15	0.29	0.13	0.19	0.17	0.1



(a)

(b)

Figure 5.1: Speech Waveform is shown in (a) with F_0 = 198Hz, (b) shows the recovered glottal waveform

Table 5.4: MSE for raw speech output

	/i/	/a/	/u/
MSE	0.13	0.15	0.12

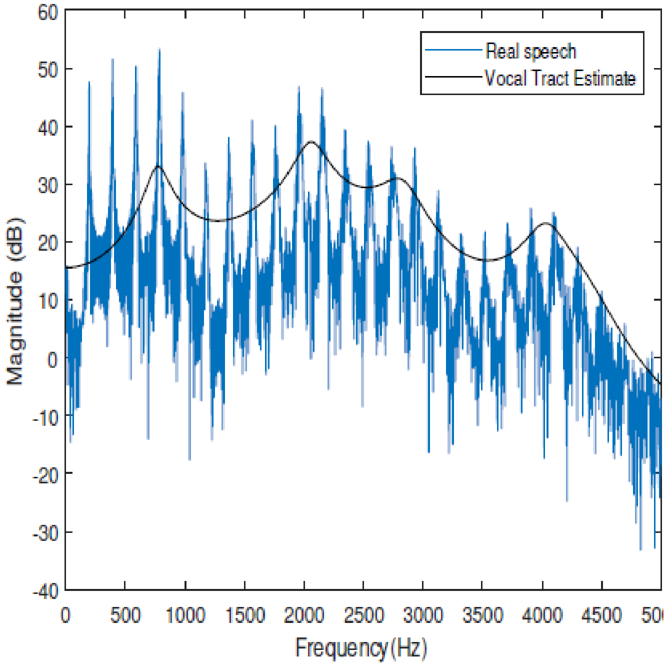


Figure 5.2: The Vocal tract estimate and the true spectrum of speech signal for the vowel /e/

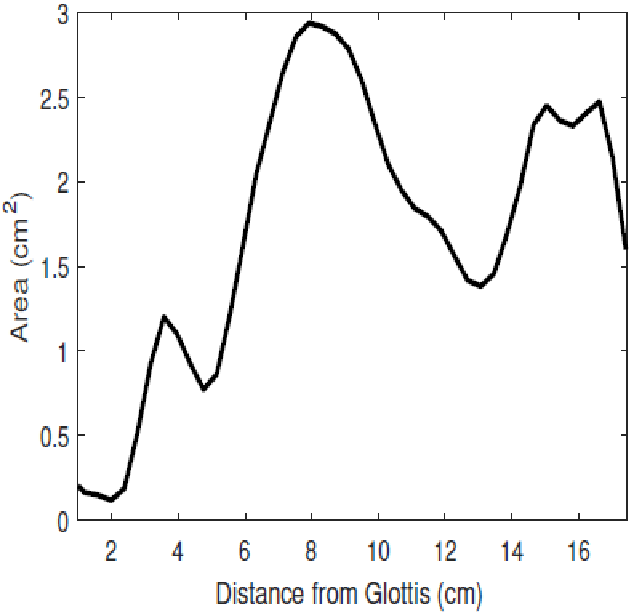


Figure 5.3: The recovered area function of the vowel /e/

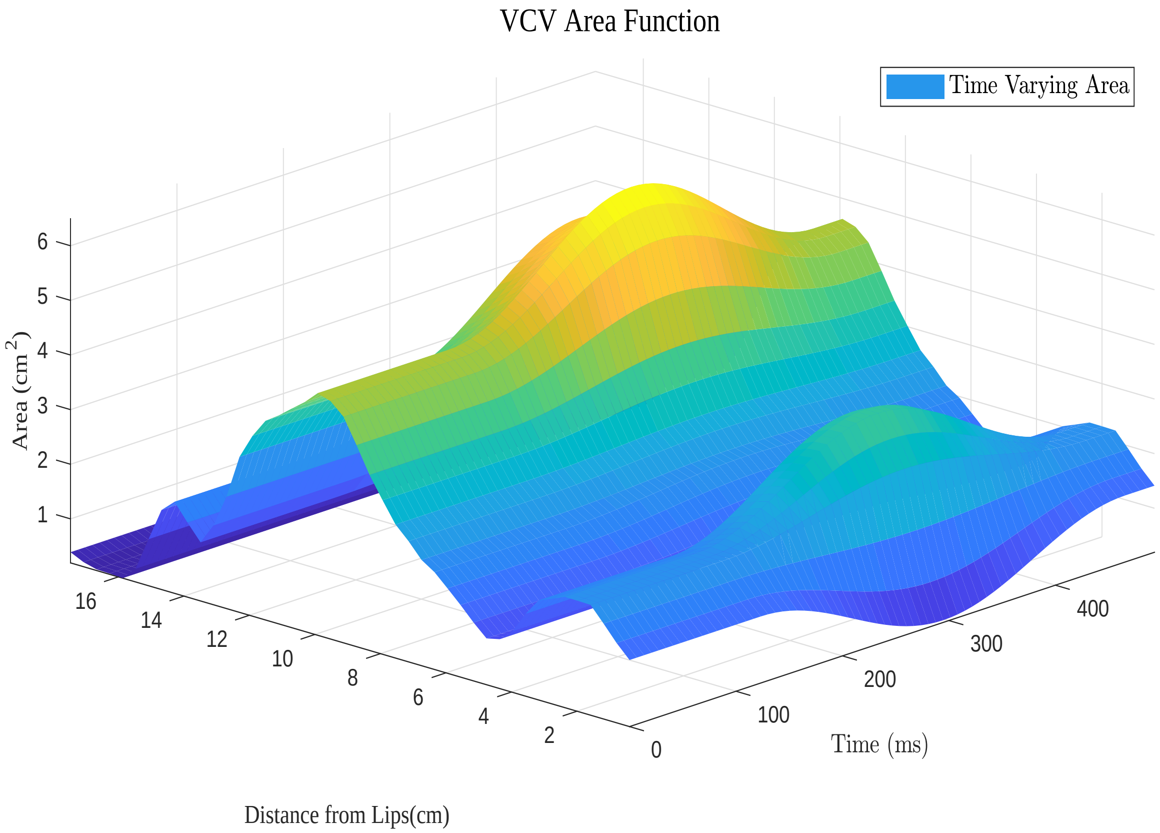


Figure 5.4: The recovered area function of the vowel-consonant-vowel transition /*a*/-/*d*/-/*a*/

Chapter 6

CONCLUSIONS

The chosen framework proves reliable and is able to decompose speech better than previous quasi-stationary methods. Computational complexity is a concern, as the pitch decreases and dimension of the state vector grows. To handle this better, one possible solution is to isolate the state-parameter augmentation and solve parameters to be independent of state estimation. This however, leads to poor performance when matching of signals is concerned. Future work will target reducing the time required for decomposition and finding alternative ways to impose constraints on particle filter. The general state space model can be extended for fricatives, consonants and stop explosives using [9]. This would allow a time varying recovery of non-nasalized vowels to help decoding all parts of speech.

References

[1] G. Degottex, *Glottal source and vocal-tract separation.* Theses, Université Pierre et Marie Curie - Paris VI, 2010.

[2] R. Korin, "Announcing the electromagnetic articulography (Day 1) subset of the mngu0 articulatory corpus," in *Proc. Interspeech*, pp. 1505–1508, 2011.

[3] J. D. Markel and A. H. Gray, *Linear Prediction of Speech*, vol. 12. Springer, 1976.

[4] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[5] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: from features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12 – 40, 2010.

[6] A. J. Gully, T. Yoshimura, D. T. Murphy, K. Hashimoto, Y. Nankaku, and K. Tokuda, "Articulatory text-to-speech synthesis using the digital waveguide mesh driven by a deep neural network," in *Proc. Interspeech*, pp. 234–238, 2017.

[7] F. Taguchi and T. Kaburagi, "Articulatory-to-speech conversion using bi-directional long short-term memory," in *Proc. Interspeech*, pp. 2499–2503, 2018.

[8] B. H. Story and K. Bunton, "Identification of stop consonants produced by an acoustically-driven model of a child-like vocal tract," *The Journal of the Acoustical Society of America*, vol. 140, no. 4, pp. 3218–3218, 2017.

[9] B. Elie and Y. Laprie, "A glottal chink model for the synthesis of voiced fricatives," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5240–5244, 2016.

[10] H. Ze, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7962–7966, 2013.

[11] S. Warhurst, P. McCabe, R. Heard, E. Yiu, G. Wang, and C. Madill, "Quantitative measurement of vocal fold vibration in male radio performers and healthy controls using high-speed videoendoscopy," *PLOS ONE*, vol. 9, pp. 1–8, 2014.

[12] V. Ramanarayanan, B. Parrell, L. Goldstein, S. Nagarajan, and J. Houde, "A new model of speech motor control based on task dynamics and state feedback," in *Proc. Interspeech*, pp. 3564–3569, 2016.

[13] B. Elie and G. Chardon, "Glottal/Supraglottal Source Separation in Fricatives Based on Non-Stationnary Signal Subspace Estimation." preprint, Apr. 2018.

[14] S. G. Lingala, B. P. Sutton, M. E. Miquel, and K. S. Nayak, "Recommendations for real-time speech MRI," *Journal of Magnetic Resonance Imaging*, vol. 43, no. 1, pp. 24–44, 2015.

[15] V. Mitra, G. Sivaraman, C. Bartels, H. Nam, W. Wang, C. Espy-Wilson, D. Vergyri, and H. Franco, "Joint modeling of articulatory and acoustic spaces for continuous speech recognition tasks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5205–5209, 2017.

[16] C. Hagedorn, M. Proctor, L. Goldstein, S. M. Wilson, B. Miller, M. L. Gorno-Tempini, and S. S. Narayanan, "Characterizing articulation in apraxic speech using real-time magnetic resonance imaging," *Journal of Speech, Language, and Hearing Research*, vol. 60, no. 4, pp. 877–891, 2017.

[17] I. R. Bleyer, L. Lybeck, H. Auvinen, M. Airaksinen, P. Alku, and S. Siltanen, "Alternating minimisation for glottal inverse filtering," *Inverse Problems*, vol. 33, no. 6, pp. 65005–65024, 2017.

[18] H. Auvinen, T. Raitio, S. Siltanen, and P. Alku, "Utilizing Markov chain Monte carlo (MCMC) method for improved glottal inverse filtering," in *Proc. of Interspeech*, pp. 1638–1641, 2012.

[19] G. A. Alzamendi and G. Schlotthauer, "Modeling and joint estimation of glottal source and vocal tract filter by state-space methods," *Biomedical Signal Processing and Control*, vol. 37, pp. 5 – 15, 2017.

[20] U. Benigno, R. Steve, and R. Korin, "A deep neural network for acoustic-articulatory speech inversion," *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

[21] J. Walker and P. Murphy, "Advanced methods for glottal wave extraction," in *Nonlinear Analyses and Algorithms for Speech Processing*, pp. 139–149, Springer, 2005.

[22] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Communication*, vol. 11, no. 2, pp. 109 – 118, 1992.

[23] V. L. Heiberger and Y. Horii, "Jitter and shimmer in sustained phonation," *Speech and Language*, vol. 7, pp. 299–332, 1982.

[24] J. Flanagan, M. Schroeder, B. Atal, R. Crochiere, N. Jayant, and J. Tribolet, "Correction to "speech coding"," *IEEE Transactions on Communications*, vol. 27, no. 6, pp. 932–932, 1979.

[25] P. Jinachitra and J. O. Smith, "Joint estimation of glottal source and vocal tract for vocal synthesis using Kalman smoothing and EM algorithm," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 327–330, 2005.

[26] P. K. Ghosh and S. S. Narayanan, "A subject-independent acoustic-to-articulatory inversion," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4624–4627, 2011.

[27] S. Sahoo and A. Routray, "A novel method of glottal inverse filtering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1230–1241, 2016.

[28] S. Panchapagesan and A. Alwan, "A study of acoustic-to-articulatory inversion of speech by analysis-by-synthesis using chain matrices and the Maeda articulatory model," *The Journal of the Acoustical Society of America*, vol. 129, no. 4, pp. 2144–2162, 2011.

[29] B. Elie and Y. Laprie, "Audiovisual to area and length functions inversion of human vocal tract," in *European Signal Processing Conference*, pp. 2300–2304, 2014.

[30] S. Pramit, S. Praneeth, and F. Sidney, "Towards automatic speech identification from vocal tract shape dynamics in real-time MRI," in *Proc. Interspeech*, pp. 1249–1253, 2018.

[31] G. Fant, *Acoustic theory of speech production with calculations based on X-ray studies of Russian articulations.* The Hague, Mouton, 1970.

[32] M. Sondhi and J. Schroeter, "A hybrid time-frequency domain articulatory speech synthesizer," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 7, pp. 955–967, 1987.

[33] G. Fant, "A four-parameter model of glottal flow," *Speech Transmission Laboratory, Quarterly Progress and Status Reports*, vol. 26, no. 4, pp. 1–13, 1985.

[34] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *Speech Synthesis Workshop*, 2016.

[35] A. Costa and M. Santesteban, "Lexical access in bilingual speech production: Evidence from language switching in highly proficient bilinguals and L2 learners," *Journal of Memory and Language*, vol. 50, no. 4, pp. 491 – 511, 2004.

[36] L. Fontan, M. Le Coz, and S. Detey, "Automatically measuring L2 speech fluency without the need of ASR: A proof-of-concept study with Japanese learners of French," in *Proc. Interspeech*, pp. 2544–2548, 2018.

[37] R. S. McGowan and M. S. Howe, "Comments on single-mass models of vocal fold vibration," *The Journal of the Acoustical Society of America*, vol. 127, no. 5, pp. 215–221, 2010.

[38] K. Ishizaka and J. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords," *Bell Syst. Tech. Journal*, vol. 51, no. 6, pp. 1233–1268, 1972.

[39] A. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *The Journal of the Acoustical Society of America*, vol. 49, no. 2B, pp. 583–590, 1971.

[40] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *The Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 820–857, 1990.

[41] G. Fant, "Vocal source analysis," *Speech Transmission Laboratory, Quarterly Progress and Status Reports*, vol. 20, no. 3-4, pp. 31–53, 1979.

[42] G. Fant, "The LF-model revisited," *Speech Transmission Laboratory, Quarterly Progress and Status Reports*, vol. 36, no. 2-3, pp. 119–156, 1995.

[43] B. Doval and C. d'Alessandro, "Spectral correlates of glottal waveform models: an analytic study," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 1295–1298, 1997.

[44] G. A. Alzamendi and G. Schlotthauer, "Modeling and joint estimation of glottal source and vocal tract filter by state-space methods," *Biomedical Signal Processing and Control*, vol. 37, pp. 5–15, 2017.

[45] B. H. Story, "Phrase-level speech simulation with an airway modulation model of speech production," *Computer Speech and Language*, vol. 27, pp. 989–1010, 2013.

[46] J. Kelly and C. C. Lochbaum, "Statistical parametric speech synthesis using deep neural networks," in *Proc. of Fourth International Congress on Acoustics*, pp. 1–4, 1962.

[47] P. Mokhtari, H. Takemoto, and T. Kitamura, "Single-matrix formulation of a time domain acoustic model of the vocal tract with side branches," *Speech Communication*, vol. 50, no. 3, pp. 179 – 190, 2008.

[48] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory.* Prentice-Hall, 1993.

[49] A. Doucet, N. de Freitas, and N. Gordon, *Sequential Monte Carlo Methods in Practice.* Springer, 2001.

[50] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods (Springer Texts in Statistics).* Berlin, Heidelberg: Springer-Verlag, 2005.

[51] G. Welch and G. Bishop, "An introduction to the Kalman filter," tech. rep., 1995.

[52] E. A. Wan and R. V. D. Merwe, "The unscented kalman filter for nonlinear estimation," in *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (Cat. No.00EX373)*, pp. 153–158, 2000.

[53] A. Doucet, S. Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering," *Statistics and Computing*, vol. 10, no. 3, pp. 197–208, 2000.

[54] N. Kantas, A. Doucet, S. S. Singh, J. Maciejowski, and N. Chopin, "On particle methods for parameter estimation in state-space models," *Statist. Sci.*, no. 3, pp. 328–351, 2015.

[55] C. M. Carvalho, M. S. Johannes, H. F. Lopes, and N. G. Polson, "Particle learning and smoothing," *Statist. Sci.*, vol. 25, no. 1, pp. 88–106, 2010.

[56] C. Nemeth, P. Fearnhead, and L. Mihaylova, "Sequential monte carlo methods for state and parameter estimation in abruptly changing environments," *IEEE Transactions on Signal Processing*, vol. 62, no. 5, pp. 1245–1255, 2014.

[57] J. Xia and M. Y. Wang, "Particle filtering with sequential parameter learning for nonlinear bold fMRI signals," *Advances and applications in statistics*, vol. 40, no. 1, pp. 61–74, 2014.

[58] O. Schleusing, T. Kinnunen, B. Story, and J. Vesin, "Joint source-filter optimization for accurate vocal tract estimation using differential evolution," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 8, pp. 1560–1572, 2013.

[59] F. Huang, Y. T. Yeung, and T. Lee, "Evaluation of pitch estimation algorithms on separated speech," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6807–6811, 2013.

[60] S. Ahmadi and A. S. Spanias, "Cepstrum-based pitch detection using a new statistical V/UV classification algorithm," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 333–338, 1999.

[61] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech coding and synthesis* (W. B. Kleijn and K. K. Paliwal, eds.), pp. 495–518, Elsevier Science, 1995.

[62] S. Maeda, "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model," *Speech Production and Speech Modelling*, pp. 131–149, 1990.

[63] B. Ebinger, N. Bouaynaya, R. Polikar, and R. Shterenberg, "Constrained state estimation in particle filters," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4050–4054, 2015.

[64] N. Amor, N. C. Bouaynaya, R. Shterenberg, and S. Chebbi, “On the convergence of constrained particle filters,” *IEEE Signal Processing Letters*, vol. 24, no. 6, pp. 858–862, 2017.