Nonstationary Time Series Modeling with Application to Speech Signal Processing

A dissertation presented

by

Daniel Rudoy

to

The Department of School of Engineering and Applied Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Engineering Sciences

Harvard University

Cambridge, Massachusetts

September 2010

Thesis advisor
Patrick J. Wolfe

Author
Daniel Rudoy

Nonstationary Time Series Modeling with Application to Speech Signal Processing

Abstract

We develop statistical methods for the analysis of nonstationary time series and apply them to a variety of problems arising in speech signal processing. Information-carrying natural sound signals such as speech exhibit a degree of controlled nonstationarity in that their statistical properties vary slowly over time. Faithfully modeling these temporal variations is extremely valuable for a wide range of applications and can be accomplished by relying on well-understood acoustic models of speech production, which motivate many of the methods developed in this thesis.

First, we make a number of contributions to the classical problem of formant tracking, in which vocal tract resonances are estimated under the assumption of their invariance on the 15–30 ms scale. Next, we relax this piecewise-stationarity constraint and model the temporal dynamics of the vocal tract using time-varying autoregressive (TVAR) models. We develop their algebraic and geometric properties, introduce several new estimators, and use TVAR models to develop a hypothesis test to detect the presence of vocal tract variation in speech waveform data. We study its asymptotic properties, and illustrate its practical efficacy by detecting vocal tract changes across different timescales of speech dynamics.

Next, we explore how standard fixed-resolution short-time Fourier representations may be generalized in order to adapt to the time-frequency structure of a speech signal. To this end, we introduce a family of adaptive, linear time-frequency representations termed superposition frames and show that they are invertible, numerically-stable, and admit fast overlap-add reconstruction akin to standard short-time Fourier techniques. The general construction proceeds via a local signal-adaptive modification of a Gabor frame. Two signal-dependent schemes for selecting an appropriate superposition frame for signal analysis are given, and the framework is illustrated in the context of speech enhancement.

Finally, we introduce a joint model of the vocal tract and the source waveform in order to take into account its quasi-periodic temporal variations during voicing. We incorporate an estimate of the source waveform into the traditional linear prediction framework via nonparametric wavelet regression; the resultant semi-parametric model is applied to various speech analysis problems including formant and source-harmonics-to-noise ratio estimation, inverse filtering, and voicing detection.

Contents

List of Figures

List of Algorithms

List of Tables

Notational Conventions

$\boldsymbol{m}$	Column vector $\boldsymbol{m}$
$\boldsymbol{M}$	Matrix $\boldsymbol{M}$
$.^T$	Transpose
$\overline{.}$	Complex conjugate
$\otimes$	Kronecker product
$*$	Convolution
$\|\cdot\|, \quad \|\cdot\|_p$	ℓ_2- and ℓ_p- norms, respectively
$\langle \cdot\,, \cdot \rangle$	inner product
$\|\boldsymbol{M}\|$	Determinant of $\boldsymbol{M}$
$\mathrm{tr}(\boldsymbol{M})$	Trace of $\boldsymbol{M}$
$\boldsymbol{M}^{\#}$	Pseudo-inverse of $\boldsymbol{M}$
$\mathrm{span}(\boldsymbol{M})$	Column span of $\boldsymbol{M}$
$P_{\boldsymbol{M}}$	Projection onto the column span of $\boldsymbol{M}$
$P_{\boldsymbol{M}}^{\perp}$	Projection onto the orthogonal complement of the column span of $\boldsymbol{M}$
$\boldsymbol{I}_n$	$n \times n$ identity matrix
$\mathbb{Z}, \quad \mathbb{Z}^n$	Spaces of one and n-dimensional integer-valued vectors, respectively
$\mathbb{R}, \quad \mathbb{R}^n$	Spaces of one and n-dimensional real-valued vectors, respectively
$\mathbb{C}, \quad \mathbb{C}^n$	Spaces of one and n-dimensional complex-valued vectors, respectively
$p(\boldsymbol{x})$	Probability density function (PDF) of $\boldsymbol{x}$
$\mathbb{E}(\cdot)$	Expectation operator
$p(\boldsymbol{x} \,\vert\, \boldsymbol{y}\,;\, \boldsymbol{\theta})$	Conditional PDF of $\boldsymbol{x}$ given $\boldsymbol{y}$ and parameterized by $\boldsymbol{\theta}$
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Multivariate Gaussian PDF with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$
$\mathcal{G}(\alpha, \beta)$	Gamma PDF with parameters α and β
$\mathcal{IG}(\alpha, \beta)$	Inverse-Gamma PDF with parameters α and β
$[n]$	Discrete index n
(t)	Continuous index t
$\mathcal{Z}$	Z-transform

Abbreviations

AIC	Akaike Information Criterion
AR	Autoregressive
ARMA	Autoregressive Moving Average
ARX	Autoregressive with eXogenous input
BIBO	Bounded Input Bounded Output
CFAR	Constant False Alarm Rate
CRLB	Cramer-Rao Lower Bound
DFT	Discrete Fourier Transform
EM	Expectation Maximization
EKF	Extended Kalman Filter
EKS	Extended Kalman Smoother
FFT	Fast Fourier Transform
GCI	Glottal Closure Instant
GLRT	Generalized Likelihood Ratio Test
GOI	Glottal Opening Instant
LF	Liljencrantz-Fant
LPC	Linear-Predictive Coding
LPCC	Linear-Predictive Cepstral Coefficients
MA	Moving Average
MAP	Maximum A Posteriori
MDL	Minimum Description Length
ML	Maximum Likelihood
MMSE	Minimum-Mean-Squared Error
PARCOR	PArtial CORrelation Coefficients
PND	Purely Non-Deterministic
RCA	Random Coefficient Autogressive
RMSE	Root-Mean-Squared Error
ROC	Receiver Operating Characteristic
SHNR	Source Harmonics-to-Noise Ratio
SNR	Signal-to-Noise Ratio
STFT	Short-Time Fourier Transform
TIMIT	Texas Instruments Massachusetts Institute of Technology
TF	Time-Frequency
TVAR	Time-Varying Autoregressive
TV-RCA	Time-Varying Random Coefficient Autoregressive
TVARX	Time-Varying Autoregressive with eXogenous input
VTR	Vocal Tract Resonance
WLS	Weighted Least Squares
WSS	Wide-Sense Stationary

Citations to Previously Published Work

- Parts of Chapter 3 appeared in
  - D. Rudoy, D. Spendley and P. J. Wolfe. "Conditionally linear Gaussian models for tracking of vocal tract resonances." In Proc. 8th Ann. Conf. Intl. Speech Commun. Ass. (Interspeech) 2007, pp. 526-529.
- Parts of Chapter 4 appeared in
  - D. Rudoy, T. F. Quatieri and P. J. Wolfe, "Estimating Stable Time-Varying Autoregressive Models: A Convex Optimization Approach." Submitted to IEEE Trans. Signal Process.
  - D. Rudoy and T. Georgiou, "Regularized Parametric Models of Nonstationary Processes." In Proc. 19th Intl. Symp. Math. Theory Networks, Systems (MTNS) 05-09 July 2010 Budapest, Hungary, to appear.
- Parts of Chapter 5 appeared in:
  - D. Rudoy, T. F. Quatieri and P. J. Wolfe, "Time-Varying Autoregressions in Speech: Detection Theory and Applications." IEEE Trans. Audio, Speech Lang., to appear.
  - D. Rudoy, T. F. Quatieri and P. J. Wolfe, "Time-varying autoregressive tests for multiscale speech analysis." In Proc. 10th Ann. Conf. Intl. Speech Commun. Ass. (Interspeech), 2009, pp. 2839–2842.
  - P. Basu, D. Rudoy and P. J. Wolfe, "A nonparametric test for stationarity based on local Fourier analysis." In Proc. IEEE Intl. Conf. Acoust. Speech Signal Process. (ICASSP) 2009, pp. 3004–3008.
- Chapter 6 appears in its entirety as:
  - D. Rudoy, P. Basu, and P. J. Wolfe, "Superposition frames for adaptive time-frequency analysis and fast reconstruction." IEEE Trans. Signal Process., vol. 58, pp. 2581–2596, 2010.

  Preliminary versions of material in Sections 6.6.1, 6.7, and 6.7.3 first appeared in
  - D. Rudoy, P. Basu, T. F. Quatieri, R. Dunn, and P. J. Wolfe, "Adaptive short-time analysis-synthesis for speech enhancement." In Proc. Intl. Conf. Acoust. Speech Signal Process. (ICASSP) 2008, pp. 4905–4908.
- Parts of Chapter 7 appeared in
  - M. A. Berezina, D. Rudoy and P. J. Wolfe, "Autoregressive Modeling of Voiced Speech," In Proc. IEEE Intl. Conf. Acoust. Speech Signal Process. (ICASSP) 2010, pp. 5042–5045.

Acknowledgments

I would have had neither the strength nor the will to start and complete my PhD studies without the love and support of my wife Melanie Rudoy. You were always there when I needed your support, and when I foolishly thought that I didn't. If I ever reached to just beyond my grasp, it was only because I stood on your shoulders... . In our lives intertwined with years in Philadelphia, Baltimore and Cambridge, from one morning to the next, with the joys and sorrows that lie in between, with growing fond memories, and eager anticipation of our unfolding adventure: it is here I love you.

Thank you to my advisor Prof. Patrick Wolfe for creating an incredible atmosphere for pursuing research, for setting the bar high by example, and for being a tireless collaborator. You have taught me to think creatively and deeply, to take risks, to trust my intuition and to carefully and meticulously proofread manuscripts, so that nary a serial comma would be missed. Thank you for taking a chance on me and for helping me reach my goals. When I think back on my graduate school experience, I will always first recall our Petsi meetings at seven a.m. and at Simons' during all hours of the day. I know you will be a lifelong mentor and friend.

Continuous support from my committee members Tom Quatieri and Prof. Vahid Tarokh was also very important to me. Tom, I will dearly cherish the many years of our lively discussions. Your love of speech science is addictive and refreshing. Thank you for all your patience and guidance. Thank you Prof. Tarokh for always being there when I needed advice for how to navigate the Harvard system.

My graduate studies were greatly enriched by my "SISL" labmates: Omar Abdala, Nick Arcolano, Betul Arda, Prabahan Basu, Ali Belabbas, Maria Berezina, David Choi, Keigo Hirakawa, Jung-Ook Hong, Jing Gu, Daryush Mehta, Patrick Perry, Daniel Spendley, Frank Tompkins, Yuhang Wang, and Chris White. Thank you for being a collective sounding board, it has been a privilege being in the trenches together. To my co-authors: Maria Berezina, Prabahan Basu, Prof. Jose Blanchet, Bob Dunn, Prof. Tryphon Georgiou, Prof. Rob Howe, Tom Quatieri, Daniel Spendley, and Shelten Yuen—your patience cannot be overemphasized, I've learned a great deal from all of you and cherish our collaboration.

Over the last eight years I have been fortunate to learn from my friends and colleagues at MIT Lincoln Laboratory. I am especially indebted to Keh-Ping Dunn who has been a tireless mentor, a vociferous advocate, and a friend. Thank you to Lou Bellaire, Kathy Bihari, Nancy Chen, Bob Dunn, Virginia Hafer, Liz Gustowt, Zahi Karam, Nicolas Malyska, Peter Parker, Andrew Silberfarb, Sung-Hyun Son, Tianyu Wang, James Williamson, Jake Wasserman, and many others in groups 32, 36, 62 and 104. My visits to LL over the last five years have helped me stay grounded and maintain my focus; I will miss the regular coffee breaks, enlightening conversations, and filled-in whiteboards.

There are many others that have helped to guide me over the years. I'm very grateful to Ankit Patel and Sourav Dey for talking me out of working on Markov chain theory or quantum mechanics; seriously, I would not have gone back to graduate school without your examples. The late-night stochastic process crew of Mehmet Akcacaya, Yves Chretien, and Peter Parker is the quintessential graduate school experience. Thank you to Tom Baran, Petros Boufounos, Julius Degesys, Sourav Dey, Shay Mamon, Prof. Radhika Nagpal, Prof. Alan Oppenheim, Ankit Patel, Charlie Rohrs, and all members of the Digital Signal Processing Group at MIT and the Self-Organizing Systems Research Group at Harvard for letting me be a "part-time member." And, a big thank you to my undergraduate advisor and mentors Profs. Max Mintz, Kostas Daniilidis and Sanjeev Khanna for instilling

in me the love for mathematics and teaching.

It would have been difficult to get through graduate school without the support of my family and great friends. Thank you to Mark and Shirley Leader, and Craig Shames for supporting Melanie and I through two PhDs! Thank you to Christi Electris, Kathryn Loving, Elizabeth Gustowt, Ankit Patel, Ziv Feldman, and Prof. Emily Fox. Thank you Ravi Goyal and Jaime Dufresne for your continual support; I have been lucky to draw strength from it over the years. Thank you for always being there for me.

Finally, I want to thank my parents. All the opportunities that I have been fortunate to have resulted from your sacrifices, perseverance and unwavering belief in me. This thesis is dedicated to you.

To my parents.

Chapter 1

Introduction

1.1 Problem Statement and Motivation

In this thesis, we develop a number of models and inference algorithms for the analysis of nonstationary time series, study their properties, and apply them to a variety of problems arising in speech signal processing. Speech and audio waveforms are canonical examples of nonstationary signals with slowly time-varying statistics, and although constructing accurate models of these temporal variations is a challenging task, it is extremely valuable for a broad range of applications.

The acoustic theory of speech production provides great insight into the physiological causes of variation. The classic source-filter model posits that a sound is produced in two stages: air is first modulated by the vocal cords, resulting in an excitation waveform (source) that is subsequently shaped by the resonant cavity between the vocal folds and the lips (vocal tract). In turn, temporal variations in the signal statistics can be explained by the variations within the source waveform (e.g., due to the changing rate of vocal cord vibration or pitch) and the vocal tract (e.g., due to continuous movement of the jaw, tongue, teeth and other articulators). Since such physical movement takes time to effect, it has been noted that the source waveform and the vocal tract shaping it are *on average* approximately time-invariant on the 15–30 ms time scale. Consequently, at this scale, the speech waveform can be modeled by a stationary stochastic process.

These observations, made early on in the history digital speech processing, have strongly influenced subsequent signal processing approaches. In particular, most modern applications, including speech recognition, coding and enhancement, rely on a common "front-end" architecture in which the signal is divided into a sequence of 15–30 ms short-time segments and various stationary time-series modeling techniques—especially spectral estimation methods—are then applied to extract information from each one.

However, the speech signal is not, in fact, a piecewise-stationary process at an *apriori*-determined time scale. For instance, plosives such as /b/ and /p/ (as in "bat" and "pat") are burst-like sounds that last for 5 ms, whereas sustained vowels can be as long as 100 ms. Some sounds result from the continuous motion of the articulators (e.g., slow opening of the jaw in /ay/ as in "my"), or from the slowly-varying fundamental frequency of vocal cord vibration (pitch).

Nevertheless, the classical front-end architecture is still largely in use, unchanged. But, many in the speech processing community agree that progress toward more accurate time-frequency modeling of the speech signal is needed and, if successful, will have simultaneous ramifications for multiple applications. Indeed, methods for nonstationary modeling of speech signals have recently received more attention in the literature, as evidenced by the many references given throughout the thesis.

Our goal is to continue along these lines, and propose a number of models and methods for improved modeling of the speech signal and the time-varying characteristics of the vocal tract and the source waveform. Specifically, we extend traditional methods including autoregressive modeling and short-time Fourier analysis, which form the basis of many speech and audio processing algorithms, by relaxing the assumption that the signal is stationary at some small, fixed scale. We also revisit the well-known problems of estimating vocal tract resonances and the time-varying source waveform. Throughout, we have aimed to design methods that are simultaneously computationally efficient, can be applied to many different problems in speech analysis, and whose properties can be formally analyzed.

1.2 Thesis Organization and Contributions

Below we provide a detailed summary of the contents and contributions described in each subsequent thesis chapter. The introductory paragraphs and sections of each chapter provide more detailed outlines.

Chapter 2: Background

We begin with an overview of the acoustic theory of speech production via the source-filter model. This classic, physiologically-faithful model posits that a speech waveform is produced in two stages: air from the lungs is first modulated by the vocal cords, resulting in a source waveform that is subsequently shaped by the physical characteristics of the vocal tract (filter) comprising the resonant cavity between the vocal folds and the lips, jaws, teeth, tongue, and other articulators. We review a number of models for the source waveform and the vocal tract, with a particular focus on how autoregressive or all-pole models capture the salient characteristics of the latter, and delineate associated statistical inference methods. Finally, we consider the sources of temporal variability in light of the source-filter framework, which in turn motivate many of the signal modeling techniques for nonstationary time series developed in this thesis. We conclude with a review of pertinent properties of stationary and nonstationary stochastic processes.

Chapter 3: Formant Tracking

Our first contribution in the direction of modeling the temporal characteristics of speech is the development of a method for estimating the trajectories of the vocal tract resonances—often termed formants—from an observed acoustic waveform, under the standard approximation of vocal tract time-invariance on the segmental (15–30 ms)

scale. Formants play a central role in the perception and analysis of speech, and tracking their locations has been a problem of interest in the speech community for over four decades. In contrast to previous approaches that rely on auxiliary algorithms such as root-finding or peak-picking, we extend earlier contributions of Deng [1,2] and develop an elegant state-space model for formant evolution, with inference realized via extended Kalman smoothing. Our approach enables the estimation of vocal tract anti-resonance (anti-formant) trajectories, which are essential for accurate spectral modeling of nasalized sounds. We propose both offline and online system identification algorithms that allow for explicit modeling of correlation structure across formants and anti-formants, extend the model to account for the absence of waveform energy during silences, and adopt a Taylor-based linearization of the formant-to-cepstrum map. Among various illustrations, results show that the proposed algorithms accurately track formants and anti-formants both in synthesized and natural speech. In particular, evaluations using a recently-constructed public database of hand-labeled formant trajectories indicate reduction of the root mean-square error relative to a benchmark formant analysis technique of up to 30% per formant.

Chapter 4: Time-Varying Autoregressive Modeling

Traditional methods that form the basis of many speech and audio processing algorithms, including autoregressive modeling and short-time Fourier analysis rely on the standard assumption that the speech signal is piecewise-stationary at a fixed (15–30 ms) time scale; we used this assumption, for instance, in our development of formant tracking algorithms in Chapter 3. In contrast, the goal of Chapters 4, 5 and 6 is to generalize and develop autoregressive modeling and short-time Fourier analysis, in the absence of this constraint, and to show that the resultant models lead to more accurate signal representations and improved algorithms for a variety of applications.

In Chapter (4), we further develop the theory of time-varying linear prediction by studying time-varying autoregressive (TVAR) processes. After reviewing the well-known class of TVAR models whose coefficient trajectories are modeled using flexible basis function expansions, we derive a closed-form expression for the covariance structure of a TVAR process, show how to apply the results to computing Cramér-Rao lower bounds, and obtain a new way of visualizing time-frequency content. We derive a new estimator of the TVAR coefficients in the case when only noisy observations are available, and develop a time-varying lattice formulation that enables us to constrain the "instantaneous" poles of the estimated process to lie inside the unit circle, which leads to frozen-time stable inverse-prediction-error autoregressive filters. In addition, two new approaches for modeling TVAR coefficient trajectories are introduced, including an approach that blends the functional-expansion approach with a stochastic evolution model, and a geometric approach based on viewing each TVAR process as a path on the manifold of AR processes, with estimators realized via convex optimization.

Chapter 5: Parametric and Nonparametric Tests for Stationarity

Prior to applying statistical techniques designed for nonstationary stochastic processes, such as time-varying AR processes, it is important to confirm whether or not the data in question is, in fact, nonstationary. To this end, we introduce parametric and nonparametric tests to establish whether a time-series of N observations contains sufficient evidence to reject the null hypothesis of stationarity, and apply these tests to problems in speech analysis and enhancement.

In the parametric case, we derive a generalized likelihood ratio test (GLRT) based on TVAR models and use it in order to detect the presence of temporal vocal tract variation in speech waveform data. We show that the GLRT can be efficiently realized and study its properties empirically for short data records, and asymptotically, leading to constant false alarm rate hypothesis tests. Next we develop two algorithms, based on the GLRT, for identifying vocal tract variations at different time scales. At the segmental level we demonstrate the sensitivity of the GLRT to vocal tract variations in whispered and voiced speech, and at the sub-segmental scale, we used it to identify both glottal opening and closuring instants. In order to test for stationarity in applications in which no suitable parametric model is available (e.g., music), we propose a method based on the characterizing the temporal variations of the empirical short-time Fourier coefficients. Our main contribution is an efficient Monte Carlo method based on the Wold representation for characterizing the null distribution of two associated test statistics. The approach is illustrated using synthetic and audio examples, and subsequently used as a method of obtaining a signal-adaptive variable-resolution Fourier analysis and a corresponding signal enhancement scheme.

Chapter 6: Superposition Frames

Chapters 4 and (5) consider TVAR modeling of speech in an effort to relax the piecewise-stationarity assumption implicit in the traditional application of linear prediction to speech analysis. Here, we address the analogous question in the nonparametric setting—how can standard fixed-resolution short-time Fourier representations be generalized in order to adapt to a signal's time-frequency structure? To this end, we introduce a family of linear, signal-adaptive time-frequency representations termed superposition frames, and show that they admit a number of efficient reconstruction methods. Though our construction is straightforward, proceeding via local signal-adaptive modification of a Gabor frame, we show it has nontrivial properties including a preservation of the original lower frame bound, a generalized constant overlap-add property that avoids explicit computation of dual windows and a means of generating new families of adaptive lapped and dyadic frames. In addition, we propose specific signal adaptation schemes based on greedy selection and dynamic programming, respectively. The framework is applied in the context of speech enhancement in order to highlight potential uses of the approach.

Chapter 7: Semiparametric Source-Filter Modeling

Formant tracking, time-varying linear prediction, and adaptive short-time Fourier—the techniques considered in the last four chapters—are designed to capture the time-

varying characteristics of the vocal tract. In this chapter, we present a number of contributions to the complementary problem of modeling the temporal variation in the source waveform during voicing. Specifically, we extend the classical linear prediction framework by incorporating model of source waveform via sparse nonparametric wavelet regression in order to take into account its quasi-periodic nature. The resultant model admits efficient linear estimators for the vocal tract transfer function, glottal flow and aspiration noise power, and exhibits robustness to pitch variation. Further, we propose algorithms for data-dependent subspace selection based on wavelet shrinkage, and consider the problem of voicing detection via hypothesis tests realized via the GLR, Wald and Rao test statistics. We show that the resultant family of models may be applied to a number of problems in speech analysis ranging from vocal tract and source-harmonics-to-noise ratio estimation to inverse filtering and voicing detection.

Chapter 8: Contributions and Recommendations
In this last chapter we summarize our contributions and outline a number of questions raised in this thesis. Even though many of the contributions to nonstationary time-series modeling were made with an eye toward applications in speech signal processing, many of the proposed models and techniques may prove useful in a broader setting and for a variety of problems related to nonstationary time-series analysis.

Chapter 2

Background

In this chapter, we first review the physiology of speech production, the classical models for speech processing it motivates, and associated inference algorithms. Second, we describe the types of temporal variability in speech waveforms and their physiological causes, which motivates many of the signal modeling techniques for nonstationary time series developed in this thesis. Last, we review relevant properties of stationary and nonstationary stochastic processes that will be used throughout. A more-in-depth treatment of some of the topics related to speech science may be found in [3], to spectral analysis in [4], and to stochastic processes in [5].

We begin in Section 2.1 by describing the physical principles governing speech production and the source-filter model—a classic, physiologically-faithful model that captures the salient features of the speech production system. The model posits that a speech waveform is produced in two stages: air is first modulated by the vocal cords, resulting in a source waveform which is then shaped by the resonant cavity between the vocal folds and the lips (vocal tract or filter). After discussing a number of source waveform models in Section 2.2, we describe, in Section 2.3, how autoregressive models capture the salient characteristics of the vocal tract, and review associated parameter estimation and model selection methods in Section 2.4. In Section 2.5, we consider the sources of temporal variability in speech waveforms, and, conclude with a review of pertinent properties of stationary and nonstationary stochastic processes in Section 2.6.

2.1 Source-Filter Model of Speech Production

Voice is an acoustic pressure wave that originates with the expulsion of air from the lungs. As the wave travels toward the lips, its characteristics are shaped by the speaker's anatomy and deliberate, controlled action of the vocal folds, velum, tongue, jaw, and lips. Extensive study of this speech production chain over the last half century has given rise to physiologically-faithful mathematical models of the speech waveform. The best-known one among them is the source-filter model of speech production and is the subject of the present section. This model is fundamental to speech processing and forms the basis of widely-used algorithms in speech recognition, synthesis, coding, and enhancement [3].

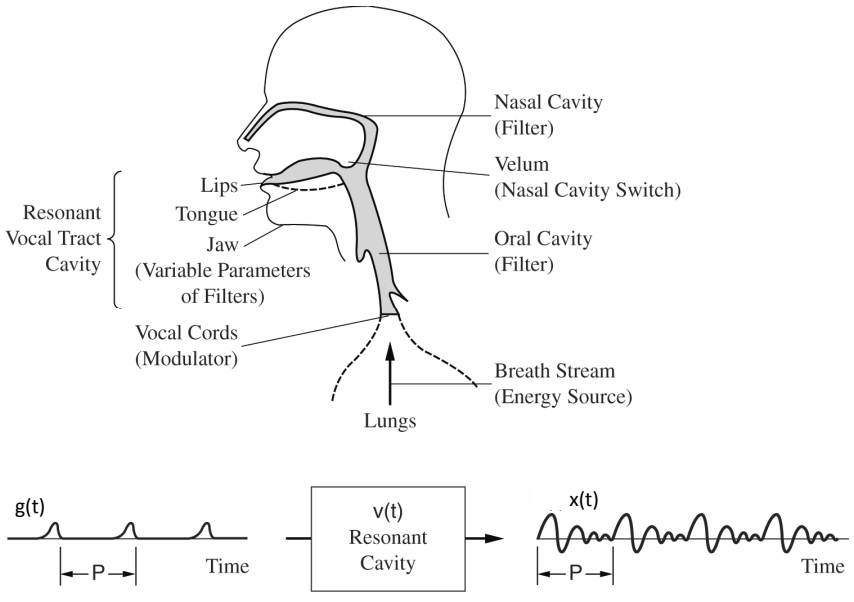


Figure 2.1: Anatomy of the human vocal system (top) and the sequence of events in the production of a steady-state vowel (bottom): a periodic waveform $g(t)$ excites a resonant vocal tract cavity resulting in the speech signal $x(t)$.

SOURCE: T. F. Quatieri, Discrete-Time Speech Signal Processing Principles and Practice [3]. ©2002, Prentice Hall PTR. Used by permission.

2.1.1 Physiology

The source-filter model is based on a linear approximation to the system of differential equations that describes the propagation of sound through the vocal system, which is shown in the top of Figure 2.1. To motivate and describe existing acoustic models, we first describe the physiology of speech production.

Consider the chain of events in the production of voiced sounds, such as the steady-state vowel [ɑ] (as in "father"). The process starts with a contraction of the diaphragm and other muscles in the rib cage that forces the expulsion of air from the lungs upward into the trachea. The expelled air travels through the glottis—an opening between two masses of vibrating tissue known as the vocal folds. The periodic opening and closing of the vocal folds modulates the glottal area and consequently the volume velocity of the airflow through the glottis.

In this manner, as the steady stream of air from the lungs exits the glottis and enters the oral cavity, it is transformed into a series of periodic puffs of air. As shown in the bottom of Figure 2.1, the puffs of air are separated by the pitch period P, which is inversely proportional to the frequency of vocal fold oscillation. The resulting wave is referred to as either the glottal airflow volume velocity, the source, or simply the excitation waveform.

Next, the source waveform excites the vocal tract—a resonant cavity that com-

prises the space between the glottis to the lips and shapes the modulated airflow through the motion of the velum, jaw, and tongue, as illustrated in Figure 2.1. In order to produce the vowel [ɑ], these articulators are in a specific configuration endowing the vocal tract with a certain set of resonant frequencies (formants) that shapes the spectral content of the excitation. The velum typically closes and separates the vocal tract from the nasal cavity, but during the production of nasal sounds—such as [m] and [n] (as in "gum" and "gun," respectively)—it also becomes part of the vocal tract and shapes the spectral content of the acoustic waveform.

The sequence of events for the production of *voiced* sounds described above changes during the production of whispered and *unvoiced* sounds. Specifically, the vocal folds are no longer vibrating and modulating the glottal airflow from the lungs. The source waveform is no longer periodic, but is instead noise-like and turbulent; any distinguishing characteristics of the sound are imparted by the vocal tract. For instance, the unvoiced fricatives [h], [s], and [f] (as in "heat", "seat," and "feat," respectively) are generated by forcing the turbulent source waveform past obstructions in the vocal tract formed, respectively, by the epilarynx, tongue, and teeth. Unvoiced plosives such as [p] and [t] (as in "pen" and "ten," respectively) are generated using a time-varying constriction: all airflow is blocked behind a constriction, leading to a buildup of air pressure. The sound is produced when the pressure is released.The excitation may also have both a periodic and a turbulent component; canonical examples include the voiced fricatives [v] and [z] (as in "voice" and "zoo," respectively), as well as the voiced plosives [b] and [d] (as in "bay" and "day," respectively).

Characteristics of the source waveform and the vocal tract it excites may change during the production of certain phonemes. Formants (vocal tract resonances) move during transitional sounds such as diphthongs (e.g., [eɪ] as in "they", moving jaw), glides (e.g., [w] as in "why", moving lips), liquids (e.g., [r] as in "room", moving tongue), and within plosives (e.g., [b] as in "boy", removal of a constriction). Moreover, the vocal tract configuration changes during transitions between different sounds, in general. The source waveform may vary as a function of changing pitch or through the introduction of aspiration.

In all the above examples, the spectral content of the excitation waveform is modified by the resonances of the oral cavity. Thus, the vocal tract can be thought of as "shaping" or "filtering" the source waveform, giving rise to the terminology *source-filter* model. As we will discuss below, this model is based on two central assumptions:

- **Independence**: The effect of subglottal physiology on the sound wave is independent of its subsequent spectral shaping by the vocal tract.
- **Linearity**: The relationship between the glottal volume velocity and the air pressure at the lips is linear.

Independence implies that the excitation waveform (source) and the vocal tract resonances (filter) may be modeled separately, while linearity suggests modeling the vocal tract as a linear operator. In a digital signal processing setting, these criteria hint at modeling the vocal tract as a linear time-varying system excited by the periodic, turbulent or impulsive glottal volume velocity waveform. Indeed, as we discuss in Section 2.1.2 below, acoustic modeling of sound propagation through the vocal system leads precisely to this model, which we describe in Section 2.1.3.

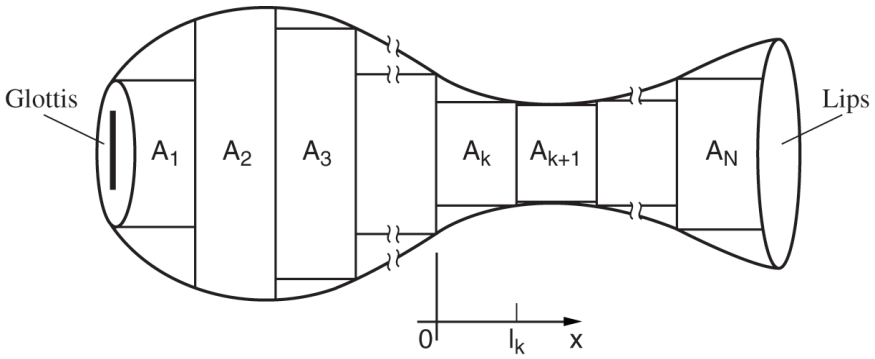


Figure 2.2: Concatenated tube model of the vocal tract: its smoothly-varying cross-section is approximated piecewise by a series of N lossless uniform acoustic tubes.

SOURCE: T. F. Quatieri, Discrete-Time Speech Signal Processing Principles and Practice [3]. ©2002, Prentice Hall PTR. Used by permission.

2.1.2 Acoustic Modeling

A complete mathematical characterization of how the source waveform propagates through the oral cavity comprises the pressure and velocity of air particles as a function of position and time. It is typically obtained as a solution of an appropriate wave equation that takes into account the physical characteristics of the propagation medium that is often modeled as a homogenous acoustic tube with space- and time-varying cross-section. Then, the relevant wave equations are given by [6]:

$$\begin{aligned} -\frac{\partial p(x,t)}{\partial x} &= \rho \frac{\partial (u(x,t)/A(x,t))}{\partial t}, \\ -\frac{\partial u(x,t)}{\partial x} &= \frac{1}{\rho c^2} \frac{\partial \left(A(x,t)p(x,t)\right)}{\partial t} + \frac{\partial A(x,t)}{\partial t}, \end{aligned} \tag{2.1}$$

where ρ is the density of air in the tube, c is the velocity of sound, $A(x,t)$ is the area function of the tube, and $p(x,t)$ and $u(x,t)$ are the variations in sound pressure and airflow volume velocity at position x and time t, respectively. Appropriate boundary conditions describing the radiation of sound at the lips and the impedance at the glottis also need to be incorporated (see e.g., [7] for a thorough discussion). It should not come as a surprise that solving the system of coupled partial differential equations in (2.1) and related generalizations (that incorporate, e.g., the effects of viscosity or the formation of vortices during aspiration) requires computationally-demanding numerical procedures.

A widely-used alternative to the differential-equation approach, which avoids this computational burden, is the concatenated tube model. The basic approach is to approximate the vocal tract by a series of short, lossless acoustic tubes as shown in Figure 2.2; each tube has length l_k and cross-sectional area of A_k. Furthermore, it is assumed that energy losses occur only at the lips due to sound radiation from the oral cavity into ambient space; it is typical to assume infinite impedance at the glottis to maintain linearity [3]. Clearly, the concatenated tube model does not capture all physical aspects of sound production such as

energy losses due to wall vibration, thermal conduction and viscosity. But it constitutes a simple model of the vocal tract that has been shown to faithfully approximate the solution to the system of differential equations that takes these quantities into account [6].

Solving the differential equations for sound propagation in each of the concatenated acoustic tubes—while ensuring that the pressure and volume velocity are continuous at the junctions, and incorporating boundary conditions and the glottis and lips—enables us to quantify the relationship between the airflow volume velocity at the glottis and the lips. Furthermore, if we consider the impulse response of an N-tube model whose parameters (areas and lengths) are time invariant, then the relationship between input and the output at the lips can be represented in the frequency domain as the (approximate) transfer function of the vocal tract. Discretizing the resultant equations in space and time leads to the following all-pole representation of the vocal tract in the $\mathcal{Z}$-domain [3, 7, 8]:

$$V(z) \triangleq \frac{\sigma}{\sum_{i=1}^{p/2}(1-\alpha_i z^{-1})(1-\overline{\alpha_i} z^{-1})}, \tag{2.2}$$

where σ is the gain, $(\alpha_i, \overline{\alpha_i})$ is a complex-conjugate pole pair and p is the number of acoustic tubes. Note that by increasing the number of acoustic tubes in the model—and consequently the number of poles in the denominator of (2.2)—we get a more faithful representation of a spatially-varying vocal tract by its transfer function.

2.1.3 Digital Filter Realization

In this section we describe a complete digital filter realization of the speech production chain. In the context of concatenated tube modeling, we have already seen that the frequency response of the vocal tract is well-approximated by an all-pole transfer function. In order to incorporate the effect of boundary conditions, we describe frequency-domain models for the glottal volume velocity and the effect of sound radiation at the lips.

Following [3], we note that a rough idealization of the glottal airflow $g[n]$ over a *single* pitch period (i.e., a single glottal pulse) is given by a convolution of two time-reversed exponentially decaying sequences, for some $|\beta| < 1$:

$$g[n] \approx \left(\beta^{-n}u[-n]\right) * \left(\beta^{-n}u[-n]\right), \tag{2.3}$$

which, in the $\mathcal{Z}$-domain, corresponds to

$$G(z) = \frac{1}{(1-\beta z)^2}.$$

Here $G(z)$ is a maximum-phase all-pole transfer function with two identical poles outside the unit circle. An example of a glottal pulse following the model of (2.3) is shown in the left panel of Figure 2.3. A simple model of the glottal airflow over *multiple* pitch periods can then be obtained by convolving the glottal pulse $g[n]$ with a periodic impulse train, with the spacing between impulses equal to the pitch period, as shown in the right panel of Figure 2.3.

The radiation impedance at the lips may be modeled by a filter whose transfer function has a single zero on the unit circle: $R(z) = 1 - z^{-1}$ because the acoustic speech

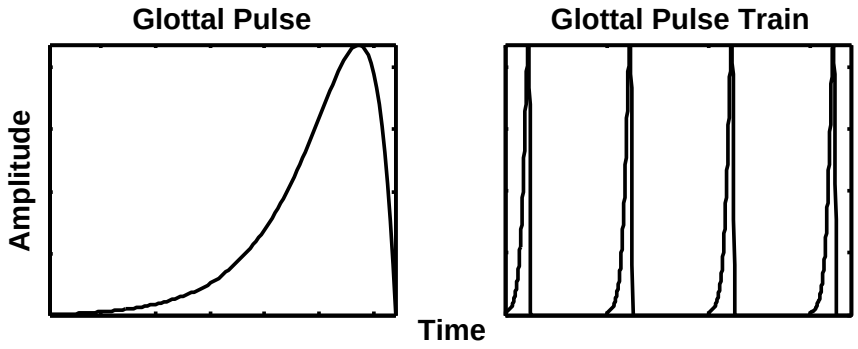


Figure 2.3: A glottal pulse modeled by two maximum-phase poles in the $\mathcal{Z}$-domain (left) is convolved with a periodic impulse train to yield a glottal volume velocity waveform during voicing (right).

pressure waveform radiated at the lips is approximately a derivative of the glottal volume velocity. Often an approximation with a zero slightly inside the unit circle as $R(z) \approx 1-\gamma z^{-1}$ is used [3]. The resulting frequency-domain model of the radiated pressure waveform $x[n]$ over a single pitch period is therefore given by:

$$X(z) \approx \sigma G(z)V(z)R(z) = \frac{\sigma(1-\gamma z^{-1})}{(1-\beta z)^2 \prod_{i=1}^{p/2}(1-\alpha_i z^{-1})(1-\overline{\alpha_i} z^{-1})}, \tag{2.4}$$

where the constant σ is a gain controlling loudness. The discrete-time difference equation corresponding to (2.4) is given by:

$$x[n] \approx \sigma g[n] * v[n] * (\delta[n] - \gamma\delta[n-1]) = \sigma(g[n] * (\delta[n] - \gamma\delta[n-1])) * v[n]. \tag{2.5}$$

Note that the linearity assumption of the source-filter model allows us to combine the effect of lip radiation together with the glottal pulse as in the second equality of (2.5); the resultant source waveform is often termed the glottal flow derivative.

The difference equation for the case of multiple pitch periods is given by

$$x[n] \approx \sigma g[n] * p[n] * v[n] * (\delta[n] - \gamma\delta[n-1]) = \sigma(g[n] * (\delta[n] - \gamma\delta[n-1])) * p[n] * v[n], \tag{2.6}$$

where $p[n]$ is a periodic impulse train. The complete discrete-time model is illustrated in Figure 2.3. Notice that the effect of radiation at the lips can be incorporated either into the source (glottal flow) or the filter (vocal tract). The model of (2.5) may also be specialized to the case of unvoiced or whispered speech according to:

$$x[n] \approx \sigma w[n] * v[n] * (\delta[n] - \gamma\delta[n-1]), \tag{2.7}$$

where $w[n]$ is the stochastic input representing turbulent noise at the glottis.

Finally, observe that the source-filter model in Figure 2.3 allows for the glottal airflow to be modeled as a linear combination of periodic and stochastic inputs. This is very useful in modeling the production of voiced fricatives and aspiration or breathiness that results from the vocal folds closing partially, rather than fully, during voicing.

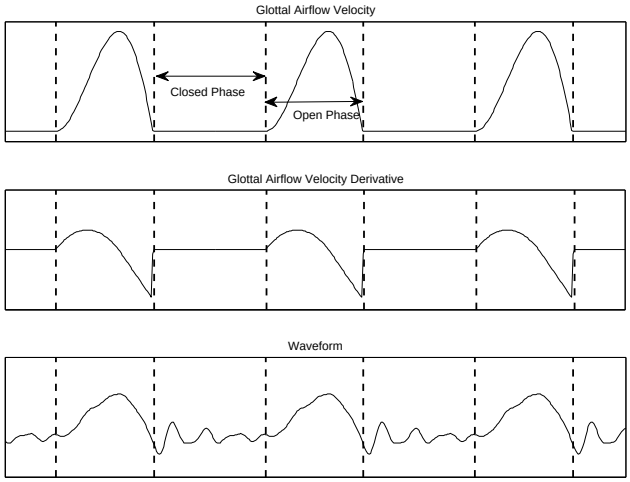


Figure 2.4: The glottal airflow closed and open phases demarcated in three pitch periods of a vowel, and superimposed on the glottal volume velocity (top), glottal volume velocity derivative (middle), and radiated pressure waveform (bottom) waveforms.

2.2 Source Models

Having broadly outlined the source-filter model, we turn to a deeper study of source and filter modeling, and the associated estimation methods. Here we focus on source modeling and begin with a more refined description of glottal flow dynamics during voicing.

2.2.1 Speech Production within a Single Voicing Period

If the glottal volume velocity were to be measured at the glottis, then a typical waveform would resemble the one shown in the top panel of Figure 2.4. The associated glottal airflow derivative—resulting from incorporating the effect of radiation at the lips into the excitation—is shown in the middle panel of Figure 5.9. Here, each pitch period is subdivided into two parts, termed the closed and open phases. The moment at which the vocal folds close, commonly referred to as the glottal closure instant (GCI), indicates the beginning of the closed phase in each pitch period. During this phase, the glottal volume velocity is zero and, as seen in the bottom panel of Figure 2.4, the acoustic output at the lips appears as exponentially damped oscillations. At the glottal opening instant (GOI), indicating the start of the open phase, the vocal folds gradually begin to open until the airflow velocity reaches its maximum amplitude; thereafter they start to close—demarcating a region sometimes called the "return phase" [3]—and shut at the next GCI.

The above description of the pitch period assumes that the vocal folds shut completely during the closed phase—known as *modal* phonation in speech science. However, during the production of certain phonemes (e.g., voiced fricatives) the vocal folds may remain partially open. This *non-modal* phonation is common in aspirated or breathy voices and, in extreme cases, could be indicative of vocal system abnormalities and disease (e.g., dysphonia). Indeed, it is well-known that partial abduction of the vocal folds is necessary

in order to synthesize natural-sounding speech [9].

During the production of unvoiced sounds, characteristics of the glottal volume velocity also depend on the state of the vocal folds including their tension and glottal area.

2.2.2 Existing Models

Mathematical models of the source waveform readily divide into two groups: the first comprises physics-based models of vocal-fold vibration that attempt to capture the mechanical forces relating to muscular tension. Examples of the former include the two-mass model of [10, 11] and the three-mass model of [12]; combining these with a model of airflow from the lungs to the glottis yields a model of the glottal volume velocity. The second class consists of models that attempt to match directly the shape of the glottal volume velocity (or its derivative, if the radiation at the lips is included), and is our focus in this section.

We have seen that during voicing the glottal volume velocity can be modeled simply as a convolution of a glottal pulse with a periodic impulse train whose impulses are regularly spaced by the pitch period, or irregularly spaced to reflect any temporal variation in pitch. Consequently, only a model for the glottal pulse (or its derivative) and an estimate of the pitch period are needed to yield a model of the glottal airflow over the duration of the entire utterance.

The Rosenberg [13] and the Liljencrantz-Fant (LF) [14] are two examples of widely-used models of the glottal pulse. The former uses second- and third-order polynomials to model the opening and closing phases of the glottal volume velocity. However, no model of the return phase is incorporated and, consequently, the resulting transition from the open to the closed phase is abrupt. The LF model of the glottal flow derivative addresses this concern by modeling the return phase with a decaying exponential, thereby allowing for a smooth transition between the open and closed phases. The exact model is given by

$$v_{\mathrm{LF}}[n] \triangleq \begin{cases} E_0 e^{\alpha n} \sin(\omega_g n) & 0 \le n < N_e \\ -\frac{E_e}{\epsilon N_a}\left(-e^{-\epsilon(n-N_e)} - e^{-\epsilon(N_c-N_e)}\right) & N_e \le n < N_c \\ 0 & N_c \le n < N_o \end{cases}, \tag{2.8}$$

where the synthesis parameters $\{E_0, E_e, \alpha, \omega_g, \epsilon\}$ describe the glottal pulse shape, and the timing parameters N_e, N_c, N_a and N_o are the instant of the maximum negative glottal flow, duration of the return phase, glottal closing, and glottal opening, respectively. An example of a glottal pulse synthesized using the LF model is shown in Figure 2.5. Similar models are described in [15–17]; subsequent refinements include the KLGLOTT88 model [18] and the R++ model [19];

An alternative to the parametric models described above is the nonparametric modeling approach initiated by [20] and improved upon by [21] and [22]. Here the glottal volume velocity $u_g[n]$ is modeled using a polynomial expansion according to:

$$u_g[n] = \sum_{j=1}^{q} f_j[n; N_o, N_c],$$

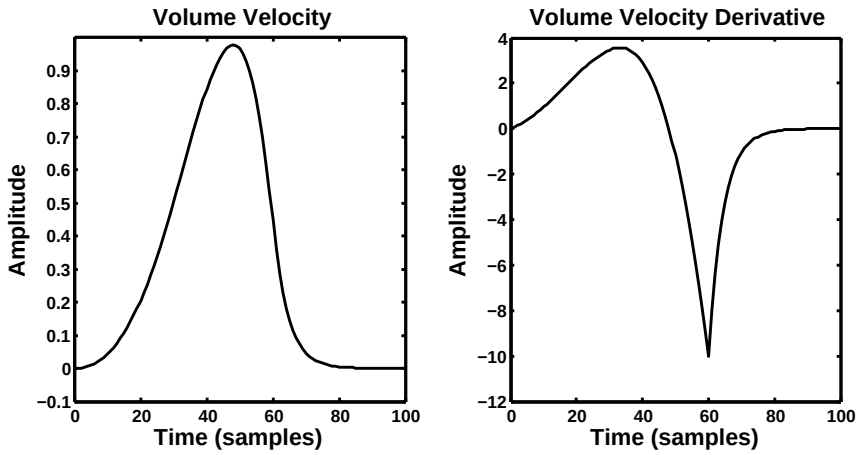


Figure 2.5: Example Liljencrantz-Fant model of the glottal volume velocity derivative (right) shown with the associated volume velocity (left).

where $\{f_j[n; N_o, N_c] \,|\, 1 \leq j \leq q\}$ is a set of suitably-selected polynomials and N_o and N_c are the glottal opening and closing instants, respectively. These nonparametric models allow for simpler estimators than the Rosenberg or LF models and are more flexible than the rigid parametric forms. They foreshadow our nonparametric approach fully developed in Chapter 7.

Source models for unvoiced and whispered speech are considerably simpler—most use white or colored noise sequences. More in-depth models can be obtained by using fluid-dynamics models for the glottal airflow through the vocal folds, which would allow to faithfully model the effects of turbulence and incorporate vortices (see e.g., [23] for an overview).

During the production mixed excitation sounds such as voiced fricatives or breathy sounds, models for voiced and unvoiced components are combined. But their combination is not always realized via a linear superposition of signal in the time domain. One of the more interesting models in this setting, described in [24], proposes to model the aspiration waveform via an amplitude-modulated white noise sequence—the modulation is proportional to the glottal airflow due to the oscillation of the vocal folds.

2.3 All-Pole Modeling of the Vocal Tract

Having discussed a variety of source models in Section 2.2, we focus on vocal tract modeling and study the all-pole vocal tract filter of (2.2). We describe the relationship among all-pole modeling, linear prediction and autoregressive processes, and study a number of equivalent, parameterizations of autoregressive processes that play an important role in a variety of speech processing applications. Then, in Section 2.4, we discuss methods for estimating vocal tract parameters from acoustic data.

The source-filter model and autoregressive modeling are closely linked. To see this, let the source waveform $u_g[n]$ comprise the glottal volume velocity (over an arbitrary number

of pitch periods) and take into account the radiation at the lips. Since $X(z) = U_g(z)V(z)$ and (2.2) together imply

$$X(z)\left(1 - \sum_{i=1}^{p} a_i z^{-i}\right) = \sigma U_g(z), \tag{2.9}$$

by taking the inverse $\mathcal{Z}$-transform of both sides and rearranging terms gives rise to the following discrete-time difference equation for the source-filter model of speech

$$x[n] = \sum_{i=1}^{p} a_i x[n-i] + \sigma u_g[n]. \tag{2.10}$$

2.3.1 Linear Prediction and Autoregressive Models

Many important properties of (2.10) are revealed under the assumption that $u_g[n] = w[n]$—a zero-mean white Gaussian noise sequence with unit variance, scaled by the gain parameter $\sigma > 0$, leading to:

$$x[n] = \sum_{i=1}^{p} a_i x[n-i] + \sigma w[n]. \tag{2.11}$$

The discrete-time difference equation of (2.11) is known as an autoregressive (AR) process and is a basic tool in time-series analysis that has been widely applied to areas ranging from economics to signal processing [4, 25, 26]. When required, the notation AR(p) is used in order to make explicit the dependence of the process on the last p lags.

Since $w[n]$ is a zero-mean Gaussian sequence, $x[n]$ is also a zero-mean Gaussian process, and its characteristics are completely determined by its second-order statistics as specified by the autocorrelation function of (2.11). The autocorrelation function is defined according to $r_{xx}[\tau] \triangleq \mathrm{E}(x[n]x[n-\tau]$ for all $\tau \in \mathbb{Z}$. Using this definition together with (2.11) it is easy to show that

$$r_{xx}[\tau] = \sum_{i=1}^{p} a_i r_{xx}[\tau - i], \tag{2.12}$$

for all $\tau \geq 1$. The recursive relationship of (2.12) is fundamental to many estimators of AR parameters that we discuss below.

Viewed from the perspective of signal processing, the difference equation of (2.11) represents a linear system with an infinite impulse response. The associated $\mathcal{Z}$-domain transfer function takes the form

$$H(z) \triangleq \frac{\sigma}{1 - \sum_{i=1}^{p} a_i z^{-i}}.$$

If all the zeros of the polynomial in the denominator are inside the unit circle then $H(z)$ is a transfer function of a bounded-input bounded-output stable filter [4]. The difference equation of (2.11) implies that this filter is causal.

One example of an AR process, often referred to in this thesis, is the AR(2) process that corresponds to a second-order digital resonator—the simplest model for a vocal tract

resonance or formant. Consider a pair of complex-conjugate poles located at $(r, \pm\theta)$, in polar coordinates. The associate transfer function is given by:

$$H(z) = \frac{1}{1 - 2r\cos(\theta)z^{-1} + r^2 z^{-2}}; \tag{2.13}$$

the associated prediction coefficients are given by $a_1 = 2r\cos(\theta)$ and $a_2 = -r^2$. The parameters r and θ are often referred to as a bandwidth and center frequency of the associated second-order digital resonator.

Observe that even though assuming that the excitation sequence $w[n]$ in (2.11) is zero-mean Gaussian white noise is only appropriate for unvoiced or whispered speech, the analysis of the resulting AR(p) model is relevant for all cases. Indeed, as we will see in Section 2.4, many vocal tract parameter estimation algorithms are derived under the assumption of a noisy innovations sequence, yet can be (and, in practice, are) directly applied to voiced speech. We return to the issue of incorporating realistic source models later, in Section 2.4.6.

2.3.2 Orthogonal Realizations and Reflection Coefficients

The difference equation of (2.11) is a direct-form realization of the all-pole filter representing the vocal tract. Next, we discuss an orthogonal filter realization that gives rise to a parametrization of the vocal tract in terms of the so-called "reflection" coefficients, in order to gain a deeper understanding of the relationship between the all-pole and acoustic tube models.

To this end, fix an integer $j \geq 1$ and define $\{a_{i,j} \,|\, 1 \leq i \leq j\}$ and $\{b_{i,j} \,|\, 1 \leq i \leq j\}$ to be the sets of forward and backward linear prediction coefficients that minimize the squared errors of predicting $x[n]$ and $x[n-j]$, respectively. These forward and backward errors are defined according to:

$$e_j^{\mathrm{f}}[n] \triangleq x[n] - \sum_{i=1}^{j} a_{i,j}\, x[n-i], \tag{2.14}$$

$$e_j^{\mathrm{b}}[n] \triangleq x[n-j] - \sum_{i=1}^{j} b_{i,j}\, x[n-j+i]. \tag{2.15}$$

If $x[n]$ is a wide-sense stationary process, then it can be shown that the forward and backward minimum-mean-squared error (MMSE)—optimal linear prediction coefficients are related via $a_{i,j} = b_{j-i,j}$ [4]. We retain this assumption in our discussion of orthogonal realizations below.

The forward error $e_j^{\mathrm{f}}[n]$ of (2.14) is the error of approximating $x[n]$ by its projection onto the space spanned by $\{x[n-1], x[n-2], \ldots, x[n-j]\}$. However, the resultant expansion ($\widehat{x}[n] \triangleq \sum_{i=1}^{j} a_{i,j} x[n-i]$) is not orthogonal since (2.11) implies that the variables $\{x[n-1], x[n-2], \ldots, x[n-j]\}$ are dependent. As an alternative to this direct-form representation of $\widehat{x}[n]$, an orthogonal realization of $\widehat{x}[n]$ may be obtained by applying the Gram-Schmidt procedure to the variables $\{x[n-1], x[n-2], \ldots, x[n-j]\}$ in order starting from $x[n-1]$ (see,

e.g., [27]). This orthogonalization yields a recursive procedure for computing the optimal linear prediction coefficients of order j from those of order $j-1$ by:

$$a_{i,j} = \begin{cases} a_{i,j-1} + \kappa_j a_{j-i,j-1} & \text{if} \quad 1 \leq i < j \\ -\kappa_j & \text{if} \quad i = j \end{cases}, \tag{2.16}$$

where the coefficients κ_j are defined via:

$$\kappa_j \triangleq -\frac{\langle e_{j-1}^{\mathrm{f}}[n], e_{j-1}^{\mathrm{b}}[n-1] \rangle}{\rho_{j-1}} = -\frac{r_{xx}[j] + \sum_{i=1}^{j-1} a_{i,j-1} r_{xx}[i]}{\rho_{j-1}}, \tag{2.17}$$

with $\rho_j \triangleq \|e_j^{\mathrm{f}}[n]\|^2 = \|e_j^{\mathrm{b}}[n-1]\|^2$. Since $-\kappa_j$ measures the partial correlation between $x[n]$ and $x[n-j]$ conditioned on the values $\{x[n-1], x[n-2], \ldots, x[n-j+1]\}$ [28], these coefficients are often called partial correlation (PARCOR) coefficients.

It turns out that the PARCOR coefficients $\{\kappa_1, \kappa_2, \ldots, \kappa_p\}$ defined in (2.17) also appear in the engineering literature, especially in acoustics, optics, and seismic sensing, wherein they are called reflection coefficients. Generally, the term "reflection coefficient" refers to the ratio of reflected and incident wave amplitudes at a discontinuity in some wave propagation medium. In transmission line theory, for example, the reflection coefficient at the boundary between the jth and $(j+1)$th sections with impedances Z_j and Z_{j+1}, respectively, is given by [8]

$$\kappa_j = \frac{Z_{j+1} - Z_j}{Z_{j+1} + Z_j}.$$

In the case of a lossless acoustic tube consisting of p sections with cross-sectional areas $\{A_1, A_2, \ldots, A_p\}$—our model for the vocal tract—the reflection coefficients reduce to

$$\kappa_j = \frac{A_{j+1} - A_j}{A_{j+1} + A_j}. \tag{2.18}$$

One of the primary reasons for interest in linear prediction in the context of speech processing is that the (negative) PARCOR coefficients defined according to (2.17) are equal to the reflection coefficients of (2.18). In other words, linear prediction can be used to solve the inverse problem of inferring acoustic tube model parameters (e.g., vocal tract areas or reflection coefficients) from observed waveform data. These relationships are carefully documented in several textbooks [3,7,8] and an excellent survey article [29], among others. This underscores, once again, how closely linear prediction is linked to acoustic tube modeling of the vocal tract.

The reflection coefficients together with the autocorrelation function at lag 0 $\{r_{xx}[0], \kappa_1, \kappa_2, \ldots, \kappa_p\}$ also provide one of the many alternative representations of an AR(p) process, since they are in a one-to-one correspondence with $\{\sigma^2, a_1, a_2, \ldots, a_p\}$—the complete set of parameters of (2.11). We discuss this equivalence and additional representations of autoregressive processes next.

2.3.3 Equivalent Parameterizations

The most common parameter set in linear predictive analysis comprises the set of predictor coefficients $\{a_1, a_2, \ldots, a_p\}$ in (2.11) and the noise variance σ^2. Yet, it is often useful to transform these parameters into other, equivalent representations. In our discussion

of orthogonal realizations of all-pole filters, we already considered the relationships among the following three sets of parameters:

$$\text{AR Parameters:} \quad \boldsymbol{P}_1 \triangleq \{\sigma^2, a_1, a_2, \ldots, a_p\}, \tag{2.19}$$

$$\text{Autocorrelation Function:} \quad \boldsymbol{P}_2 \triangleq \{r_{xx}[0], r_{xx}[1], \ldots, r_{xx}[p]\}, \tag{2.20}$$

$$\text{Reflection Coefficients:} \quad \boldsymbol{P}_3 \triangleq \{r_{xx}[0], \kappa_1, \kappa_2, \ldots \kappa_p\}. \tag{2.21}$$

These three parameterizations are equivalent, and efficient algorithms are available for transforming among them. The autocorrelation function parametrization of (2.20), for instance, can be transformed to both the reflection and AR coefficient representations of (2.19) and (2.21), respectively, using the Levinson-Durbin recursion summarized in Algorithm 2.1.

Algorithm 2.1 Levinson-Durbin Algorithm

- Initialization:
$$a_1[1] = -r_{xx}[1]/r_{xx}[0] \qquad \rho_1 = (1 - |a_1[1]|^2) r_{xx}[0] \tag{2.22}$$
- For $j = 2, \ldots, p$
  - Compute reflection coefficient
$$\kappa_j = -\frac{r_{xx}[j] + \sum_{i=1}^{j-1} a_{i,j-1} r_{xx}[i]}{\rho_{j-1}}$$
  - Compute linear prediction coefficients $\{a_{i,j} \,|\, 1 \leq i < j\}$ of appropriate order using (2.16)
  - Update error variance
$$\rho_j = (1 - |\kappa_j|^2)\rho_{j-1}$$
- Set the innovations variance as $\sigma^2 = \rho_p$

The Levinson-Durbin recursion of Algorithm 2.1 may be reversed, resulting in the so-called step-down procedure, in order to calculate the first $p + 1$ lags of the autocorrelation function from either the AR or the reflection coefficients. A summary of different transformations among $\boldsymbol{P}_1$, $\boldsymbol{P}_2$ and $\boldsymbol{P}_3$ can be found in [4].

In addition to these parameterizations of an AR process, there is a number of other equivalent representations—each arising in various speech processing applications. Following [7], we summarize them here for completeness and briefly discuss their use.

- Roots of predictor polynomial: The predictor polynomial associated to the AR coefficients is
$$D(z) \triangleq 1 - \sum_{k=1}^{p} a_i z^{-i} = \prod_{i=1}^{p} \left(1 - \alpha_i z^{-1}\right),$$
where $\{\alpha_1, \alpha_2, \ldots, \alpha_p\}$ are the associated roots that are either real or come in complex-conjugate pairs. This parametrization is central to the analysis of linear systems and is used for studying the frequency response, stability, and observability of the underlying system.

Every complex-conjugate pole pair $(\alpha, \overline{\alpha})$ may be parameterized in terms of a frequency and bandwidth pair (f, b) according to

$$\alpha = \exp\left(-\pi b/f_s + 2\pi i f/f_s\right) \qquad \text{and} \qquad \overline{\alpha} = \exp\left(-\pi b/f_s - 2\pi i f/f_s\right),$$

where $i \triangleq \sqrt{-1}$ and f_s is the sampling frequency in Hertz. This parametrization of poles will be especially useful in our discussion of formant tracking in Chapter 3 of this thesis.

- Line spectral frequencies: Consider two polynomials $P(z)$ and $Q(z)$ formed by adding to or subtracting from $D(z)$ the associated time-reversed system function $z^{-(p-1)}D(z^{-1})$ according to

$$P(z) \triangleq D(z) + z^{-(p+1)}D(z^{-1}) \qquad \text{and} \qquad Q(z) \triangleq D(z) - z^{-(p+1)}D(z^{-1}).$$

  The line spectral frequencies correspond to the phases of the roots of these two polynomials that lie strictly within the open interval $(0, \pi)$. This parametrization is very popular in low-bit rate coders due to its favorable quantization properties [30].

- Infinite impulse response: The impulse response $\{h(n) | n \geq 0\}$ is obtained recursively via

$$h(n) = \sum_{i=1}^{p} a_i h(n-i) + \sigma\delta[n],$$

  under the assumption that $h(n) = 0$ for all $n < 0$.

- Linear predictive coding (LPC) cepstrum: The cepstrum of the impulse response $(h[n] \,|\, n \geq 1)$ denoted by $(c_n \,|\, n \geq 1)$ can be obtained recursively via

$$c_n = \begin{cases} a_1 & \text{if} \quad n = 1 \\ a_n + \sum_{i=1}^{n-1} \left(\frac{i}{n}\right) a_{n-i} c_i & \text{if} \quad 1 < n \leq p \\ \sum_{i=n-p}^{n-1} \left(\frac{i}{n}\right) a_{n-i} c_i & \text{if} \quad p < n. \end{cases}$$

  We rely heavily on this relationship in our treatment of formant tracking in Chapter 3, where we provide a detailed derivation.

- Log-area ratio coefficients: Let the jth and $(j+1)$th lossless tube in the acoustic tube model have cross-sectional areas of A_j and A_{j+1}, respectively. Then the jth log-area ratio coefficient γ_j is given by

$$\gamma_j \triangleq \log\left(\frac{A_{j+1}}{A_j}\right) = \log\left(\frac{1-k_j}{1+k_j}\right) = -2\tanh^{-1}(-k_j),$$

  where the first equality follows from (2.18). Clearly the log-area ratios are in one-to-one correspondence with the reflection coefficients. We will revisit this representation in our discussion of stability in time-varying systems in Chapter 4.

2.4 Vocal Tract Estimation

We now turn to the inverse problem of estimating vocal tract parameters from observed speech data. First, we assume that the glottal flow is noisy and turbulent, as is the case for unvoiced as well as whispered speech. Therefore, the glottal flow $g[n]$ is well-modeled by white Gaussian noise, reducing the model of (2.10) to the classical AR(p) process of (2.11). This simplification justifies the application of AR parameter estimation methods that are standard in statistical signal processing. Specifically, we describe maximum likelihood and method-of-moments estimators or, as they are known in the speech literature, the covariance and autocorrelation methods of linear prediction. We also discuss the exact maximum likelihood and reflection coefficient estimators and highlight the differences among these methods.

Next, in Section2.4.5, we explain how the above methods can be extended to work with voiced speech, when the glottal flow is periodic. In addition, we touch on the question of model order selection. Note that we focus only on batch estimation algorithms, which assume that the entire time-series has been observed. Most of the methods can be adapted to the sequential setting; see [31] for a good overview.

2.4.1 Conditional Maximum Likelihood Estimation

We begin by deriving a maximum-likelihood (ML) estimator for the AR coefficients $\boldsymbol{a} \triangleq \begin{pmatrix} a_1 & a_2 & \cdots & a_p \end{pmatrix}^T$ given a single time-series of N observations, partitioned according to

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_p & | & \boldsymbol{x}_{N-p} \end{pmatrix}^T \triangleq \begin{pmatrix} x[0] & \cdots & x[p-1] & | & x[p] & \cdots & x[N-1] \end{pmatrix}^T.$$

The *unconditional* joint probability density function of $\boldsymbol{a}$ and σ^2 is given by:

$$p(\boldsymbol{x}\,;\,\boldsymbol{a},\sigma^2) = p(\boldsymbol{x}_{N-p}\,|\,\boldsymbol{x}_p\,;\,\boldsymbol{a},\sigma^2)p(\boldsymbol{x}_p\,;\,\boldsymbol{a},\sigma^2), \tag{2.23}$$

but its maximization with respect to $\boldsymbol{a}$ results in a nonlinear (in the data) estimator that must be implemented via a computationally-demanding iterative procedure [4]. In practice, this issue is resolved by approximating the data likelihood of (2.23) by the *conditional* likelihood $p(\boldsymbol{x}_{N-p}\,|\,\boldsymbol{x}_p\,;\,\boldsymbol{a},\sigma^2)$, whose maximization yields an estimator that converges to the exact (unconditional) ML estimator as $N \to \infty$ if the model order p remains fixed. We return to the issue of exact ML estimation in Section 2.4.2 below.

The assumption that $w[n]$ in (2.11) is a zero-mean Gaussian process and the fact that a linear combination of Gaussian random variables is Gaussian imply that the conditional likelihood is given by

$$p(\boldsymbol{x}_{N-p}\,|\,\boldsymbol{x}_p;\boldsymbol{a},\sigma^2) = \frac{1}{(2\pi\sigma^2)^{(N-p)/2}} \exp\left(-\sum_{n=p}^{N-1}\frac{e^2[n]}{2\sigma^2}\right),$$

where $e[n] \triangleq x[n] - \sum_{i=1}^{p} a_i x[n-i]$ is the prediction error. The log-likelihood is given by

$$\ln p(\boldsymbol{x}_{N-p}\,|\,\boldsymbol{x}_p;\boldsymbol{a},\sigma^2) = -\frac{N-p}{2}\ln(2\pi\sigma^2) - \frac{\|\boldsymbol{x}_{N-p} - \boldsymbol{H}_{\boldsymbol{x}}\boldsymbol{a}\|^2}{2\sigma^2} \tag{2.24}$$

where the nth row of the matrix $\boldsymbol{H}_{\boldsymbol{x}} \in \mathbb{R}^{(N-p)\times p}$ is given by $\big(x[n-1] \;\; x[n-2] \;\; \cdots \;\; x[n-p]\big)$. Therefore, maximizing (2.24) with respect to $\boldsymbol{a}$ yields the *least-squares* solution of the following linear regression problem:

$$\boldsymbol{x}_{N-p} = \boldsymbol{H}_{\boldsymbol{x}}\boldsymbol{a} + \sigma\boldsymbol{w}, \tag{2.25}$$

where $\boldsymbol{w} \triangleq (w[p] \;\; \cdots \;\; w[N-1])^T$. Consequently, the conditional ML estimates of $\boldsymbol{a}$ and σ^2 follow from (2.24) and (2.25) as

$$\begin{aligned} \widehat{\boldsymbol{a}} &= \left(\boldsymbol{H}_{\boldsymbol{x}}^T\boldsymbol{H}_{\boldsymbol{x}}\right)^{-1}\boldsymbol{H}_{\boldsymbol{x}}^T\boldsymbol{x}_{N-p} \\ \widehat{\sigma^2} &= \frac{1}{N-p}\sum_{n=p}^{N-1}\left(x[n]x[n] - \sum_{i=1}^{p} a_i x[n]x[n-i]\right)^2 = \widehat{r}_{xx}[0] + \sum_{i=1}^{p} a_i \widehat{r}_{xx}[i], \end{aligned} \tag{2.26}$$

where $r_{xx}[\tau] \triangleq \mathbb{E}(x[n]x[n-\tau])$ is the autocorrelation function of $x[n]$. Estimating parameters of the AR process of (2.11) via the conditional ML estimator of (2.26) is often referred to as the *covariance method* of linear prediction in the speech processing community.

It is important to note that even though the ML formulation relied on the assumption that the innovations sequence was Gaussian, the resultant estimator is identical to the least-squares estimator derived starting from (2.10) directly, where no Gaussian assumption is explicit. As we discuss in Section 2.4.6 below, recognizing that this assumption is implicit in the usual treatment of the covariance method (see e.g., [3]) is valuable to understanding its shortcomings in the case of voiced speech, when the glottal flow is no longer turbulent, but is instead periodic.

2.4.2 Exact Maximum Likelihood Estimation

The conditional maximum likelihood estimator converges to the exact ML estimator as the number of observations N tends to infinity. But in the finite-sample-size setting, maximizing the *unconditional* joint density of $\boldsymbol{a}$ and σ^2 yields a different answer from that of the covariance method. To demonstrate this, we first explicitly write out the joint density of (2.23) as

$$p(\boldsymbol{x}\,;\,\boldsymbol{a},\sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}|\boldsymbol{R}|^{1/2}}\exp\left(-\frac{1}{2\sigma^2}\boldsymbol{x}^T\boldsymbol{R}^{-1}\boldsymbol{x}\right), \tag{2.27}$$

where $\boldsymbol{R}_{ij} = \mathbb{E}(x[n-i]x[n-j])\,/\sigma^2$ is the *filter* autocorrelation function—a nonlinear function of the AR coefficients $\boldsymbol{a}$, and $|\boldsymbol{M}|$ is the determinant of the matrix $\boldsymbol{M}$. Maximizing $p(\boldsymbol{x}\,;\,\boldsymbol{a},\sigma^2)$ with respect to σ^2 and substituting the resultant estimator back into (2.27) shows that the quantity

$$\left|\boldsymbol{R}^{-1}\right|^{1/N}\left(\frac{1}{N}\boldsymbol{x}^T\boldsymbol{R}^{-1}\boldsymbol{x}\right)^{-1} \tag{2.28}$$

must be maximized in order to obtain the exact ML estimator of $\boldsymbol{a}$ [32]. Unfortunately, the objective function in (2.28) is nonlinear in the parameters and its maximization requires computationally-demanding numerical procedures. In the case of an AR(1) process, for

instance, the exact maximum-likelihood estimator of $a[1]$ is a root of the following *cubic* polynomial:

$$a^3[1] + \frac{N-2}{N-1}\frac{S_{01}}{S_{11}}a^2[1] - \frac{S_{00}+NS_{11}}{(N-1)S_{11}}a[1] - \frac{NS_{01}}{(N-1)S_{11}} = 0,$$

whose coefficients are *nonlinear* functions of empirical cross-correlations with

$$S_{ij} \triangleq \sum_{n=0}^{N-1-i-j} x[n+i]x[n+j].$$

In fact, algorithms for exact ML estimation detailed in [32] and [33] entail solving a sequence of cubic and quartic equations in order to obtain the desired estimates.

One argument for ignoring $\left|\boldsymbol{R}^{-1}\right|^{1/N}$ in (2.28) and, instead, minimizing an approximation to $\frac{1}{N}\boldsymbol{x}^T\boldsymbol{R}^{-1}\boldsymbol{x}$ is that we can recover the conditional ML estimator for large N since

$$\lim_{N\to\infty}\left|\boldsymbol{R}^{-1}\right|^{1/N} = 1, \tag{2.29}$$

by properties of Toeplitz determinants [34].

Nevertheless, the computational expense may be worthwhile in certain cases. For example, consider an AR(1) process—the determinant of its inverse covariance matrix is $\left|\boldsymbol{R}^{-1}\right| = (1-a_1^2)^{1/N}$, which approaches 1 as $N \to \infty$. The closer a_1 is to unity, the slower the convergence of the sequence of determinantal powers in (2.29). As a result, we expect to see differences between the exact and conditional ML estimators when the sample size is small and the spectrum has sharp peaks. This is the setting in which the exact ML estimator is preferable to the conditional ML approach.

2.4.3 Method-of-Moments Estimation

Another popular estimator, termed the *autocorrelation* method of linear prediction, is derived using a method-of-moments argument as follows. Computing the second moment of $x[n]$ via (2.11) leads to:

$$r_{xx}[j] = \mathbb{E}(x[n]x[n-j]) = \begin{cases} \sum_{i=1}^{p} a_i r_{xx}[j-i] + \sigma^2 & \text{if} \quad j = 0 \\ \sum_{i=1}^{p} a_i r_{xx}[j-i] & \text{if} \quad 1 \le j \le p \end{cases}. \tag{2.30}$$

Rewriting (2.30) in matrix form results in the Normal or Yule-Walker equations:

$$\left(\begin{array}{c|ccc} r_{xx}[0] & r_{xx}[1] & \cdots & r_{xx}[p] \\ \hline r_{xx}[1] & r_{xx}[0] & \cdots & r_{xx}[p-1] \\ \vdots & \vdots & \ddots & \vdots \\ r_{xx}[p] & r_{xx}[p-1] & \cdots & r_{xx}[0] \end{array}\right)\begin{pmatrix} 1 \\ a_1 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} \sigma^2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{2.31}$$

In order to estimate the AR coefficients $\boldsymbol{a}$ and σ^2 we use a plug-in estimate of $r_{xx}[\tau]$ in (2.31) computed using sample averages via either

$$\widehat{r}_{xx}[\tau] = \frac{1}{N}\sum_{n=0}^{N-1} x[n]x[n-\tau] \quad \text{or} \quad \widehat{r}_{xx}[\tau] = \frac{1}{N-\tau}\sum_{n=\tau}^{N-1} x[n]x[n-\tau], \tag{2.32}$$

which correspond to the biased and unbiased estimators, respectively. Even though these estimators are asymptotically equal, the main difference in the finite-sample-size regime is that using the biased estimator guarantees that the estimated poles lie inside the unit circle [35].

It is well known that the Yule-Walker estimator has high variance and, consequently, a number of modifications to address this issue have been proposed. The least-squares modified Yule-Walker (LSMYWE) method [4], for instance, solves an overdetermined system of equations in an attempt to reduce estimator variance. Let $q \geq 0$ and $N \gg M - q$, then the LSMYWE estimator for the AR coefficients is obtained by solving:

$$\begin{pmatrix} r_{xx}[q] & r_{xx}[q-1] & \cdots & r_{xx}[q-p+1] \\ r_{xx}[q+1] & r_{xx}[q] & \cdots & r_{xx}[q-p+2] \\ \vdots & \vdots & \ddots & \vdots \\ r_{xx}[M-1] & r_{xx}[M-2] & \cdots & r_{xx}[M-p] \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} r_{xx}[q+1] \\ r_{xx}[q+2] \\ \vdots \\ r_{xx}[M] \end{pmatrix}.$$

The estimate of σ^2 can then be obtained via (2.30). The asymptotic properties of the LSMYWE estimator have been studied, and it was shown to be asymptotically unbiased with variance that decreases as $M - q$ increases [36]. It is also known that the performance of this estimator depends heavily on the pole locations, with the best performance typically achieved for sharply-peaked spectra.

2.4.4 Reflection Coefficient Estimation

The algorithms we have discussed so far estimate the predictor coefficients directly from the observation sequence. In order to obtain estimates of other parameter sets, the estimated AR coefficients need to be transformed using one of the methods described in Section 2.3.3. An important exception to this is Burg's method that, in essence, is used to estimate the *reflection* coefficients directly from observed data. Estimates of the AR coefficients can then be obtained via the Levinson recursion of (2.16).

To begin, observe that by substituting (2.16) into (2.14) and (2.15) and using the relation $a_{i,j} = b_{j-i,j}$, we obtain the following recursive relationship:

$$\begin{pmatrix} e_j^{\mathrm{f}}[n] \\ e_j^{\mathrm{b}}[n] \end{pmatrix} = \begin{pmatrix} 1 & \kappa_j \\ \kappa_j & 1 \end{pmatrix} \begin{pmatrix} e_{j-1}^{\mathrm{f}}[n] \\ e_{j-1}^{\mathrm{b}}[n-1] \end{pmatrix}. \tag{2.33}$$

Next suppose that the first $j-1$ reflection coefficients and consequently the prediction errors $e_{j-1}^{\mathrm{f}}[n]$ and $e_{j-1}^{\mathrm{b}}[n]$ are known. Then by using (2.33), we estimate κ_j to be the minimizer of the sum of squared forward and backward prediction errors given by:

$$\widehat{\kappa}_j^B \triangleq \operatorname*{argmin}_{\kappa_j} \sum_{n=j}^{N-1} \left(\|e_j^{\mathrm{f}}[n]\|^2 + \|e_j^{\mathrm{b}}[n]\|^2 \right). \tag{2.34}$$

It is easy to see that solving the least-squares problem of (2.34) for each $j = 1, \ldots, p$, starting with the natural, initial condition of $\hat{e}_0^{\mathrm{f}}[n] = \hat{e}_0^{\mathrm{b}}[n] = x[n]$, yields an estimator for the reflection coefficients. This approach in known as Burg's method [4] and is summarized in Algorithm 2.2.

Algorithm 2.2 Burg's Algorithm

- Initialize: For all $n = 0, \ldots, N-1$, set $\hat{e}_0^{\mathrm{f}}[n] = \hat{e}_0^{\mathrm{b}}[n] = x[n]$ and estimate $r_{xx}[0]$ via

$$\widehat{r}_{xx}[0] = \frac{1}{N}\sum_{n=0}^{N-1} |x[n]|^2.$$

- For $j = 1, \ldots, p$ compute

$$\widehat{\kappa}_j^B = \frac{-2\sum_{n=j}^{N-1} \hat{e}_{j-1}^{\mathrm{f}}[n]\hat{e}_{j-1}^{\mathrm{b}}[n]}{\sum_{n=j}^{N-1} |\hat{e}_{j-1}^{\mathrm{f}}[n]|^2 + |\hat{e}_{j-1}^{\mathrm{b}}[n]|^2}. \tag{2.35}$$

- Estimate the noise variance by:

$$\widehat{\sigma^2} = \widehat{r}_{xx}[0]\prod_{j=1}^{p}(1-\widehat{\kappa}_j^2).$$

- Return the estimated reflection coefficients and noise variance: $\{\widehat{\kappa}_1, \widehat{\kappa}_2, \ldots, \widehat{\kappa}_p, \sigma^2\}$.

Another estimator similar to that of (2.35) is the Itakura-Saito estimator given by [8]

$$\widehat{\kappa}_j^I = \frac{-\sum_{n=j}^{N-1} \hat{e}_{j-1}^{\mathrm{f}}[n]\hat{e}_{j-1}^{\mathrm{b}}[n]}{\sqrt{\sum_{n=j}^{N-1} |\hat{e}_{j-1}^{\mathrm{f}}[n]|^2}\sqrt{\sum_{n=j}^{N-1} |\hat{e}_{j-1}^{\mathrm{b}}[n]|^2}}. \tag{2.36}$$

Though (2.36) is not derived from minimizing a least-squares criterion, it has a pleasing interpretation as a method-of-moments estimator since from (2.17) we have that

$$\kappa_j \triangleq -\frac{\langle e_{j-1}^{\mathrm{f}}[n], e_{j-1}^{\mathrm{b}}[n-1]\rangle}{\|e_j^{\mathrm{f}}[n]\|^2} = -\frac{\langle e_{j-1}^{\mathrm{f}}[n], e_{j-1}^{\mathrm{b}}[n-1]\rangle}{\|e_j^{\mathrm{f}}[n]\|\|e_j^{\mathrm{b}}[n-1]\|}. \tag{2.37}$$

The Itakura-Saito estimator of (2.36) is then obtained by replacing the unknowns in (2.37) with their estimates obtained using sample averages.

Applying the Cauchy-Schwartz inequality to (2.36) shows that $|\widehat{\kappa}_j^I| \leq 1$ for all $1 \leq j \leq p$. Moreover, since the geometric mean of two numbers is smaller than their arithmetic mean, we see that $|\widehat{\kappa}_j^B| \leq |\widehat{\kappa}_j^I| \leq 1$. Consequently, poles estimated either by the Burg or the Itakura-Saito estimators lie inside or on the unit circle. Many other variations of these algorithms exist and there is a strong connection between them and the literature on lattice and ladder filters (see, e.g., [37] for an excellent overview), but their differences are too minor to be significant in practice.

2.4.5 Comparing AR Estimation Methods

The conditional maximum likelihood (covariance) and method-of-moments (autocorrelation) estimators converge to the exact maximum likelihood estimator as the number

of observations N tends to infinity. In addition, a careful argument, based on the Levinson recursion, has been used to show that many reflection coefficient estimators (including the Burg and Itakura-Saito estimators of Section 2.4.4) also converge to the exact maximum likelihood estimate [38]. Therefore, all these estimators share the properties of the maximum-likelihood estimator—they are asymptotically consistent and efficient. In particular, the estimates converge to:

$$\begin{pmatrix} \widehat{\boldsymbol{a}_{\mathrm{ML}}} \\ \widehat{\sigma^2}_{\mathrm{ML}} \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{a} \\ \sigma^2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{R}_{xx}/N & \mathbf{0}_{p\times 1} \\ \mathbf{0}_{1\times p} & 2\sigma^4/N \end{bmatrix} \right). \tag{2.38}$$

The main differences among the various AR estimators arise in the finite-sample-size setting, but there is no clear winner and a number of issues needs to be considered. Application of the autocorrelation method (using biased estimates of r_{xx}) or the reflection coefficient estimators results in minimum-phase spectral estimates (i.e., estimated poles lie inside or on the unit circle); this property isn't guaranteed for either of the maximum-likelihood methods, but is an important consideration in a number of applications such as speech coding and synthesis. On the other hand, maximum likelihood methods have better spectral resolution than method-of-moments estimators [4]. When the number of observations is not much larger than the model order (and especially if some of the poles are close to the unit circle resulting in a peaked spectrum), then the exact maximum likelihood estimator is preferable.

Finally, we point out that, with the exception of the exact maximum likelihood estimator, all the other methods yield *linear* estimators of the predictor or reflection coefficients, respectively. Unfortunately, these are the only parameter sets that may be linearly estimated; their estimates are appropriately transformed if other representations (e.g., those discussed in Section 2.3.3) are required.

2.4.6 Extension to Voiced Speech

All the estimators presented are derived under the assumption that the innovations sequence in (2.11) is a stationary Gaussian process. During voiced speech, however, this assumption is inappropriate since the glottal airflow is periodic. But the covariance and autocorrelation methods are still widely applied in this case, albeit with ad hoc modifications designed to lessen the effect of the resultant estimator bias.

A simple and common way of addressing this implicit model mismatch is based on the observation that two maximum-phase (i.e., outside the unit circle) poles provide a rough model of the glottal pulse shape [3]. Indeed, increasing the order of the estimated AR model by at least two greatly improves the model fit, but at the expense of blurring the distinction among the parameters that describe the transfer function of the vocal tract and those that model the glottal flow waveform. This method reduces the estimator bias for the poles used to model the vocal tract, but does not eliminate it.

In order to illustrate this explicitly, we synthesized the vowel [ɪ] (as in beet) using a formant synthesizer in which the vocal tract is represented by a cascade of three second-order digital resonators, corresponding to an AR(6) model. The filter is excited by a deterministic voicing component $g[n]$ generated using the Rosenberg-Klatt model [13]. We

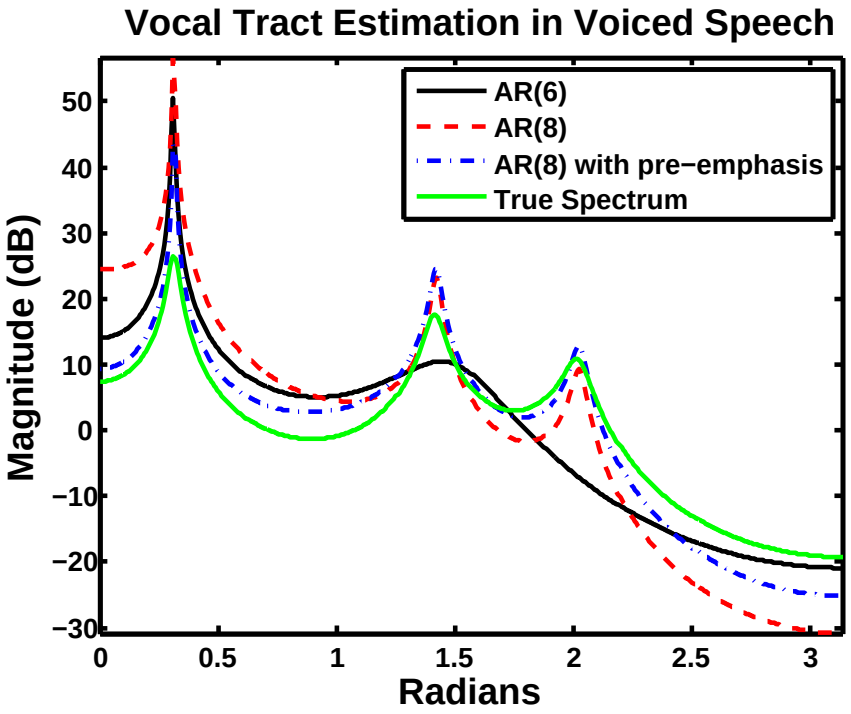


Figure 2.6: Comparison of AR spectral estimation methods on 1 second of the synthetic vowel [ɪ] generated at an 8 kHz sampling rate. Log-spectra for the true AR (solid, green) and fitted AR(6) (solid, black) and AR(8) models. The latter is fit both with (dashed-dot, blue) and without (dashed, red) pre-emphasis.

fit the waveform data with AR models of different order using the covariance (conditional ML) method of linear prediction; the estimated all-pole (log) spectra are shown in Figure 2.6. Fitting an AR(6) model and ignoring the non-Gaussian nature of the input leads to severe bias as evidenced by the locations of the estimated spectral peaks. One of the peaks is missed altogether! Adding two poles improves the overall fit, as does adding a pre-emphasis step (a standard step used in order to remove the spectral tilt resulting from the glottal pulse shape). But the estimators are clearly biased even in this, essentially asymptotic, regime with $N = 8000$ and $p = 8$.

Another approach to mitigating these problems is to derive estimators based on a more robust error criterion than least-squares [39]. This is equivalent to assuming a heavy-tailed innovations sequence in (2.11) and may help to model the impulsive nature of the excitation during glottal closures more accurately, but does not result a physiologically-faithful model of the glottal flow since its periodicity is disregarded. This technique and related variations are summarized and compared in [40].

An entirely different class of methods for estimating filter parameters during voiced speech is based on jointly modeling the glottal flow and vocal tract and estimating their parameters simultaneously. The approach is to assume that a speech signal $x[n]$ can be modeled as a linear-time invariant system driven by a Gaussian process with a *time-varying*

mean $\mu[n]$ according to

$$x[n] = \sum_{i=1}^{p} a_i x[i] + \mu[n] + \sigma w[n].$$

The time-varying mean is modeled as a convolution of a periodic pulse train with one of the glottal templates previously described in Section 2.2. The vocal tract parameters are then estimated jointly with the glottal pulse parameters; examples include the Rosenberg-Klatt [18] and LF [14] models in [41] and [42, 43], respectively. The parameters of the resultant models are nonlinearly related to the observed data, which leads to inferential procedures based on iterative estimation methods that have no guarantee of convergence.

Nonparametric models of the glottal flow pulse have also been proposed [20, 21], but suffer from some of the same inferential difficulties as the parametric approaches. In Chapter 7, we present a novel approach to this problem through nonparametric modeling of the glottal flow waveform over multiple pitch periods—a marked departure from the template-based methods—which leads to efficient linear estimators for the resultant semi-parametric model.

2.4.7 Model Order Selection in Linear Prediction

In our discussion of AR estimation methods thus far, we assumed that the number of poles p is known and fixed. Of course, when applying linear prediction to speech data, we need to know how p should be set. There are two ways to approach this model order selection problem.

The first approach is more general because it does not take into account any domain-level knowledge (of, e.g., speech production) and primarily relies on asymptotic statistical arguments in order to strike a balance between goodness of fit to the observed data and parsimony. Two of the most popular approaches are based on the Akaike information criterion (AIC) [44] and the minimum description length (MDL) principle [45]. Both methods attempt to estimate the model order through minimizing a cost function of the data and the number of fitted parameters. The cost functions are given by

$$\begin{aligned} \text{AIC:} \quad & N \log\left(\widehat{\sigma^2}\right) + 2p \\ \text{MDL:} \quad & N \log\left(\widehat{\sigma^2}\right) + p \log N, \end{aligned}$$

where N is the number of observations, p is the number of parameters, and $\widehat{\sigma^2}$ is the maximum likelihood estimate of the innovations noise variance. In both cases, the magnitude of the second term grows and that of the first term decreases with increasing p, since increasing model order always improves goodness of fit. Thus, in most cases, a minimum should exist. In fact, it has been shown that MDL (but not AIC) is a consistent estimator of the model order for AR processes. In other words as N tends to infinity the estimated model order $\widehat{p}$ converges to the true model order p with probability one. Many other similar criteria exist including the Bayesian information criterion [46], the Deviance information criterion [47], and AR-specific criteria described in [4].

The second approach to model order selection is based on knowledge about the speech production mechanism and does not solely rely on parsimony. The well-known rule

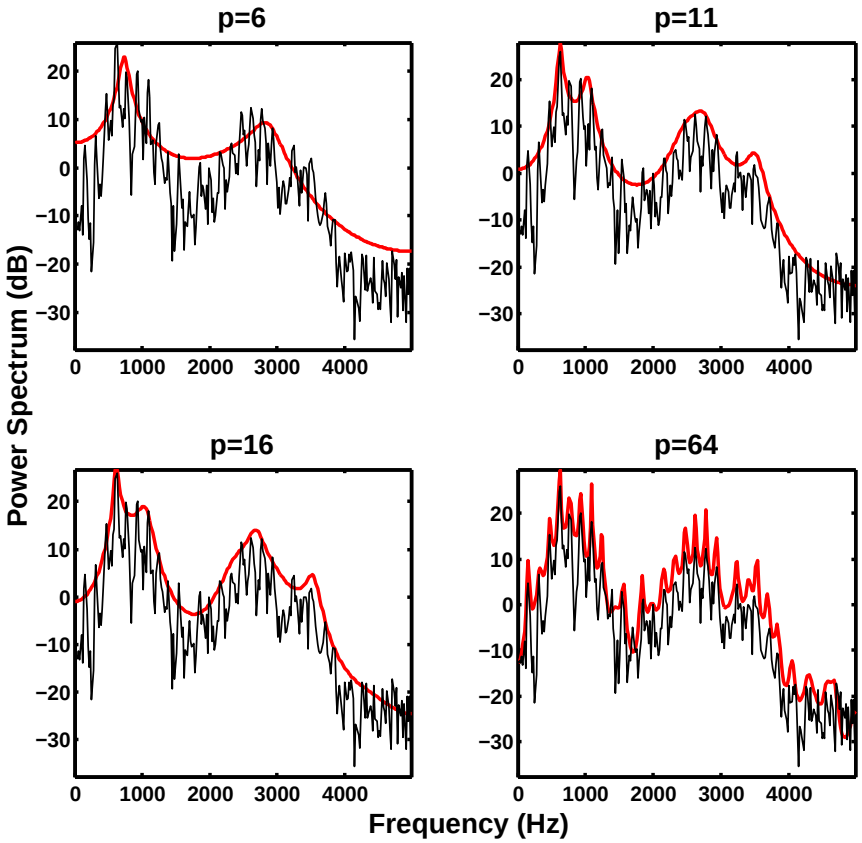


Figure 2.7: Linear-predictive analysis of the vowel [o] using the covariance method with different model orders. The estimated AR spectra are plotted (thick, red) lines for $p = 6$ (top left), $p = 16$ (top right), $p = 32$ (bottom left), and $p = 64$ (bottom right). The periodogram estimate of the power spectrum is also shown (thin, black) for reference.

of thumb, which depends on the sampling rate, is that there is one vocal tract resonance per 1000 Hz on average. Thus, supposing a bandwidth of 5000 Hz, 10 poles are required to represent the vocal tract. The radiation load at the lips is typically modeled by a zero, but it has been found that an all-zero spectrum could be sufficiently characterized by about 4 poles [29]. Two additional poles would serve to account for the glottal pulse shape as discussed in Section 2.4.6, suggesting a total of 16 poles for modeling the vocal tract.

Figure 2.7 shows an example of the covariance method of linear prediction applied to a vowel [o] (as in 'boy', recorded at 10 kHz) for different model orders including $p = 11$ (as determined by AIC and MDL methods) and $p = 16$ (as suggested by the speech-specific) approach. A nonparametric spectral estimate shown in all the panels reveals harmonic structure due to the periodicity of the glottal flow.

Clearly, when $p = 64$, the fitted all-pole spectrum begins to estimate the harmonic structure and models spectral contributions of the glottal flow together with those of the vocal tract. On the other hand, the spectral estimate obtained using for $p = 6$ is over-smoothed and misses the second and fourth formants that are clearly visible if $p = 11$ or

$p = 16$ were used instead.

In practice, AR-synthesized speech sounds "musical" with randomly appearing and disappearing tones if p estimated to be too large (estimated spectrum exhibits harmonic structure), but sounds "muffled" if p is underestimated (estimated spectrum is oversmoothed). Outside the extremes of grossly overestimating or underestimating p, there is no accepted way of deciding whether to use the MDL-based or rule-of-thumb estimates of the model order. If common practice is any guide, most applications are based on the simpler data-independent rule-of-thumb approach. To an engineer, it has the added benefit of being interpreted in the context of speech production rather than being rooted in arguments about asymptotic estimator behavior that are preferred in statistics.

2.5 Temporal Variation in Speech

All the methods for estimating vocal tract parameters from a speech waveform, described in Section 2.4, were derived under the assumption that the vocal tract is *not changing* over the duration of the entire signal. This is transparent in the all-pole model of the vocal tract transfer function—the time invariance of its coefficients in (2.2) and the induced autoregressive model of (2.10) imply that the vocal tract resonances are fixed. Even though it is justifiable for sustained sounds such as steady vowels, this assumption is clearly inconsistent with the time-varying nature of natural speech: words comprise sequences of phonemes, each associated with a unique vocal tract configuration.

How, then, do we estimate the time-varying configuration of a vocal tract from a single utterance? First, observe that the vocal tract cannot change instantaneously because any temporal variation results from the physical movement of articulators, which takes time for our muscles to effect. Indeed, at very short time scales there will is little-to-no articulator movement, and therefore the vocal tract transfer function may be reasonably approximated by an all-pole model with time-invariant coefficients, which can then be estimated by any of the methods described in Section 2.4. Thus, the remaining issue is to determine at which time scales is the time-invariance of the vocal tract observable given the sampling rate of the acoustic recording[1]. Our discussion of speech production suggests two answers: vocal tract variation at the *segmental* (tens of milliseconds) and *subsegmental* (less than one pitch period) levels.

2.5.1 Segmental Vocal Tract Variation

Time invariance of the vocal tract at the segmental level depends on the phoneme uttered and its duration. It is widely accepted that over different phonemes and speakers, the vocal tract is time invariant on the order of 15 to 30 ms at the segmental level [3,7,8,29]. This, in turn, motivates a short-time speech analysis procedure we describe below; it is used in nearly every downstream application ranging from speech recognition and coding to time-scale modification, enhancement, and many others [3].

[1]Sampling rates of most speech recordings typically range from 5 kHz to 44.1 kHz, implying anywhere between 50 and 5 samples per millisecond of speech.

The speech waveform is tiled using overlapping translates of a smooth, time-localized 15–30 ms window. Multiplying the waveform by each of these windows induces a sequence of short-time segments within which the vocal tract is assumed to be time invariant. Next, each short-time segment can be used to infer the vocal tract parameters (e.g., using the covariance method of linear prediction) or other features of interest may be obtained. The extracted parameters may then be employed in a downstream application. If necessary the short-time segments (either modified or intact) can be recombined using the so-called overlap-add procedure (see e.g., [3] and Chapter 6 in this thesis) to reconstitute a speech waveform.

An example is shown in Figure 2.8 using the waveform "Reading in poor light gives you eye strain," taken from the TIMIT speech corpus [48]; it was recorded by a female speaker at 16 kHz. The waveform is tiled using a sequence of 20 ms Hamming windows overlapping by 50 % with their neighbors—a cartoon is shown in Figure 2.8(b). The spectral content of each of the resulting short-time segments can then be analyzed. The estimated power spectrum of one such short-time segment is shown in Figure 2.8(c) along with an estimate of the spectral envelope obtained using the conditional maximum likelihood method of Section 2.4.1. If one were to arrange the power-spectral estimates derived from the short-time segments in a two-dimensional display, then a time-frequency representation called a spectrogram, shown in Figure 2.8(d), is obtained. The vocal tract resonances are indicated by higher intensities in the spectrogram. Formant positions can also be derived from an all-pole envelope such as the one shown in Figure 2.8(c). The locations of the first three vocal tract resonances derived in this manner using the Wavesurfer toolbox [49] are superimposed on the spectrogram in Figure 2.8(d), resulting in so-called formant tracks—functions describing the temporal evolution of each vocal tract resonance.

The assumption that the vocal tract is time invariant in short-time segments whose length is 15–30 ms is reasonable, on average. But, in fact, the duration of many phonemes can significantly differ from this average. For instance, plosives last on the order of 5 ms, whereas while certain vowels may last as long as 100 ms. Consequently, the length of the analysis can have a profound effect on subsequent speech analysis as shown in Figure 2.9, displaying 5, 20 and 80 ms spectrograms (Hamming windows, 50% overlap) of the TIMIT utterance shown earlier in Figure 2.8. Clearly, the plosives "d", "g", and "t" are best resolved using short 5 ms windows; their spectrum is smeared when longer windows are employed. On the other hand, using longer windows yields high spectral resolution and the harmonic structure in voiced parts of the words "poor light" is transparent. Spectral resolution decreases with shorter windows, as expected by the Fourier uncertainty principle, and when 5 ms windows are used the harmonic lines can no longer be distinguished. This motivates the use of variable-resolution spectral analysis—we discuss such methods in Chapter 6 of this thesis.

2.5.2 Subsegmental and Suprasegmental Vocal Tract Variation

Short-time speech analysis at the segmental scale is based on the assumption that the vocal tract is time invariant on the scale of 15–30 ms. In voiced speech, however, the vocal tract is also time invariant at the subsegmental scale (within a single pitch period) during the closed phase of the glottal airflow. This invariance is truly time-localized because

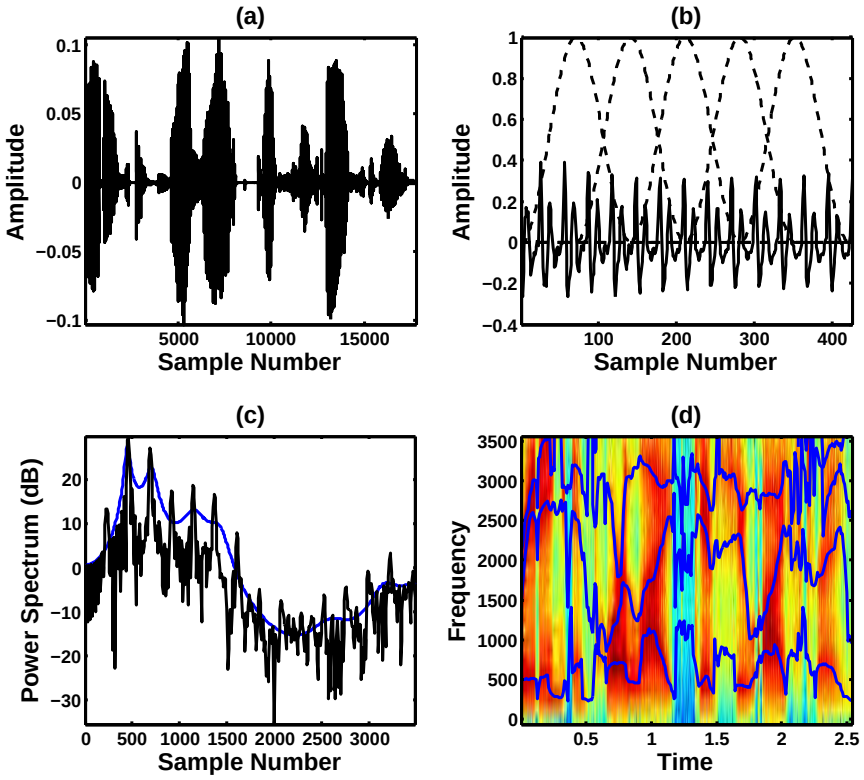


Figure 2.8: Example of using short-time analysis to infer the time-varying vocal tract configuration from acoustic data: (a) Time-domain representation of the waveform "Reading in poor light gives you eye strain." (b) Cartoon showing how windowing induces short-time segments. (c) Estimate of the power spectrum in one short-time segment obtained using a discrete Fourier transform (black) overlaid with an estimate obtained using all-pole modeling (blue) (d) A spectrogram comprising a sequence of power spectra (one per short-time segment), estimated using the discrete Fourier transform, is overlaid with tracks of the first three formants (blue) derived from the all-pole model using Wavesurfer [49].

the closed phase is short—its length may range anywhere from 2 to 4 ms [2]. Moreover, the vocal tract configuration changes at the end of the closed phase as the vocal folds start to separate at the glottal opening instant, the effective length of the vocal tract gradually increases, resulting in a change in the frequency and bandwidth of the first formant [50].

The time-invariance of the vocal tract during the closed phase and its subsequent motion during the open phase is demonstrated in the top two panels of Figure 2.10. The closed phase of a single pitch period of the vowel [ɑ] (as in "father") is shown in Figure 2.10(a), and is demarcated on its left by a solid black line and on its right by a dashed black line, corresponding to the GCI and GOI, respectively. Figure 2.10(b) shows the

[2]The average pitch frequency of human voices varies from 100 Hz (males) to 250 − 300 Hz (women and children) and the length of the closed phase is one quarter to one half of the length of a single period.

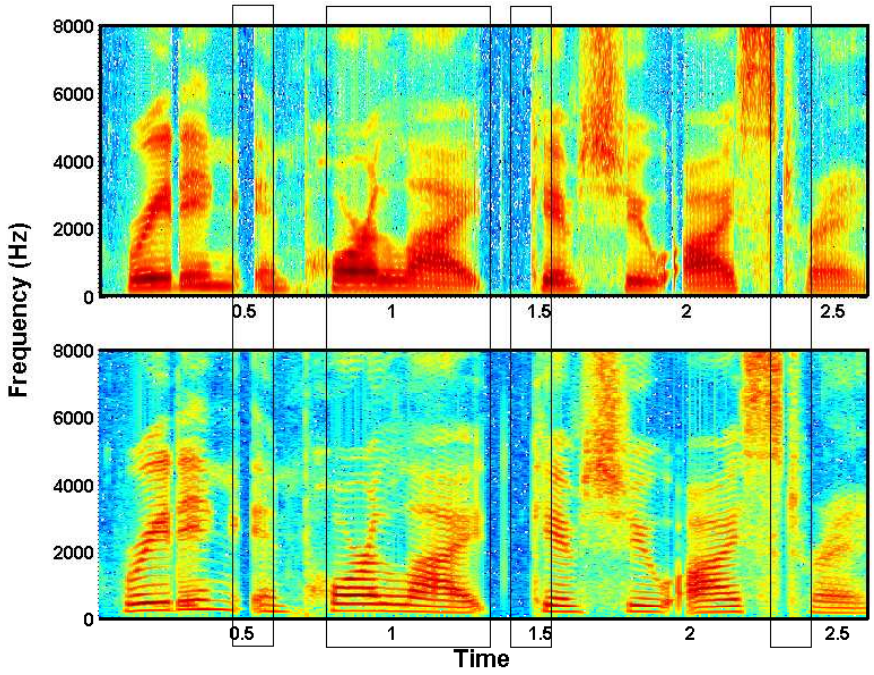


Figure 2.9: Effect of analysis window length on short-time spectral analysis of speech. Spectrograms of the utterance "Rea*d*ing in p*oor light* *g*ives you eye s*t*rain," associated to 5, 20 and 80 ms Hamming windows are shown in the top, middle, and bottom panels, respectively. The italicized parts of sentence are highlighted in the spectrogram using boxes and emphasize differences among the estimated spectra.

estimated formant trajectories for each position of a short, rectangular window initially left-aligned with the GCI and sliding one sample to the right through the GOI. As expected, the estimated AR coefficients are not changing during the closed phase, but go through a marked transformation around the glottal opening due to both a change in the frequency/bandwidth of the first formant and a gradual increase in the glottal volume velocity.

Thus far we have seen that it is meaningful to discuss the temporal dynamics of the vocal tract at multiple time scales, as shown in Figure 2.10. We can observe vocal tract variation due to the oscillation of the vocal folds and at the segmental scale due to the motion of the articulators. One can even study vocal tract variation at the supra-segmental scale (hundreds of milliseconds) arising from the prosody of speech comprising effects such as rhythm, stress, emotion, intonation, and other long-term linguistic effects. One example, adapted from [3] is shown in Figure 2.10(d) and (e) where clear differences in the formant structure of the word "today" arise from an increase in loudness and stress on the last syllable.

From the source-filter point of view, temporal variation in speech can arise not only from vocal tract (filter) variations at multiple scales, but also from time-varying characteristics of the glottal flow (source). At the subsegmental scale glottal pulses may be

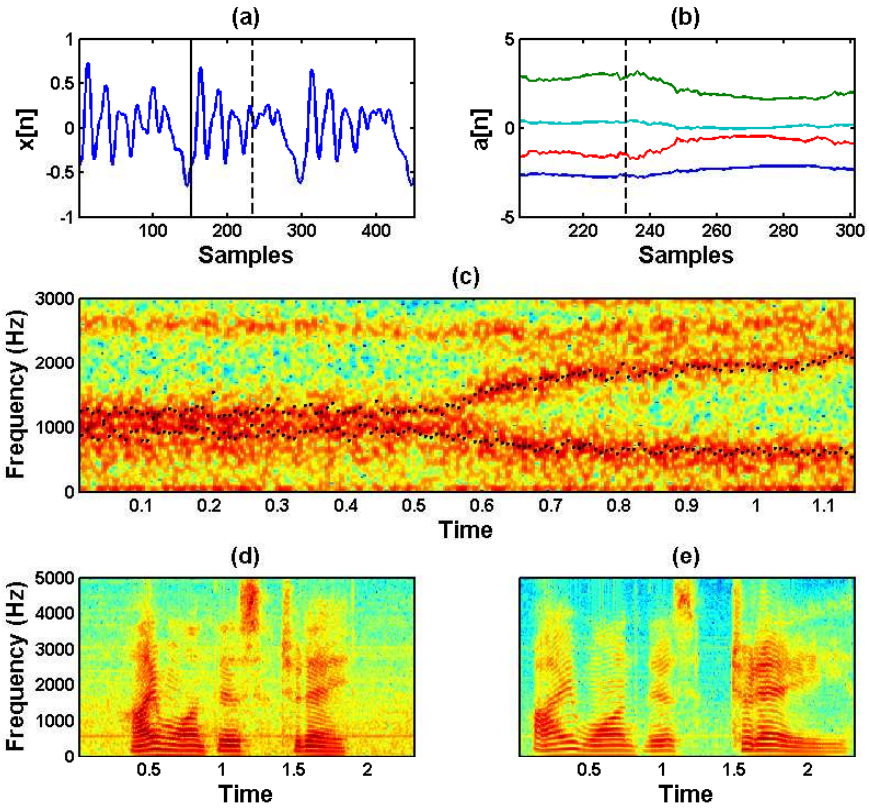


Figure 2.10: Variation of the vocal tract on multiple time-scales. (a) Subsegmental vocal tract variation within one pitch period of the vowel [ɑ] is seen through (b) the trajectories of AR coefficients computed for every position of a sliding window are overlaid with estimates of the GCI (solid black line) and GOI (dashed black line) (c) Segmental vocal tract variation during the whispered waveform [ɑaɪ] is seen through the time-varying positions of the first two formants, derived through the all-pole model using Wavesurfer [49] and overlaid on a spectrogram computed using 16 ms Hamming windows with 50% overlap. Suprasegmental vocal tract variation is seen in the differences between the spectral content of the word "today" spoken normally (d) and with stress on the last syllable (e).

irregularly spaced with their shapes and amplitudes varying from one pulse to the next—such *non-modal* phonation occurs in normal speakers resulting in effects called "creak, " "diplophonia", and "vocal fry" among others [3]. Pitch variation also exists on the segmental and suprasegmental scales. In tonal languages—a group which includes all Chinese and most sub-Saharan African languages—pitch inflections are used to ascribe different meanings to phonemes (sometimes appropriately called tonemes). At the supra-segmental scale pitch is used to convey emotion and to place emphasis. Increasing pitch at the end of a question is a classic example.

2.6 Nonstationary Signal Modeling

In Section 2.5 above, we have described temporal speech variation in terms of the physiology of the speech production system and the evolution of specific model parameters such as formants or autoregressive coefficients. However, we can make these concepts much more precise in the framework of stochastic processes. At a high level, stationary and non-stationary stochastic processes can be used to model time invariant and time-varying aspects of the speech waveform, respectively. Consequently, in this section we first review important properties of discrete-time stationary stochastic process and then discuss classes of nonstationary processes which we will use throughout this thesis.

2.6.1 Stationary Processes

A discrete-time stochastic process $\{x[n] \,|\, n \in \mathbb{Z}\}$ is called *strict-sense stationary* (SSS) if its statistics are invariant to time shifts: The processes $x[n]$ and $x[n+m]$ have the same statistics for any $m \in \mathbb{Z}$. A useful relaxation of strict-sense stationarity is an invariance condition only on the first- and second-order statistics of a stochastic process. A discrete-time stochastic process $\{x[n] \,|\, n \in \mathbb{Z}\}$ is called *wide-sense stationary* (WSS) if it has a constant mean:

$$\mathbb{E}(x[n]) = m_x, \tag{2.39}$$

and if its autocorrelation function $\mathbb{E}(x[n]x[m])$ depends only on the lag $n-m$ according to:

$$r[n,m] \triangleq \mathbb{E}(x[n]x[m]) = \mathbb{E}(x[n-m]x[0])\,. \tag{2.40}$$

Accordingly, we denote $r[n,m]$ by $r[n-m]$. Note that the condition of (2.39) may be relaxed, thereby allowing for a time-varying mean. Consider a zero-mean WSS stationary process $x[n]$ and construct the process $\widetilde{x}[n] = x[n] + \eta[n]$ where $\eta[n]$ is an arbitrary deterministic function. The resultant process is nearly wide-sense stationary (NWSS)—its mean and autocovariance function depend on time, but its autocorrelation function does not [51]. We will encounter NWSS processes in our modeling of the source waveform in Chapter 7.

The definition of a WSS stationary process implies that its average power $r[0] \triangleq \mathbb{E}(x[n]x[n])$ is time invariant. It is often revealing to study the distribution of the average power over a discrete set of frequencies. This *power-spectral density* or *power-spectrum* is defined to be the discrete Fourier transform of the autocorrelation function

$$S(\omega) \triangleq \sum_{m=-\infty}^{\infty} r[m]e^{-im\omega}. \tag{2.41}$$

It is easy to check whether an autoregressive processes is wide-sense stationary. If $x[n]$ is a discrete-time AR process satisfying (2.10), then it is wide-sense stationary *if and only if* the roots of the associated predictor polynomial $A(z) \triangleq 1 - \sum_{i=1}^{p} a_i z^{-i}$ lie strictly inside the unit circle in the complex plane [4].

Every WSS process $x[n]$ can be written as an infinite moving average of an uncorrelated innovations sequence $e[n]$ according to:

$$x[n] = \sum_{m=-\infty}^{n} h[n-m]e[m] + \eta[n] = \sum_{m=0}^{\infty} h[n]e[n-m] + \eta[n], \tag{2.42}$$

where $h[n]$ is the impulse-response of a minimum-phase moving-average filter, and $\eta[n]$ is an arbitrary deterministic trend. If $\eta[n] = 0$ for all $n \in \mathbb{Z}$, then $x[n]$ is called a *purely nondeterministic* (PND) process. The MA(∞) representation of (2.42) is known as the Wold representation.

The spectral representation is another very useful tool for studying stationary processes. A discrete-time WSS process $x[n]$ may be written as:

$$x[n] = \int_{-\pi}^{\pi} e^{i\omega n} Z(\omega) d\omega, \tag{2.43}$$

where the stochastic integral is interpreted as a mean-square limit [52], and $Z(\omega)d\omega$ is an orthogonal increments process, which means that if $S(\omega)$ is the power spectral density of $x[n]$, then:

$$\mathbb{E}(Z(\omega)Z(\nu)) = \delta(\omega - \nu)S(\omega). \tag{2.44}$$

It may be shown that a process $x[n]$ is WSS *if and only if* it admits the representation of (2.43). A rigorous treatment of the spectral representation may be found in [52].

2.6.2 Controlled Departure from Stationarity

It is not constructive to talk about the class of nonstationary processes because nonstationarity is a non-property—it describes any stochastic process that is not wide-sense stationary. A more practical and useful approach is to relax the definition of stationarity in a controlled way so that the set of resulting nonstationary processes can be quantitatively described by how much its members depart from their stationary counterparts. In this vein, the search for representations of nonstationary processes that are amenable to theoretical analysis and useful for modeling real-world signals has led to extensive literatures in statistics, signal processing, and probability. In this section, we discuss a number of these ideas. While not exhaustive, our presentation aims to survey the most relevant results for this thesis.

We first describe an approach that depends on relaxing the spectral representation of (2.43), but captures only a small class of non-stationary processes. A method based on relaxing the Wold representation of (2.42) that captures a much broader set of processes is described next. We show that it, in turn, inspires a class of parametric time-varying autoregressive moving average models.

2.6.2.1 Priestley's Evolutionary Spectrum

Since a process $x[n]$ is wide-sense stationary if and only if it admits the spectral representation of (2.43), the relaxation of this characterization provides a natural approach to defining a class of nonstationary processes. Priestley [53] proposed to relax (2.43) as follows:

$$x[n] = \int_{-\pi}^{\pi} e^{i\omega n} Z(n, \omega) d\omega, \tag{2.45}$$

where $Z(n, w)$ are time-varying, orthogonal increments:

$$\mathbb{E}(Z(n, \omega)Z(n, \nu)) = \mathrm{ES}(n, \omega)\delta(\omega - \nu), \tag{2.46}$$

and where $\mathrm{ES}(n,\omega)$ denotes the value of the *evolutionary spectrum* at time instant n and frequency ω.

Priestley studied the properties of the evolutionary spectrum for the class of so-called oscillatory processes defined, implicitly, by the assumption that each orthogonal increment in (2.46) is of the form:

$$Z(n,\omega) = A(n,\omega)N(\omega), \tag{2.47}$$

where $N(\omega)$ is stationary white noise with normalized average intensity and $A(n,\omega)$ is a deterministic, complex-valued modulation function. Substituting (2.47) into (2.45) results in the following spectral representation for oscillatory processes:

$$x[n] = \int_{-\pi}^{\pi} e^{i\omega n} A(n,\omega)N(\omega)d\omega. \tag{2.48}$$

Combining (2.46) with (2.47) and (2.48), it is easy to see that the evolutionary spectrum of an oscillatory process is defined entirely through the modulation function $A(n,\omega)$ according to:

$$\mathrm{ES}(n,\omega) = |A(n,\omega)|^2, \tag{2.49}$$

which agrees with our intuition since oscillatory processes are the result of time-varying amplitude modulation of each frequency in $(-\pi,\pi)$. If $A(n,\omega)$ were *slowly* time-varying then $A(n,\omega)e^{i\omega n}$ may be viewed as a narrowband signal (as a function of n) localized around ω. The slower the time-variation in the modulation function $A(n,w)$, the closer we are to the stationary case, and the more appropriate it is to interpret $\mathrm{ES}(n,\omega)$ as the signal power at a frequency ω at time instant n. Moreover, one can think of the evolutionary spectrum as a spectral distribution of time-varying average power—in analogy to interpreting the power spectrum of a stationary process as a distribution of average power—since (2.48) implies that:

$$\mathbb{E}(x[n]x[n]) = \int_{-\pi}^{\pi} \mathrm{ES}(n,\omega)d\omega.$$

Another pleasing interpretation of the evolutionary spectrum can be obtained if we assume that $A(n,\omega)$ has an inverse Fourier transform $h[n,u]$:

$$A(n,\omega) = \sum_{u=-\infty}^{\infty} e^{-i\omega u} h[n,u],$$

where the spectral representation of the stationary white noise sequence $e[n]$ is given by:

$$e[n] = \int_{-\pi}^{\pi} e^{i\omega n} N(\omega)d\omega. \tag{2.50}$$

Then, we may write $x[n]$ as an output of a time-varying linear filter with impulse response $h[n,u]$ that is excited by the input $e[n]$ since:

$$\begin{aligned} x[n] &= \int_{-\pi}^{\pi} e^{i\omega n} A(n,\omega)N(\omega)d\omega = \int_{-\pi}^{\pi} \sum_{u=-\infty}^{\infty} e^{-i\omega u} h[n,u] e^{i\omega n} N(\omega)d\omega \\ &= \sum_{u=-\infty}^{\infty} h[n,u] \left(\int_{-\pi}^{\pi} e^{i\omega(n-u)} N(\omega)d\omega \right) = \sum_{u=-\infty}^{\infty} h[n,u]e[n-u]. \end{aligned}$$

Despite some of the nice properties we have discussed, Priestley's evolutionary spectrum has a number of weaknesses. Even though oscillatory processes may be written as outputs of a time-varying linear filter, the converse is not true—many processes that are naturally represented as outputs of a time-varying linear filter excited by a stationary input (e.g., ARMA) are not oscillatory. It is also difficult to verify whether or not a given stochastic process is oscillatory.

2.6.2.2 Tjøstheim's Evolutionary Spectrum

Inspired by the time-varying filtering interpretation of Priestley's evolutionary spectrum, but aiming to overcome some of its limitations, Tjøstheim [54] proceeded by relaxing the Wold representation of (2.42), rather than the spectral representation of (2.43), according to:

$$x[n] = \sum_{m=-\infty}^{n} h[n,m]e[m], \tag{2.51}$$

where the process is represented as the output of a *time-varying* moving-average filter with impulse response $h[n,m]$. The representation of (2.51) was first developed by Cramer [55].

Substituting the spectral representation of $e[m]$ given by (2.50) into (2.42) yields:

$$x[n] = \int_{-\pi}^{\pi} \sum_{m=-\infty}^{n} h[n,m]e^{i\omega m}N(\omega)d\omega. \tag{2.52}$$

Using (2.52), Tjøstheim defined an evolutionary power spectrum $\Theta[n,\omega]$ according to:

$$\Theta[n,\omega] \triangleq \left| \sum_{m=-\infty}^{n} h[n,m]e^{i\omega m} \right|^2, \tag{2.53}$$

which is everywhere nonnegative and reduces to the usual definition of the power spectrum in the stationary case when $h[n,m]$ is a function only of the lag $n-m$. Clearly, the time-varying MA(∞) representation of (2.51) implies that Tjøstheim's evolutionary spectrum generalizes the approach of Priestley to a broader class of nonstationary processes.

Tjøstheim's spectrum also has a very nice frequency-domain interpretation. A change of variables shows that:

$$\left| \sum_{m=-\infty}^{n} h[n,m]e^{i\omega m} \right|^2 = \left| \sum_{m=0}^{\infty} h[n,n-m]e^{-i\omega m} \right|^2, \tag{2.54}$$

and we may recognize the term on the right side of (2.54) as Zadeh's generalized transfer function [56]

$$H(n,z) = \sum_{m=0}^{\infty} h[n,n-m]z^{-m}, \tag{2.55}$$

evaluated on the unit circle at $z = e^{i\omega}$, which captures the response of a time-varying filter to complex exponentials. Consequently, the Zadeh's and Tjøstheim's generalizations of the power spectrum to the time-varying setting coincide, since $\Theta[n,\omega] = |H(n,\omega)|$.

Even though Tjøstheim's approach circumvents some of the problems of Priestley's evolutionary spectrum, a major practical obstacle remains. Indeed, an infinite number of parameters may be required to parameterize a stochastic process obeying (2.51). In practice, this translates into the need of estimating a very large number of parameters.

2.6.2.3 Parametric Modeling

A more practical and flexible approach was put forward by Grenier [57], who proposed to realize (or approximate) the time-varying MA(∞) representation of (2.51) using an ARMA(p, q) model with time-varying coefficients according to:

$$x[n] = \sum_{i=1}^{p} a_i[n-i]x[n-i] + \sum_{i=0}^{q} b_i[n-i]e[n-i], \tag{2.56}$$

where $e[n]$ is a stationary white sequence for all $n \in \mathbb{Z}$. Grenier [58] shows that a nonstationary system with time-varying impulse response $h[n, m]$ admits a representation of (2.56) if and only if there exist some $p, q \in \mathbb{Z}^+$ and functions $\{a_i[n] \,|\, 1 \leq i \leq p\}$ so that:

$$h[n, n-m] + \sum_{i=1}^{p} a_i[n-i]h[n-i, n-m] = 0 \qquad \text{for} \quad m > q. \tag{2.57}$$

For processes that satisfy (2.51) and (2.57), the representation of (2.56) is exact, otherwise it provides a useful but approximate realization.

Notice that the time indices of the AR and MA coefficients in (2.56) are synchronous with the associated random variables—we say the model is written in *synchronous* form. An alternative is the *shifted* form given by:

$$x[n] = \sum_{i=1}^{p} a_i[n]x[n-i] + \sum_{i=0}^{q} b_i[n]e[n-i]. \tag{2.58}$$

The synchronous and shifted forms of (2.56) and (2.58) are related by a simple transformation, but, as we will see in Chapter (4), they lead to substantially different algebraic structure in associated parameter estimators. Both models can be used to obtain a doubly-infinite sequence of *tangential* or *frozen-in-time* stochastic processes using the coefficients $\{a_1[n], a_2[n], \ldots, a_p[n]\}$ and $\{b_0[n], b_1[n], \ldots, b_q[n]\}$ at each time instant $n \in \mathbb{Z}$ in order to define a frozen-in-time *rational* spectrum that, in the case of (2.58), is given by:

$$|H(n, z)|^2 = \left.\frac{B_n(z)B_n(z^{-1})}{A_n(z)A_n(z^{-1})}\right|_{z=e^{-2\pi i\omega}} \tag{2.59}$$

with the polynomials $A_n(z)$ and $B_n(z)$ defined according to

$$A_n(z) \triangleq \sum_{i=1}^{p} a_i[n]z^{-i}, \qquad \text{and} \qquad B_n(z) \triangleq \sum_{i=0}^{q} b_i[n]z^{-i},$$

respectively. The tangential spectrum of (2.59) preserves the majority of properties that are satisfied by the Tjøstheim's evolutionary spectrum [5].

This definition of a frozen-in-time spectrum raises an important question. Is the set of processes defined using difference equations with time-varying coefficients, as in (2.56), equivalent to the set of processes whose transfer function, suitably generalized to the time-varying setting, takes the rational form of (2.59)? In the time-invariant case, the answer is in the affirmative. But there is no unique extension of the $\mathcal{Z}$-transform to the time-varying setting,[3] and among the known generalizations none leads to a representation that captures the same class of processes as the time-domain representation of (2.56).

The most complete results relating to this issue are presented in [59, 60] and generalize, as well as essentially complete, the earlier work of [57, 58, 61–63]. Following [60], we consider the relationship among four representations of a purely nondeterministic process satisfying the MA(∞) representation of (2.51) with time-varying impulse response $h[n, m]$, and assume that there exists some function $c(n)$ such that:

$$\sum_{m=-\infty}^{n} |h[n, m]| < c[n] < \infty. \tag{2.60}$$

The four representations, defined below, include two frequency-domain representations (rational and rational-adjoint) and two time-domain representations (ARMA and ARMA-adjoint).

- **Rational.** A PND process $x[n]$ is rational if there exists some $P \in \mathbb{Z}^+$ and functions $\{\alpha_i[n] \,|\, 1 \leq i \leq P\}$, and $\{\beta_i[n] \,|\, 1 \leq i \leq P-1\}$ such that the frequency-response function $H(n, e^{j\omega})$ of the signal model can be expressed as:

$$H(n, e^{j\omega}) = \frac{\sum_{i=0}^{P-1} \beta_i[n] e^{-j\omega i}}{1 - \sum_{i=1}^{P} \alpha_i[n] e^{-j\omega i}}, \tag{2.61}$$

  where $H(n, e^{j\omega})$ is Zadeh's generalized frequency-response function defined in (2.55).

- **Rational Adjoint.** A PND process $x[n]$ is rational-adjoint if there exists some $M \in \mathbb{Z}^+$ and functions $\{\gamma_i[n] \,|\, 1 \leq i \leq M\}$, and $\{\zeta_i[n] \,|\, 1 \leq i \leq M-1\}$ such that the conjugate-adjoint frequency-response function $H^*(n, e^{j\omega})$ of the signal model can be expressed as:

$$H^*(n, e^{j\omega}) = \frac{\sum_{i=0}^{M-1} \zeta_i[n] e^{-j\omega i}}{1 - \sum_{i=1}^{M} \gamma_i[n] e^{-j\omega i}}, \tag{2.62}$$

  where $H^*(n, e^{j\omega})$ is closely related to the frequency-response function of (2.55), [4] and is defined by:

$$H^*(n, e^{j\omega}) \triangleq \sum_{i=0}^{\infty} h(n+i, n) e^{-j\omega i}. \tag{2.63}$$

[3] Consequently, there are many definitions for poles and zeros of a linear time-varying system.

[4] In fact, using the conjugate-adjoint frequency-response function in order to define the evolutionary spectrum in the manner of Tjøstheim leads to the so-called transitory evolutionary spectrum—a time-frequency dual of Priestley's approach [63].

- **ARMA.** A PND process $x[n]$ is ARMA if there exists some $Q \in \mathbb{Z}^+$ and functions $\{a_i[n] \,|\, 1 \leq i \leq Q\}$, and $\{b_i[n] \,|\, 1 \leq i \leq Q-1\}$ such that $x[n]$ can be expressed in the form:

$$x[n] = \sum_{i=1}^{Q} a_i[n]x[n-i] + \sum_{i=0}^{Q-1} b_i[n]e[n-i], \tag{2.64}$$

  where $e[n]$ is a stationary white sequence.

- **ARMA Adjoint.** A PND process $x[n]$ is ARMA adjoint if there exists some $N \in \mathbb{Z}^+$ and functions $\{\phi_i[n] \,|\, 1 \leq i \leq N\}$, and $\{\chi_i[n] \,|\, 1 \leq i \leq N-1\}$ such that $x[n]$ can be expressed in the form:

$$x[n] = \sum_{i=0}^{N-1} \chi_i[n]u[n-i] \qquad \text{where} \qquad u[n] = \sum_{i=1}^{N} \phi_i[n]u[n-i] + e[n], \tag{2.65}$$

  and $e[n]$ is a stationary white sequence.

In the stationary case, when $a_i[n] = \alpha_i[n] = \phi_i[n] = \gamma_i[n] = a_i$ and $b_i[n] = \beta_i[n] = \zeta_i[n] = \chi_i[n] = b_i$ for all i and all $n \in \mathbb{Z}$, the four representations are equivalent and it is easy to transform among them. But this exchangeability does not generalize to the time-varying setting.

Under a natural assumption that ensures that the parameters P, M, Q, N are as small as possible, so that unique representations based on each of the four parameterizations are compared, it can be shown that the four parameterizations describe distinct classes of nonstationary processes [60]. The classes overlap under certain conditions as summarized in Figure 2.11 (adapted from [60]).

When the coefficients in all the representations are constant, then the classes formally overlap; when the coefficients vary, the amount of overlap among classes depends on the underlying rate of temporal variation, as should be expected.

2.7 Summary

In this chapter we have reviewed the physiology of speech production, discussed how it explains the nonstationarity of speech waveforms, and described a number of the time-series models it motivates. In the context of the source-filter model of speech, we discussed autoregressive or all-pole modeling of the vocal tract. These techniques form inform our contributions to formant tracking in Chapter 3 and form the basis for our treatment of time-varying autoregressive models in Chapters 4 and 5. Our discussion of fixed-resolution short-time Fourier or spectrographic speech analysis serves to motivate the development of variable-resolution signal-adaptive time-frequency signal representations in Chapter 6. In order to foreshadow the semiparametric approach developed in Chapter 7, we described a number of approaches to modeling the source waveform. Finally, we collected a number of useful results concerning stationary and nonstationary processes that will be needed throughout this thesis.

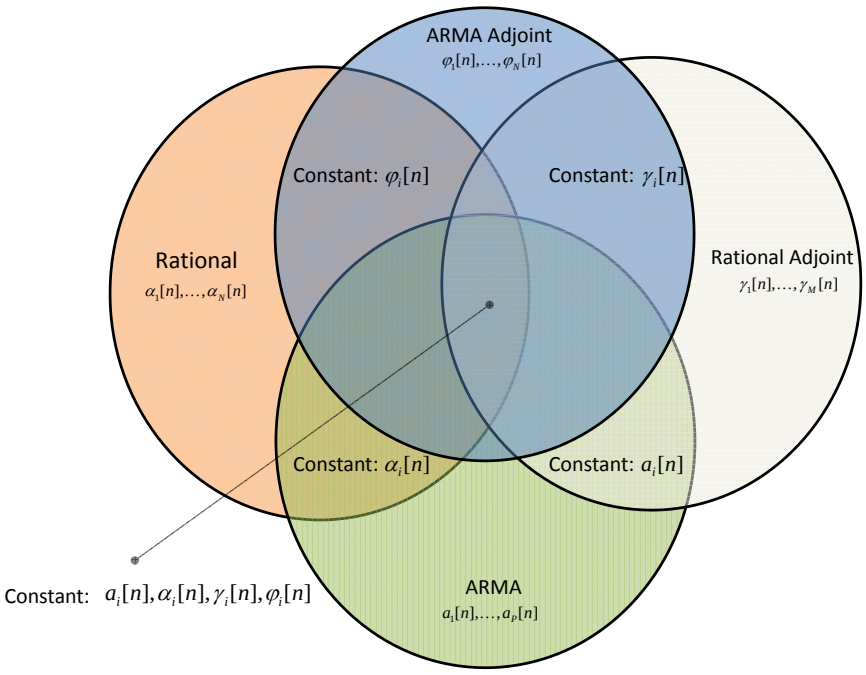


Figure 2.11: Four sets of unique minimum-order nonstationary parametric processes and some of conditions necessary for their intersection.

Chapter 3

Formant Tracking

Modern speech processing algorithms for a broad range of applications, ranging from coding and enhancement to recognition and time-scale modification, rely on tracking the time-varying characteristics of the source waveform and the vocal tract over the duration of each utterance. Typically, this is accomplished by assuming that the vocal tract is approximately time invariant on the segmental (15–30 ms) scale and estimating a sequence of desired features from a series of short-time speech segments induced by a smooth, time-localized sliding window; this amounts to approximating a nonstationary process with a piecewise-stationary one. In this setting, an important subproblem is estimating the temporal trajectories of the vocal tract resonances from an observed acoustic waveform. The present chapter comprises a number of contributions toward this formant tracking problem.

3.1 Introduction

Vocal tract resonances—often termed formants—play a central role in the perception and analysis of speech sounds [3, 7, 64, 65]. The problem of accurately estimating the temporal trajectory of vocal tract resonances over the duration of an utterance—commonly known as formant tracking—has generated especially high interest in the speech community over the last four decades [1, 66–78], and remains of interest today [2, 79–83].

All the approaches to formant-tracking aim to estimate the unobserved vocal tract resonance (VTR) frequencies and bandwidths directly from observed speech waveform data. However, auxiliary algorithms such as root-finding or peak-picking [67, 69, 70, 73, 76, 77] are typically required in order to infer the hidden variables from the observations, and hence do not readily yield analytic formulations. Nevertheless, these algorithms are often implemented in practical systems such as WaveSurfer [49], which employs a Viterbi-based algorithm due to Talkin [70]. In contrast, Deng et. al, [1, 2] recently showed the relationship among the formant frequencies and bandwidths and the linear predictive coding (LPC) cepstrum of speech enables the specification of a *generative probabilistic model* for VTR trajectories, which leads to an elegant linear-Gaussian state space model for formant tracking, with inference realized via an extended Kalman smoother (EKS).

In this chapter, we generalize the approach of [1, 2] in a number of directions and benchmark the resultant algorithms using both synthetic and recorded speech waveform

data. While retaining the conditional linear Gaussian formulation, we extend the entire framework to enable estimation of vocal tract anti-resonance (anti-formant) trajectories—essential for modeling the effects of oral and nasal cavity coupling [3]. To automate the requisite parameter estimation step, we propose both offline and online system identification algorithms that allow for explicit modeling of correlation structure across formants and anti-formants. In addition, we extend the model to account for the absence of waveform energy during silences by way of a censored likelihood formulation; adopt a Taylor-based linearization of the formant-to-cepstrum map, in contrast to the piecewise linearization approach of [2]; and benchmark the resultant EKS-based tracking algorithm using particle filtering.

Results show that the framework accurately tracks formants and anti-formants both in synthesized and natural speech. Evaluations using a recently introduced public database of formant trajectories [78] indicate reduction of the root mean-square error relative to a benchmark formant analysis technique of up to 30% per formant. In addition, we include a case study focused on the feasibility of estimating VTR bandwidths, which provides evidence to support the known difficulties of bandwidth estimation [84, 85].

The rest of this chapter is organized as follows. In Section 3.2 we formulate the precise model to be considered and contrast it to classical formant tracking approaches; the required transformations from polynomial coefficients and formant frequencies/bandwidths to the LPC cepstrum are derived in Section 3.3. Next, in Section 3.4 we describe two inference algorithms for the proposed model based on extended Kalman and particle filtering, respectively. In Section 3.5, we propose offline and online approaches to system identification, and show how to incorporate a speech activity detector into our framework in Section 3.6. In Section 3.7, we illustrate the performance of the proposed model using synthetic and natural speech data, including a recently released public database [78] of formant trajectories. A short summary follows in Section 3.8.

3.2 Model Formulation

At its core, formant tracking is nothing more than estimating the temporal trajectory of the speech spectral envelope over the duration of an utterance. Consequently, a large majority of approaches to formant tracking involve selecting an appropriate representation of the spectral envelope and modeling the temporal evolution of the associated parameters. We review the classical approach to formant tracking from this point of view, including the specific parametrization employed, in Section 3.2.1, and contrast it with our approach delineated in Sections 3.2.2, 3.2.3, and 3.2.4.

3.2.1 Classical Approach

The classical approach to formant tracking is based on the assumption that the spectral envelope of 15–30 ms short-time speech segment is well characterized by $p/2$ complex-conjugate pole pairs each of which corresponds to a second-order digital resonator, and $q/2$ complex-conjugate zero pairs. As discussed earlier in Chapter 2, the poles are used for modeling the vocal tract and, to first order, the source waveform. The zeros are used

for modeling spectral effects introduced by the coupling of the oral and nasal cavities, the back cavity formed during constrictions, and the radiation load [3]—most formant tracking algorithms, however, do not take spectral zeros into account (i.e., $q = 0$).

Specifically, let $\boldsymbol{\alpha}_t, \overline{\boldsymbol{\alpha}}_t \in \mathbb{C}^{p/2\times 1}$ and $\boldsymbol{\beta}_t, \overline{\boldsymbol{\beta}}_t \in \mathbb{C}^{q/2\times 1}$ denote the vectors of complex-conjugate pole and zero pairs, respectively. Then the speech spectrum at *frame time index* t is parameterized according to:

$$H(t,z) \triangleq \frac{\prod_{j=1}^{q/2}(1-\boldsymbol{\beta}_t[j]z^{-1})(1-\overline{\boldsymbol{\beta}_t[j]}z^{-1})}{\prod_{i=1}^{p/2}(1-\boldsymbol{\alpha}_t[i]z^{-1})(1-\overline{\boldsymbol{\alpha}_t[i]}z^{-1})}, \tag{3.1}$$

with the associated power spectrum given by $|H(t,z))|^2$.

Next, in keeping with the second-order digital resonator analogy, we parameterized each complex-conjugate pair of poles and zeros by a frequency and 3-dB bandwidth (both in units of Hertz). Specifically, the ith pole pair $(\boldsymbol{\alpha}_t[i], \overline{\boldsymbol{\alpha}_t[i]})$ and jth zero pair $(\boldsymbol{\beta}_t[j], \overline{\boldsymbol{\beta}_t[j]})$ at frame time index t are parameterized by the frequency-bandwidth pairs $(f_t^p[i], b_t^p[i]) \in \mathbb{R}_+^2$ and $(f_t^z[i], b_t^z[j]) \in \mathbb{R}_+^2$, respectively. The frequencies corresponding to the pole and zero pairs are often informally referred to as formant and anti-formant frequencies. These two parameterizations are related via:

$$\begin{aligned}\alpha_t[i] &= \exp\left(-\pi \boldsymbol{b}_t^p[i]/f_s - 2\pi \mathtt{i}\, \boldsymbol{f}_t^p[i]/f_s\right)\\ \beta_t[j] &= \exp\left(-\pi \boldsymbol{b}_t^z[j]/f_s - 2\pi \mathtt{i}\, \boldsymbol{f}_t^z[j]/f_s\right),\end{aligned} \tag{3.2}$$

where f_s is the sampling frequency of the acoustic waveform, $\mathtt{i} \triangleq \sqrt{-1}$ is the complex unit, and $\boldsymbol{f}_t^p,\ \boldsymbol{b}_t^p \in \mathbb{R}_+^{p/2}$ and $\boldsymbol{f}_t^z,\ \boldsymbol{b}_t^z \in \mathbb{R}_+^{q/2}$ are vectors of formant and anti-formant frequencies/bandwidths, respectively.

Using the relationship of (3.2), classical formant tracking algorithms (see, e.g., [66, 67, 70]) proceed as follows. The speech waveform is pre-emphasized and windowed to produce a sequence of T short-time speech segments. An ARMA(p,q) model is then fitted to each short-time segment and polynomials defined by the estimated coefficients are factored, using an appropriate numerical procedure, to produce estimates of poles and zeros. In turn, estimates of the formant frequencies and bandwidths are obtained via (3.2). Finally, the sequence of estimates across the entire speech utterance is smooothed in order to remove outliers—a variety of techniques can be employed for this including dynamic programming as popularized by [70].

The entire process is illustrated in Figure 3.1. Even though there is no generative model underlying the above procedure, the two components corresponding to intra-frame parameter estimation and to inter-frame smoothing are readily discernible. Note, in particular, that the required root-finding (or, equivalently, peak-picking) procedure must be implemented numerically, and cannot be written in closed form. This renders statistical analysis of the resultant estimators very difficult.

3.2.2 LPC Cepstral Approach

Our approach is based on a fundamentally different idea, originally proposed by [1, 2]. As before, the speech is pre-emphasized and windowed to produce a sequence

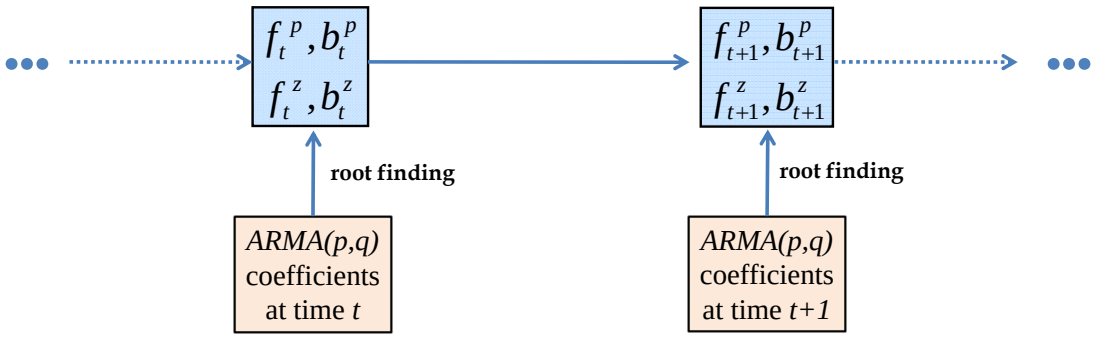


Figure 3.1: Illustration of the classical approach to formant tracking. For each frame time index t an ARMA(p, q) model is fit to the associated short-time speech segment and the estimated coefficients are converted to formant frequency/bandwidth estimates via root-finding procedure. Estimates are then smoothed across the entire speech utterance.

of T short-time speech segments, an ARMA(p, q) model is then fitted to each short-time segment. However, in lieu of a root finding or peak-picking procedure, the estimated ARMA coefficients are instead converted to N LPC cepstral coefficients, where the nth coefficient is defined by the inverse $\mathcal{Z}$-transform of the log-spectrum according to:

$$c_n \triangleq \frac{1}{2\pi} \int_{z=e^{iw}} \log H(t, z) z^{n-1} dz, \tag{3.3}$$

where the integral is taken counterclockwise along the unit circle. In practice, the transformation of (3.3) can be efficiently implemented using the recursive relation described in Section 3.3.2 below.

Next, the spectral envelope parameters $\boldsymbol{f}_t^p$, $\boldsymbol{b}_t^p$ and $\boldsymbol{f}_t^z$, $\boldsymbol{b}_t^z$ are obtained from the LPC cepstral coefficients through a nonlinear mapping $h(\cdot)$ we specify below in *closed form*. We will show that a linearization of $h(\cdot)$ leads to an efficient minimum-mean-squared error (MMSE) estimator. The resultant estimates are then smoothed across time through an explicit model constraining the temporal evolution of the spectral envelope parameters. The entire process is illustrated in Figure 3.2.

The closed-form specification of $h(\cdot)$, coupled with an explicit model for the temporal evolution of the spectral envelope parameters, readily lends itself to an interpretation in terms of a *generative probabilistic model* for vocal tract resonance trajectories, and yields an analytic formulation in a state-space modeling framework as we detail next.

3.2.3 State-Space Model

We now formalize our approach using a state-space model. The spectral envelope at frame time index t is parameterized by a vector $\boldsymbol{x}_t \in \mathbb{R}_+^{(p+q)}$, where

$$\boldsymbol{x}_t \triangleq \begin{pmatrix} \boldsymbol{f}_t^{pT} & \boldsymbol{b}_t^{pT} & \boldsymbol{f}_t^{zT} & \boldsymbol{b}_t^{zT} \end{pmatrix}^T,$$

and the vectors $\boldsymbol{f}_t^p$, $\boldsymbol{b}_t^p \in \mathbb{R}_+^{p/2\times 1}$ and $\boldsymbol{f}_t^z$, $\boldsymbol{b}_t^z \in \mathbb{R}_+^{q/2\times 1}$ represent pole and zero frequencies/bandwidths, respectively. The transfer function $H(t, z)$ associated to frame time index

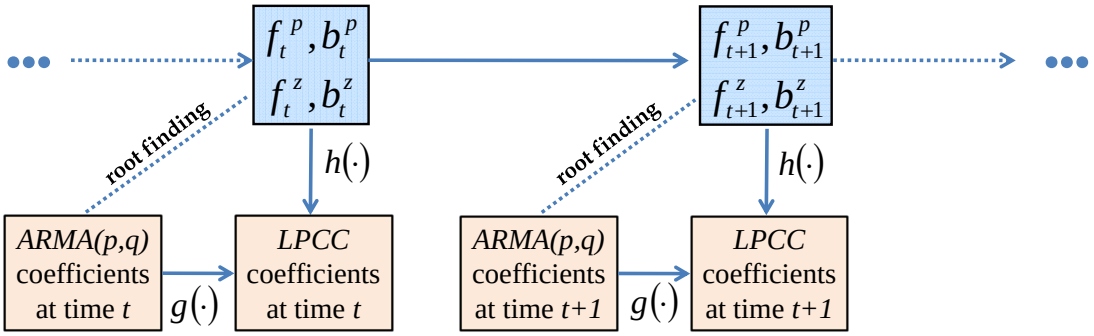


Figure 3.2: Proposed formant tracking approach. For each frame time index t an ARMA(p,q) model is fit to the associated short-time speech segment and the estimated coefficients are then converted to LPC cepstral coefficients using a recursive relationship $g(\cdot)$. Formant frequency/bandwidth estimates are then obtained from the LPC cepstral coefficients using an MMSE estimator obtained by a Taylor approximation to $h(\cdot)$. Estimates are then smoothed across the entire speech utterance.

t may be computed from $\boldsymbol{x}_t$ using (3.2) and is given by (3.1).

We model the *temporal* evolution of the spectral envelope parameters $\boldsymbol{x}_t$ according to a discrete-time Gauss-Markov process, and assume that at each frame time index t we observe a *nonlinear* function of the state $h(\boldsymbol{x}_t) \in \mathbb{R}^N$ embedded in additive white Gaussian noise. The function $h : \mathbb{R}_+^{p+q} \to \mathbb{R}^N$ is a vector-valued nonlinear mapping of the state $\boldsymbol{x}_t$ to the first N coefficients of the LPC cepstrum[1] and is defined *coordinate-wise* by:

$$\boldsymbol{c}_t[n] = \frac{2}{n}\sum_{i=1}^{p/2} \exp\left(-\frac{\pi n}{f_s}\boldsymbol{b}_t^p[i]\right)\cos\left(\frac{2\pi n}{f_s}\boldsymbol{f}_t^p[i]\right) - \frac{2}{n}\sum_{j=1}^{q/2} \exp\left(-\frac{\pi n}{f_s}\boldsymbol{b}_t^z[j]\right)\cos\left(\frac{2\pi n}{f_s}\boldsymbol{f}_t^z[j]\right). \tag{3.4}$$

The explicit form of (3.4) is derived in detail in Section 3.3.1 below. This leads to the following state-space model:

$$\begin{aligned} \boldsymbol{x}_{t+1} &= \boldsymbol{F}\boldsymbol{x}_t + \boldsymbol{w}_t, \\ \boldsymbol{y}_t &= h\left(\boldsymbol{x}_t\right) + \boldsymbol{v}_t, \end{aligned} \tag{3.5}$$

where $\boldsymbol{F} \in \mathbb{R}^{(p+q)\times(p+q)}$ is the state transition matrix and $\boldsymbol{w}_t \in \mathbb{R}^{(p+q)}$ is a white Gaussian noise sequence satisfying $\mathbb{E}(\boldsymbol{w}_t\boldsymbol{w}_{t'}^T) = \boldsymbol{Q}\delta[t-t']$, with $\boldsymbol{Q}$ denoting the process noise covariance. We also assume that $\boldsymbol{v}_t \in \mathbb{R}^N$ is a white Gaussian sequence satisfying $\mathbb{E}(\boldsymbol{v}_t\boldsymbol{v}_{t'}^T) = \boldsymbol{R}\delta[t-t']$, with $\boldsymbol{R}$ denoting the observation noise covariance. The process and observation noise sequences are uncorrelated and satisfy $\mathbb{E}(\boldsymbol{v}_t\boldsymbol{w}_{t'}^T) = \boldsymbol{0}$ for all t and t'. We also assume that the initial state $\boldsymbol{x}_0$ follows a Normal distribution with mean $\boldsymbol{\mu}_0 \in \mathbb{R}_+^{(p+q)}$ and covariance $\boldsymbol{\Sigma}_0 \in \mathbb{R}^{(p+q)\times(p+q)}$.

In summary, the state-space model of (3.5) is parameterized by the set $\boldsymbol{\theta}$ defined by:

$$\boldsymbol{\theta} \triangleq \{\boldsymbol{F}, \boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\}. \tag{3.6}$$

[1]Not counting the 0th cepstral coefficient.

Observe that in contrast to the popular dynamic programming approaches previously cited, this statistical formulation will allow us to characterize the uncertainty (i.e., covariance) in the point estimates of formant and anti-formant parameters.

3.2.4 Model Constraints

Careful consideration of the nonlinear mapping of (3.4) reveals that the frequency values of the poles and zeros $\boldsymbol{f}_t^p$ and $\boldsymbol{f}_t^z$ outside the interval $[0\ f_s/2)$ will be mapped to *nonzero* values of the complex cepstrum due to the periodicity of the cosine term. In practice, this implies that the first formant may suddenly begin to track a reflection of itself about 0 Hz since $\cos(\pi n \boldsymbol{f}_t^p[1]/f_s) = \cos(-\pi n \boldsymbol{f}_t^p[1]/f_s)$ for any $1 \leq n \leq N$. Similarly, the top formant may begin to track a reflection of itself about the Nyquist frequency of $f_s/2$ Hz since $\cos(2\pi n \boldsymbol{f}_t^p[1]/f_s) = \cos(2\pi n(kf_s - \boldsymbol{f}_t^p[1])/f_s)$, for any $1 \leq n \leq N/2$ and $k \in \mathbb{Z}$. Consequently, it may be desirable to constrain the state so that all elements of $\boldsymbol{f}_t^p$ and $\boldsymbol{f}_t^z$ lie in the interval $[0, f_s/2)$ for each t.

Another constraint of interest is to require that the poles and zeros are ordered and separated from their neighbors by some minimal distance such as:

$$\begin{aligned}\boldsymbol{f}_t^p[1] < \boldsymbol{f}_t^p[2] + \delta < \ldots < \boldsymbol{f}_t^p[p-1] + \delta < \boldsymbol{f}_t^p[p],\\ \boldsymbol{f}_t^z[1] < \boldsymbol{f}_t^z[2] + \delta < \ldots < \boldsymbol{f}_t^z[q-1] + \delta < \boldsymbol{f}_t^z[q],\end{aligned}$$

where $\delta \geq 0$ is a minimum separation (in units of Hz) among the formant frequencies. The constrained estimation problem that results from the incorporation of such constraints inevitably requires more sophisticated inference algorithms, such as particle filters, which we discuss in Section 3.4.3.

3.3 Transformations to the LPC Cepstrum

Having described the state-space model of (3.5), we now derive a number of transformations necessary for its implementation. In Section 3.3.1, we derive the mapping between the state variable $\boldsymbol{x}_t$, comprising frequency/bandwidth pairs for poles and zeros, and the LPC cepstrum. Then, in Section 3.3.2, we derive a recursive relationship between ARMA coefficients and the LPC cepstrum.

3.3.1 From Frequencies and Bandwidths to LPC Cepstrum

The relationship of (3.4) between the formant-bandwidth representation of the spectral envelope and the LPC cepstrum has been derived in the all-pole case (see e.g., [1, 74]); here we generalize it to the pole-zero setting. To begin, observe that the complex

logarithm of the spectral envelope of (3.1) is given by:

$$\begin{aligned}\log(H(t,z)) &= \log\left(\frac{\prod_{j=1}^{q/2}(1-\boldsymbol{\beta}_t[j]z^{-1})(1-\overline{\boldsymbol{\beta}_t[j]}z^{-1})}{\prod_{i=1}^{p/2}(1-\boldsymbol{\alpha}_t[i]z^{-1})(1-\overline{\boldsymbol{\alpha}_t[i]}z^{-1})}\right)\\ &= \sum_{j=1}^{q/2}\log(1-\boldsymbol{\beta}_t[j]z^{-1}) + \sum_{j=1}^{q/2}\log(1-\overline{\boldsymbol{\beta}_t[j]}z^{-1})\\ &\quad - \sum_{i=1}^{p/2}\log(1-\boldsymbol{\alpha}_t[i]z^{-1}) - \sum_{i=1}^{p/2}\log(1-\overline{\boldsymbol{\alpha}_t[i]}z^{-1}),\end{aligned}$$

and using the associated Taylor-series expansion yields:

$$\begin{aligned}\log(H(t,z)) &= -\sum_{j=1}^{q/2}\sum_{n=1}^{\infty}\frac{\boldsymbol{\beta}_t^n[j]z^{-n}}{n} - \sum_{j=1}^{q/2}\sum_{n=1}^{\infty}\frac{(\overline{\boldsymbol{\beta}_t[j]})^n z^{-n}}{n} + \sum_{i=1}^{p/2}\sum_{n=1}^{\infty}\frac{\boldsymbol{\alpha}_t^n[i]z^{-n}}{n} + \sum_{i=1}^{p/2}\sum_{n=1}^{\infty}\frac{(\overline{\boldsymbol{\alpha}_t[i]})^n z^{-n}}{n}\\ &= -\sum_{j=1}^{q/2}\sum_{n=1}^{\infty}\frac{\left(\boldsymbol{\beta}_t^n[j] + (\overline{\boldsymbol{\beta}_t[j]})^n\right)z^{-n}}{n} + \sum_{i=1}^{p/2}\sum_{n=1}^{\infty}\frac{\left(\boldsymbol{\alpha}_t^n[i] + (\overline{\boldsymbol{\alpha}_t[i]})^n\right)z^{-n}}{n}.\end{aligned}$$

Let $c_t[0]$ denote the 0th cepstral coefficient at frame time index t. Since $\log(H(t,z)) = c_t[0] + \sum_{n=1}^{\infty}\boldsymbol{c}_t[n]z^{-n}$, by construction, it is easy to see that equating the coefficients of powers of z^{-1} leads to:

$$\begin{aligned}\boldsymbol{c}_t[n] &= \frac{1}{n}\sum_{i=1}^{p/2}\left(\boldsymbol{\alpha}_t^n[i] + \overline{\boldsymbol{\alpha}_t[i]}^n\right) - \frac{1}{n}\sum_{j=1}^{q/2}\left(\boldsymbol{\beta}_t^n[j] + \overline{\boldsymbol{\beta}_t[j]}^n\right)\\ &= \frac{2}{n}\sum_{i=1}^{p/2}\exp\left(-\frac{\pi n}{f_s}\boldsymbol{b}_t^p[i]\right)\cos\left(\frac{2\pi n}{f_s}\boldsymbol{f}_t^p[i]\right) - \frac{2}{n}\sum_{j=1}^{q/2}\exp\left(-\frac{\pi n}{f_s}\boldsymbol{b}_t^z[j]\right)\cos\left(\frac{2\pi n}{f_s}\boldsymbol{f}_t^z[j]\right),\end{aligned}$$

where the last equality was obtained by substituting (3.2) and applying Euler's formula. This yields the exact form of the observation equation in the state space model of (3.5).

3.3.2 Relating ARMA Coefficients to LPC Cepstrum

In order to infer formant/anti-formant frequencies and bandwidths using the state-space model of (3.5), it is necessary to obtain a set of N LPC cepstral coefficients $\boldsymbol{y}_t \triangleq \left(\boldsymbol{y}_t[1]\ \ \boldsymbol{y}_t[2]\ \ \cdots\ \ \boldsymbol{y}_t[N]\right)^T$ from each tth short-time speech segment (see e.g., Figure 3.2). The general approach proceeds by fitting an ARMA(p,q) model to the acoustic waveform and, subsequently transforming the *rational* transfer function parameterized by the resultant coefficients to the cepstral domain. This transformation was detailed for the case of all-pole models in [1], here it is generalized to the pole-zero setting.

Consider a *minimum-phase*, rational transfer function $H(z)$ corresponding to a system with impulse response given by the sequence $(h_n \,|\, n \in \mathbb{Z})$. As a first step, we derive

the well-known recursion from $(h_n \,|\, n \in \mathbb{Z})$ to the complex cepstrum $\{c_n \,|\, n \in \mathbb{Z}\}$. Let $C(z)$ denote the $\mathcal{Z}$-transform of the complex cepstrum as in:

$$C(z) \triangleq \log\left(H(z)\right) = \sum_{n=-\infty}^{\infty} c_n z^{-n}. \tag{3.7}$$

Then using properties of the complex logarithm and the chain rule one can show that

$$\frac{dC(z)}{dz} = \frac{1}{H(z)} \frac{dH(z)}{dz}, \tag{3.8}$$

which is equivalent to $nh_n = (nc_n) * h_n$ in the time-domain [86]. Writing out this convolution explicitly as a summation yields (for $n > 0$):

$$h_n = \sum_{k=-\infty}^{n} \left(\frac{k}{n}\right) c_n h_{n-k} = \sum_{k=1}^{n} \left(\frac{k}{n}\right) c_k h_{n-k}, \tag{3.9}$$

where the second equality follows since $H(z)$ has no poles or zeros outside the unit circle (i.e., it is minimum-phase) and the complex cepstrum is, consequently, *right-sided* (i.e., $h_n = 0$ for all $n < 0$). The recursive relationship of (3.9) can be reversed to yield:

$$c_n = \begin{cases} \frac{h_n}{h_0} - \sum_{k=0}^{n-1} \left(\frac{k}{n}\right) c_k \frac{h_{n-k}}{h_0} & \text{if} \quad n > 0 \\ \log(\sigma^2) & \text{if} \quad n = 0 \\ 0 & \text{if} \quad n < 0 \end{cases}, \tag{3.10}$$

where we have used the Szego-Kolmogorov formula:

$$\sigma^2 = \exp\left(\frac{1}{2\pi} \int_{z=e^{iw}} \log\left(|H(z)|\right) dz\right),$$

with σ^2 denoting the average power of the system.

A recursive transformation similar to (3.10) may be derived from the predictor coefficients of an ARMA(p, q) model to the complex cepstrum. In the all-pole case, the transfer function is defined by:

$$H(z) \triangleq \frac{1}{\prod_{i=1}^{p}(1 - \alpha_i z^{-1})},$$

where $\{\alpha_i, 1 \leq i \leq p\}$ are poles *inside* the unit circle.[2] Then using (3.8) we have that:

$$\frac{dC(z)}{dz^{-1}} = \frac{1}{H(z)} \frac{dH(z)}{dz^{-1}} = \frac{\sum_{i=1}^{p} i a_i z^{-i+1}}{1 - \sum_{i=1}^{p} a_i z^{-i}},$$

which together with (3.7) implies that

$$\frac{\sum_{i=1}^{p} i a_i z^{-i+1}}{1 - \sum_{i=1}^{p} a_i z^{-i}} = \sum_{n=1}^{\infty} c_n \frac{d}{dz^{-1}}\left(z^{-n}\right) = \sum_{n=1}^{\infty} n c_n z^{-n+1}.$$

[2]Note that the presence of complex-conjugate pole pairs (or zero pairs below) need not be assumed for this development.

Rearranging the terms above we obtain:

$$\sum_{n=1}^{\infty} nc_n z^{-n+1} = \sum_{i=1}^{p} ia_i z^{-i+1} + \sum_{i=1}^{p} a_i z^{-i} \sum_{n=1}^{\infty} nc_n z^{-n+1}. \tag{3.11}$$

Using (3.11), we can match the coefficients of terms with equal exponents. In the constant-coefficient case (associated to z^0) we have: $c_1 = a_1$. Fixing some $1 < n \leq p$, we obtain:

$$\begin{aligned} nc_n z^{-n+1} &= na_n z^{-n+1} + \sum_{i=1}^{n-1} (n-i) a_i z^{-i} c_{n-i} z^{i-n+1} \\ &= na_n z^{-n+1} + \sum_{i=1}^{n-1} (n-i) a_i c_{n-i} z^{-n+1}, \end{aligned}$$

which leads to

$$c_n = a_n + \sum_{i=1}^{n-1} \frac{(n-i)}{n} a_i c_{n-i} = a_n + \sum_{i=1}^{n-1} \left(1 - \frac{i}{n}\right) a_i c_{n-i}.$$

On the other hand, if $n > p$, then (3.11) implies that:

$$c_n = \sum_{i=n-p}^{n-1} \frac{(n-i)}{n} a_i c_{n-i} = \sum_{i=1}^{n-1} \left(1 - \frac{i}{n}\right) a_i c_{n-i}.$$

In summary, we have obtained the following relationship between the prediction polynomial coefficients and the complex cepstrum:

$$c_n = \begin{cases} a_1 & \text{if} \quad n = 1 \\ a_n + \sum_{i=1}^{n-1} \left(\frac{n-i}{n}\right) a_i c_{n-i} & \text{if} \quad 1 < n \leq p \\ \sum_{i=n-p}^{n-1} \left(\frac{n-i}{n}\right) a_i c_{n-i} & \text{if} \quad p < n. \end{cases} \tag{3.12}$$

An alternative, perhaps more prevalent, formulation can be obtained by substituting k for $n - i$ and then i for k to yield:

$$c_n = \begin{cases} a_1 & \text{if} \quad n = 1 \\ a_n + \sum_{i=1}^{n-1} \left(\frac{i}{n}\right) a_{n-i} c_i & \text{if} \quad 1 < n \leq p \\ \sum_{i=n-p}^{n-1} \left(\frac{i}{n}\right) a_{n-i} c_i & \text{if} \quad p < n. \end{cases} \tag{3.13}$$

In the case of rational functions with poles and zeros *inside* the unit circle, $H(z)$ is defined by:

$$H(z) \triangleq \frac{\prod_{j=1}^{q}(1 - \beta_j z^{-1})}{\prod_{i=1}^{p}(1 - \alpha_i z^{-1})},$$

where $\{\alpha_i, 1 \leq i \leq p\}$ and $\{\beta_j, 1 \leq j \leq q\}$ are poles and zeros, respectively. The desired recursive relationship—analogous to (3.13) in the all-pole case—is derived as follows. First, we separate the contributions of the poles and zeros to each cepstral coefficient c_n via:

$$c_n = \mathcal{Z}^{-1} \log \left(\frac{B(z)}{A(z)}\right) = \mathcal{Z}^{-1} \log \left(\frac{1}{A(z)}\right) - \mathcal{Z}^{-1} \log \left(\frac{1}{B(z)}\right) = c_n^p - c_n^z.$$

Clearly, the recursion of (3.13) implies that

$$c_n^p = \begin{cases} a_1 & \text{if} \quad n = 1 \\ a_n + \sum_{i=1}^{n-1} \left(\frac{i}{n}\right) a_{n-i} c_i^p & \text{if} \quad 1 < n \leq p \\ \sum_{i=n-p}^{n-1} \left(\frac{i}{n}\right) a_{n-i} c_i^p & \text{if} \quad p < n, \end{cases} \tag{3.14}$$

and that

$$c_n^z = \begin{cases} b_1 & \text{if} \quad n = 1 \\ b_n + \sum_{j=1}^{n-1} \left(\frac{j}{n}\right) b_{n-j} c_j^z & \text{if} \quad 1 < n \leq q \\ \sum_{j=n-q}^{n-1} \left(\frac{j}{n}\right) b_{n-j} c_j^z & \text{if} \quad q < n. \end{cases} \tag{3.15}$$

3.3.3 Summary of Pre-Processing Steps

In summary, in order to obtain the set of N LPC cepstral coefficients $\boldsymbol{y}_t$ from waveform data in the tth short-time frame, the predictor coefficients of an ARMA(p, q) LTI system are first obtained and, subsequently, used to calculate the coefficients of the associated complex-cepstrum representation through the recursions of (3.14) and (3.15). Following pre-emphasis and windowing steps, a number of standard spectral estimation techniques (see e.g., [4]) could be employed in order to fit the associated ARMA(p, q) model. However, it is important to point out that the estimated poles and zeros *must* lie inside the unit circle since the relationships of (3.14) and (3.15) were derived under this minimum-phase assumption. In the all-pole case, for instance, this suggests using the autocorrelation method of linear prediction as it is guaranteed to yield pole-location estimates constrained to lie in the unit disc [3].

Given a sequence of LPC cepstral coefficients corresponding to a sequence of short-time speech segments, we may now use the state-space model of (3.5) to infer the underlying formant/anti-formant parameter trajectories as described in the next section.

3.4 Inference

Let $\boldsymbol{y}_{1:T} \triangleq \begin{pmatrix} \boldsymbol{y}_1 & \boldsymbol{y}_2 & \cdots & \boldsymbol{y}_T \end{pmatrix}$ and $\boldsymbol{x}_{1:T} \triangleq \begin{pmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_T \end{pmatrix}$ be the matrices of observations and latent variables, respectively. Our aim, in this section, is to compute a minimum-mean-squared error (MMSE) estimate of $\boldsymbol{x}_{1:T}$ from the observed data $\boldsymbol{y}_{1:T}$ under the assumption that all the parameters $\boldsymbol{\theta}$ of the state-space model of (3.5) are known. The system identification question of how to estimate $\boldsymbol{\theta}$ is addressed in Section 3.5 below.

The desired MMSE estimate of the state $\boldsymbol{x}_t$ is the mean of the posterior density $p(\boldsymbol{x}_t|\boldsymbol{y}_{1:T}; \boldsymbol{\theta})$; its dispersion characterizes the associated estimator uncertainty. If the state-space model of (3.5) were *linear* and Gaussian, then it would follow that $p(\boldsymbol{x}_t \,|\, \boldsymbol{y}_{1:T}\,; \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{m}_{t|T}, \boldsymbol{P}_{t|T})$ with the quantities $\boldsymbol{m}_{t|T}$ and $\boldsymbol{P}_{t|T}$ defined according to:

$$\boldsymbol{m}_{t|T} \triangleq \mathbb{E}(\boldsymbol{x}_t \,|\, \boldsymbol{y}_{1:T}; \boldsymbol{\theta})\,, \quad \text{and} \qquad \boldsymbol{P}_{t|T} \triangleq \mathbb{E}\big((\boldsymbol{x}_t - \boldsymbol{m}_{t|T})(\boldsymbol{x}_t - \boldsymbol{m}_{t|T})^T \,|\, \boldsymbol{y}_{1:T}; \boldsymbol{\theta}\big)\,.$$

In such a case, the quantities $\boldsymbol{m}_{t|T}$ and $\boldsymbol{P}_{t|T}$ would be easily obtained by the Kalman smoother. This approach is not possible, however, due to the nonlinearity of the observation equation in (3.5).

Numerical approaches to approximating the resultant non-Gaussian posterior density are possible using Monte Carlo methods such as sequential importance sampling (i.e., particle filtering) and Markov chain Monte Carlo (see e.g., [74]). However, due to the relatively-smooth nature of (3.4), we employ a simpler approximate inference method based on a Taylor approximation of the observation equation.[3] If, in addition, none of the constraints described in Section 3.2.4 are enforced, an extended Kalman smoother can be used to obtain the desired MMSE estimates as detailed in Sections 3.4.1 and 3.4.2 below. However, we will reconsider the use of particle filtering in Section 3.4.3, in order to evaluate the quality of the linearization on which the extended Kalman smoother relies and to compute *constrained* MMSE estimates.

3.4.1 Extended Kalman Smoother

In order to approximate the posterior density $p(\boldsymbol{x}_t \,|\, \boldsymbol{y}_{1:T}\,;\,\boldsymbol{\theta})$, we employ the extended Kalman smoother (EKS), whereby at each frame time index t we *linearize* the mapping from $\boldsymbol{x}_t$ to the LPC cepstrum of (3.4) about the conditional mean $\boldsymbol{m}_{t|t-1} \triangleq \mathbb{E}(\boldsymbol{x}_t \,|\, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{t-1}\,;\,\boldsymbol{\theta})$ using a first-order Taylor expansion via:

$$h(\boldsymbol{x}_t) \approx h(\boldsymbol{m}_{t|t-1}) + \boldsymbol{H}_t(\boldsymbol{x}_t - \boldsymbol{m}_{t|t-1}). \tag{3.16}$$

The posterior mean $\boldsymbol{m}_{t|T}$ and covariance $\boldsymbol{P}_{t|T}$ are then computed using a two-pass procedure consisting of filtering (forward) and smoothing (backward) Kalman recursions as summarized in Algorithm 3.1. The specifics of how the observation matrix $\boldsymbol{H}_t$ in the Taylor expansion of (3.16) is calculated are described in Section 3.4.2 below. Note that in addition to the standard extended Kalman smoother recursions, Algorithm 3.1 computes the *lag-one* filtering and smoothing covariances $\boldsymbol{P}_{t,t-1|t}$ and $\boldsymbol{P}_{t,t-1|T}$ defined by

$$\begin{aligned}
\boldsymbol{P}_{t,t-1|t} &\triangleq \mathbb{E}\big((\boldsymbol{x}_t - \boldsymbol{m}_{t|t})(\boldsymbol{x}_{t-1} - \boldsymbol{m}_{t-1|t-1})^T \,|\, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_t\,;\,\boldsymbol{\theta}\big)\,, \text{ and} \\
\boldsymbol{P}_{t,t-1|T} &\triangleq \mathbb{E}\big((\boldsymbol{x}_t - \boldsymbol{m}_{t|T})(\boldsymbol{x}_{t-1} - \boldsymbol{m}_{t-1|T})^T \,|\, \boldsymbol{y}_{1:T}\,;\,\boldsymbol{\theta}\big)
\end{aligned}$$

are also computed. These quantities will be required in our discussion of system identification using the expectation-maximization algorithm in Section 3.5.2 below.

3.4.2 Linearization

In order to linearize $h(\boldsymbol{x}_t)$ according to the first-order Taylor expansion of (3.16), we need to compute the derivatives of the observation equation with respect to the frequencies and bandwidths of poles and zeros. Since the mapping $h(\cdot)$ is specified coordinate-wise

[3] However, we use a Taylor approximation to the nonlinearity, rather than the nonstandard piecewise-constant approximation employed in contrast to [2], which results in a simpler, more accurate and computationally efficient procedure.

Algorithm 3.1 Extended Kalman Smoother

- Initialize: Set $\boldsymbol{m}_{0|0} = \boldsymbol{\mu}_0$ and $\boldsymbol{P}_{0|0} = \Sigma_0$.
- Filtering: Repeat for $t = 1, \ldots, T$

$$\begin{aligned}
\boldsymbol{m}_{t|t-1} &= \boldsymbol{F}\boldsymbol{m}_{t-1|t-1} \\
\boldsymbol{P}_{t|t-1} &= \boldsymbol{F}\boldsymbol{P}_{t-1|t-1}\boldsymbol{F}^T + \boldsymbol{Q} \\
\boldsymbol{K}_t &= \boldsymbol{P}_{t|t-1}\boldsymbol{H}_t^T \left(\boldsymbol{H}_t\boldsymbol{P}_{t|t-1}\boldsymbol{H}_t^T + \boldsymbol{R}\right)^{-1} \\
\boldsymbol{m}_{t|t} &= \boldsymbol{m}_{t|t-1} + \boldsymbol{K}_t(\boldsymbol{y}_t - h(\boldsymbol{m}_{t|t-1})) \\
\boldsymbol{P}_{t|t} &= \left(\boldsymbol{I} - \boldsymbol{K}_t\boldsymbol{H}_t\right)\boldsymbol{P}_{t|t-1} \\
\boldsymbol{P}_{t,t-1|t} &= \left(\boldsymbol{I} - \boldsymbol{K}_t\boldsymbol{H}_t\right)\boldsymbol{F}\boldsymbol{P}_{t|t-1}
\end{aligned}$$

- Smoothing: Repeat for $t = T, \ldots, 1$

$$\begin{aligned}
\boldsymbol{S}_t &= \boldsymbol{P}_{t-1|t-1}\boldsymbol{F}^T\boldsymbol{P}_{t|t-1}^{-1} \\
\boldsymbol{m}_{t-1|T} &= \boldsymbol{m}_{t-1|t-1} + \boldsymbol{S}_t\left(\boldsymbol{m}_{t|T} - \boldsymbol{F}\boldsymbol{m}_{t-1|t-1}\right) \\
\boldsymbol{P}_{t-1|T} &= \boldsymbol{P}_{t-1|t-1} + \boldsymbol{S}_t\left(\boldsymbol{P}_{t|T} - \boldsymbol{P}_{t-1|t-1}\right)\boldsymbol{S}_t^T \\
\boldsymbol{P}_{t,t-1|T} &= \boldsymbol{P}_{t,t-1|t} + \left(\boldsymbol{P}_{t|T} - \boldsymbol{P}_{t|t}\right)\boldsymbol{P}_{t|t}^{-1}\boldsymbol{P}_{t,t-1|t}
\end{aligned}$$

in (3.4), we compute the derivatives corresponding to each coordinate to be

$$\begin{aligned}
\frac{\partial \boldsymbol{y}_t[n]}{\partial \boldsymbol{f}_t^p[i]} &= -\frac{4\pi}{f_s}\exp\left(-\frac{\pi n}{f_s}\boldsymbol{b}_t^p[i]\right)\sin\left(\frac{2\pi n}{f_s}\boldsymbol{f}_t^p[i]\right) \\
\frac{\partial \boldsymbol{y}_t[n]}{\partial \boldsymbol{b}_t^p[i]} &= -\frac{2\pi}{f_s}\exp\left(-\frac{\pi n}{f_s}\boldsymbol{b}_t^p[i]\right)\cos\left(\frac{2\pi n}{f_s}\boldsymbol{f}_t^p[i]\right) \\
\frac{\partial \boldsymbol{y}_t[n]}{\partial \boldsymbol{f}_t^z[j]} &= \frac{4\pi}{f_s}\exp\left(-\frac{\pi n}{f_s}\boldsymbol{b}_t^z[j]\right)\sin\left(\frac{2\pi n}{f_s}\boldsymbol{f}_t^z[j]\right) \\
\frac{\partial \boldsymbol{y}_t[n]}{\partial \boldsymbol{b}_t^z[j]} &= \frac{2\pi}{f_s}\exp\left(-\frac{\pi n}{f_s}\boldsymbol{b}_t^z[j]\right)\cos\left(\frac{2\pi n}{f_s}\boldsymbol{f}_t^z[j]\right).
\end{aligned}$$

Consequently, the matrix $\boldsymbol{H}_t$ appearing in the Taylor expansion of (3.16) and used in the extended Kalman recursions of Algorithm 3.1 is defined as

$$\boldsymbol{H}_t \triangleq \begin{pmatrix} \boldsymbol{H}^T(\boldsymbol{f}^p) & \boldsymbol{H}^T(\boldsymbol{b}^p) & \boldsymbol{H}^T(\boldsymbol{f}^z) & \boldsymbol{H}^T(\boldsymbol{b}^z) \end{pmatrix}, \tag{3.17}$$

where the matrices $\boldsymbol{H}(\boldsymbol{f}^p), \boldsymbol{H}(\boldsymbol{b}^p) \in \mathbb{R}^{p/2\times N}$ are defined via:

$$\boldsymbol{H}(\boldsymbol{f}^p) \triangleq \begin{pmatrix} \frac{\partial \boldsymbol{y}_t[1]}{\partial \boldsymbol{f}_t^p[1]} & \frac{\partial \boldsymbol{y}_t[2]}{\partial \boldsymbol{f}_t^p[1]} & \cdots & \frac{\partial \boldsymbol{y}_t[N]}{\partial \boldsymbol{f}_t^p[1]} \\ \frac{\partial \boldsymbol{y}_t[1]}{\partial \boldsymbol{f}_t^p[2]} & \frac{\partial \boldsymbol{y}_t[2]}{\partial \boldsymbol{f}_t^p[2]} & \cdots & \frac{\partial \boldsymbol{y}_t[N]}{\partial \boldsymbol{f}_t^p[2]} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \boldsymbol{y}_t[1]}{\partial \boldsymbol{f}_t^p[p/2]} & \frac{\partial \boldsymbol{y}_t[2]}{\partial \boldsymbol{f}_t^p[p/2]} & \cdots & \frac{\partial \boldsymbol{y}_t[N]}{\partial \boldsymbol{f}_t^p[p/2]} \end{pmatrix}, \quad \text{and} \quad \boldsymbol{H}(\boldsymbol{b}^p) \triangleq \begin{pmatrix} \frac{\partial \boldsymbol{y}_t[1]}{\partial \boldsymbol{b}_t^p[1]} & \frac{\partial \boldsymbol{y}_t[2]}{\partial \boldsymbol{b}_t^p[1]} & \cdots & \frac{\partial \boldsymbol{y}_t[N]}{\partial \boldsymbol{b}_t^p[1]} \\ \frac{\partial \boldsymbol{y}_t[1]}{\partial \boldsymbol{b}_t^p[2]} & \frac{\partial \boldsymbol{y}_t[2]}{\partial \boldsymbol{b}_t^p[2]} & \cdots & \frac{\partial \boldsymbol{y}_t[N]}{\partial \boldsymbol{b}_t^p[2]} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \boldsymbol{y}_t[1]}{\partial \boldsymbol{b}_t^p[p/2]} & \frac{\partial \boldsymbol{y}_t[2]}{\partial \boldsymbol{b}_t^p[p/2]} & \cdots & \frac{\partial \boldsymbol{y}_t[N]}{\partial \boldsymbol{b}_t^p[p/2]} \end{pmatrix}, \tag{3.18}$$

with $\boldsymbol{H}(\boldsymbol{f}^z)$ and $\boldsymbol{H}(\boldsymbol{b}^z) \in \mathbb{R}^{q/2\times N}$ defined analogously.

3.4.3 EKF Benchmarking and Constrained MMSE Inference

The extended Kalman smoother of Algorithm 3.1 linearizes the nonlinear observation function of (3.4) about the state and, in this manner, computes an approximation to the true posterior density $p\left(\boldsymbol{x}_{1:T} \,|\, \boldsymbol{y}_{1:T}\,;\, \boldsymbol{\theta}\right)$. However, it is well-known that the first-order approximations can introduce large errors in the posterior mean and covariance estimates, which may lead to sub-optimal performance or even filter divergence [87]. In addition, there is no principled way, within the EKF framework, to obtain MMSE estimates subject to the constraints discussed in Section 3.2.4.

On the other hand, a particle filtering approach—based on propagating a Monte Carlo approximation to the posterior density using sequential importance sampling—does not suffer from the above limitations. Though computationally much more demanding, the particle filtering approach allows us to address both of these issues simultaneously. In particular, the approximation converges to the true posterior density as the number of samples (i.e., particles) increases. Assuming (as before) that the state-space model parameters $\boldsymbol{\theta}$ are known, the particle filtering algorithm[4] is summarized in Algorithm 3.2. The forms of the likelihood and transition densities $p\left(\boldsymbol{y}_t \,|\, \boldsymbol{x}_{t|t-1}^{(m)}\right)$ and $p\left(\boldsymbol{x}_{t|t-1}^{(m)} \,|\, \boldsymbol{x}_{t-1|t-1}^{(m)}\right)$ follow directly from the model of (3.5) and are given by:

$$\begin{aligned} p\left(\boldsymbol{y}_t \,|\, \boldsymbol{x}_{t|t-1}^{(m)}\right) &\triangleq \mathcal{N}\left(\boldsymbol{y}_t; h(\boldsymbol{x}_{t|t-1}^{(m)}), \boldsymbol{R}\right) \\ p\left(\boldsymbol{x}_{t|t-1}^{(m)} \,|\, \boldsymbol{x}_{t-1|t-1}^{(m)}\right) &\triangleq \mathcal{N}\left(\boldsymbol{x}_{t|t-1}^{(m)}; \boldsymbol{F}\boldsymbol{x}_{t-1|t-1}^{(m)}, \boldsymbol{Q}\right). \end{aligned}$$

The specification of the proposal distribution $q\left(\boldsymbol{x}_{t|t-1}^{(m)} \,|\, \boldsymbol{x}_{t-1|t-1}^{(m)}, \boldsymbol{y}_t\right)$ is left up to the user. A popular choice is to let the proposal equal the transition density, thereby simplifying the computation of the importance weights in (3.19). The constraints described in Section 3.2.4 can also be incorporated as part of a proposal distribution. For instance, no frequency (or bandwidth) values can lie outside the range $[0 \;\; f_s/2)$.

We now use the particle filter of Algorithm 3.2 to evaluate the quality of the extended Kalman filter approximation to the true posterior density $p(\boldsymbol{x}_t \,|\, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_t; \boldsymbol{\theta})$. To this end, we simulate 100-sample data sequences according to the model of (3.5) with four complex-conjugate pole pairs (i.e., $p = 8$, $q = 0$) and $N = 15$ cepstral coefficients, and compare the EKF of Algorithm 3.1 (without the smoothing pass) and Algorithm 3.2 in terms of the root-mean-squared error (RMSE, in Hz) averaged over all frequency trajectories. The true bandwidths were provided to both algorithms and were not tracked. The results, summarized in Figure 3.3, show that as the number of particles increases the error rates of the particle filter approach those of the EKF. This confirms that the EKF approximation to the true posterior is quite good since the particle filter does not improve upon the EKF RMSE even when the number of samples is large. Similar results hold over a broad range of simulations.

---

[4] This is the most basic particle filtering approach, sometimes termed the "bootstrap filter." For a variety of more sophisticated extensions, we point the reader to the excellent text of [88].

Algorithm 3.2 Particle Filter

- Initialize: For all $m = 1, \ldots, M$
  - Draw initial particles $\boldsymbol{x}_0^{(m)} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$
  - Set the initial weights: $w_0^{(m)} = 1$
- Filtering: Repeat for $t = 1, \ldots, T$
  - Importance Sampling: Repeat for $m = 1, \ldots, M$
    - Propagate samples:
    $$\boldsymbol{x}_{t|t-1}^{(m)} \sim q\left(\boldsymbol{x}_{t|t-1}^{(m)} \,|\, \boldsymbol{x}_{t-1|t-1}^{(m)}, \boldsymbol{y}_t\right)$$
    - Evaluate importance weights:
    $$\widetilde{w}_t^{(m)} = w_{t-1}^{(m)} \frac{p\left(\boldsymbol{y}_t \,|\, \boldsymbol{x}_{t|t-1}^{(m)}\right) p\left(\boldsymbol{x}_{t|t-1}^{(m)} \,|\, \boldsymbol{x}_{t-1|t-1}^{(m)}\right)}{q\left(\boldsymbol{x}_{t|t-1}^{(m)} \,|\, \boldsymbol{x}_{t-1|t-1}^{(m)}, \boldsymbol{y}_t\right)} \tag{3.19}$$
    - Normalize weights: $w_t^{(m)} = \widetilde{w}_t^{(m)} / \sum_{m=1}^{M} \widetilde{w}_t^{(m)}$
  - Calculate number of effective particles (measure of degeneracy):
  $$N_{eff} = \left(\sum_{m=1}^{M} \left(w_t^{(m)} \right)^2 \right)^{-1}$$
  - If $N_{eff} \leq N/2$ resample particles with replacement. Repeat for $m = 1, \ldots, M$:
  $$\boldsymbol{x}_{t|t}^{m} \sim \sum_{m=1}^{M} w_t^{(m)} \delta\left(\boldsymbol{x}_{t|t} - \boldsymbol{x}_{t|t-1}^{(m)}\right); \quad w_t^{(m)} = 1$$

3.5 System Identification

We now turn to the question of estimating the parameters of the state space model $\boldsymbol{\theta}$ of (3.6) when they are unknown. First, in Section 3.5.1, we describe an offline, plug-in estimator for the parameters $\boldsymbol{\theta}$ based, in part, on output of the popular formant tracking software called WaveSurfer [49]. Then, in Section 3.5.2, we propose an online approach based on the expectation-maximization (EM) algorithm. Both approaches are illustrated using real speech data in Section 3.7.

Note that in lieu of estimating the state transition matrix $\boldsymbol{F}$, it is perhaps easiest to set it to the identity matrix as was done in [1, 2]. However, formants do not evolve independently of one another, and their temporal trajectories are not independent in frequency. For instance, in synthesis of front vowels, it is common practice to employ a *linear regression* of formant F3 onto F1 and F2 (see, e.g., [71])! As a preliminary step in our analysis, we empirically estimated the correlation structure amongst all three hand-corrected formant trajectories in the VTR database [78]. We observed the empirical formant cross-correlation function to decay slowly, implying that a set of formant values at frame t may

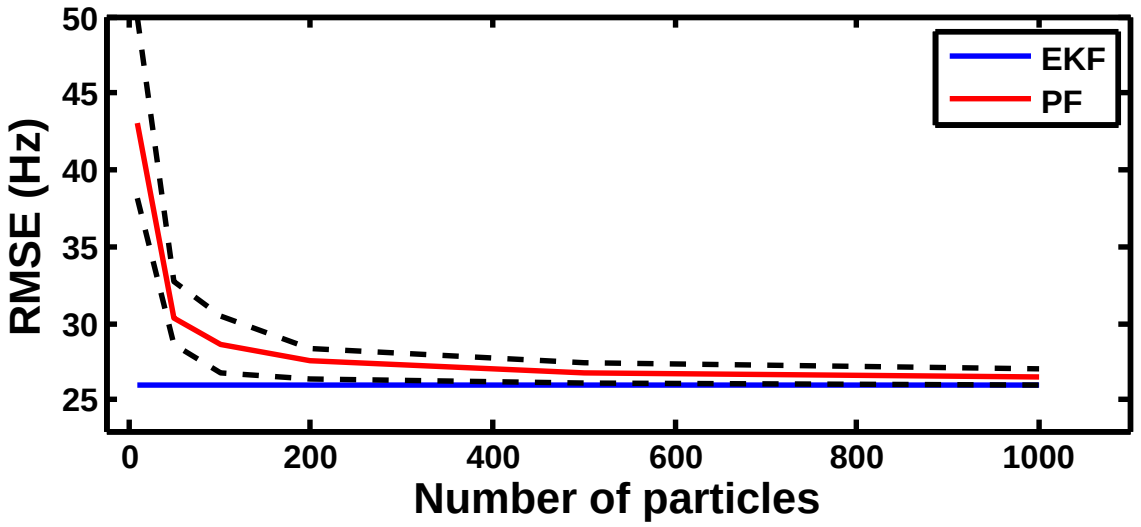


Figure 3.3: Comparison of the EKF and PF tracking performance in terms of root-mean-squared error averaged over 25 Monte Carlo trials and reported together with 95% posterior intervals.

be helpful in predicting values of all formants at frame $t+1$. This observation suggests that off-diagonal terms of the state transition matrix $\boldsymbol{F}$ can be used to incorporate the structure of the formant cross-correlation function at lag 1;[5] this allows for potentially more accurate estimation of VTR parameters.

3.5.1 Offline Approach via Plug-in Estimators

A simple, offline approach to system identification is to use the output of another formant tracker, such as WaveSurfer [49], as a means of empirically estimating all model parameters, a strategy we pursued in [79]. Thus, the state transition matrix $\boldsymbol{F}$ for a particular utterance is estimated from the WaveSurfer formant tracks corresponding to that utterance, using a linear-least-squares estimator [26]. Note that this is equivalent to fitting an order-1 vector autoregressive process to WaveSurfer formant trajectories.

The process noise covariance $\boldsymbol{Q}$ can likewise be estimated by computing the sample covariance of the first differences of the WaveSurfer trajectories. Finally, we fix the observation noise covariance matrix $\boldsymbol{R}$ to be diagonal, with terms given by $\boldsymbol{R}_{nn} = 1/n$ for $n \in \{1, 2, \ldots, N\}$. Empirically, we observed this to be in reasonable agreement with the variance of the residual vector of the LPC cepstral coefficients derived separately for WaveSurfer-generated VTR parameters from nvarious acoustic waveform data.

3.5.2 Online Approach via Expectation-Maximization

We now describe how to obtain an *online* maximum-likelihood estimate of the unknown parameters $\boldsymbol{\theta}$ from the observed data $\boldsymbol{y}_{1:T}$. Since it is difficult to work with

[5] Cross-correlation at lags greater than one may be incorporated by appropriately augmenting the state vector.

Algorithm 3.3 Formant Tracking: Online System Identification via EM Algorithm

- Initialize $\widehat{\boldsymbol{\theta}}^{(0)}$ using any appropriate method, set the number of iterations M
- For m = 1:M
  - E-Step: Compute $Q(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(m-1)})$
  - M-Step: Obtain the next estimate $\boldsymbol{\theta}^{(m)}$ according to:

$$\widehat{\boldsymbol{\theta}}^{(m)} = \sup_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(m-1)}) \tag{3.22}$$

- Return $\widehat{\boldsymbol{\theta}}^{(M)}$

the likelihood function $L(\boldsymbol{y}_{1:T}; \boldsymbol{\theta})$ directly—its exact form corresponds to that of a multivariate ARMA model [89]—we approach the problem in the context of the expectation-maximization framework [89,90]. If we were to view the latent state variable $\boldsymbol{x}_{0:T}$ as "missing data", the complete data log likelihood $L(\boldsymbol{\theta}; \boldsymbol{x}_{0:T}, \boldsymbol{y}_{1:T})$ immediately follows from (3.5) as

$$\begin{aligned} L(\boldsymbol{\theta}\,;\, \boldsymbol{x}_{0:T}, \boldsymbol{y}_{1:T}) \triangleq \log p(\boldsymbol{x}_{0:T}, \boldsymbol{y}_{1:T}\,;\, \boldsymbol{\theta}) &= -\frac{1}{2}\log|\boldsymbol{\Sigma}_0| - \frac{1}{2}(\boldsymbol{x}_0 - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{x}_0 - \boldsymbol{\mu}_0) \\ &- \frac{T}{2}\log|\boldsymbol{Q}| - \frac{1}{2}\sum_{t=1}^{T}(\boldsymbol{x}_t - \boldsymbol{F}\boldsymbol{x}_{t-1})^T \boldsymbol{Q}^{-1}(\boldsymbol{x}_t - \boldsymbol{F}\boldsymbol{x}_{t-1}) \\ &- \frac{T}{2}\log|\boldsymbol{R}| - \frac{1}{2}\sum_{t=1}^{T}(\boldsymbol{y}_t - h(\boldsymbol{x}_t))^T \boldsymbol{R}^{-1}(\boldsymbol{y}_t - h(\boldsymbol{x}_t)), \end{aligned} \tag{3.20}$$

which is a linear function of $\boldsymbol{\theta}$; this is in contrast to nonlinear relationship between $\boldsymbol{\theta}$ and the observed data likelihood $L(\boldsymbol{y}_{1:T}; \boldsymbol{\theta})$.

Let $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$ denote the expectation of the complete log likelihood defined according to:

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) &\triangleq \mathbb{E}_{p(\boldsymbol{x}_{0:T}\,|\,\boldsymbol{y}_{1:T};\boldsymbol{\theta}^{(m)})}\left(L(\boldsymbol{\theta}; \boldsymbol{x}_{0:T}, \boldsymbol{y}_{1:T})\right) \\ &= \int_{\boldsymbol{x}_{0:T}} L(\boldsymbol{\theta}; \boldsymbol{x}_{0:T}, \boldsymbol{y}_{1:T}) p(\boldsymbol{x}_{0:T}\,|\,\boldsymbol{y}_{1:T}; \boldsymbol{\theta}^{(m)}) d\boldsymbol{x}_{0:T}, \end{aligned} \tag{3.21}$$

where the expectation is with respect to $p(\boldsymbol{x}_{0:T}|\boldsymbol{y}_{1:T}; \boldsymbol{\theta}^{(m)})$—the conditional distribution of the missing data given the observed data, parameterized by $\boldsymbol{\theta}^{(m)}$. Then, the EM algorithm proceeds as described in Algorithm 3.3.

In the remainder of this section we discuss the details of how to evaluate the "E" and "M" steps of Algorithm 3.3. The primary difference between what follows and the standard approach described in [89] is that the nonlinearity in the observation equation of (3.5) needs to be linearized using a Taylor expansion so that the necessary conditional expectations can be calculated.

In order to evaluate $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$ in the E-step of Algorithm 3.3, we first use the Taylor expansion of (3.16) to approximate the nonlinearity in the likelihood function of (3.20)

as follows:

$$\boldsymbol{y}_t - h(\boldsymbol{x}_t) \approx \boldsymbol{y}_t - h(\boldsymbol{m}_{t|t-1}) - \boldsymbol{H}_t\left(\boldsymbol{x}_t - \boldsymbol{m}_{t|t-1}\right) = \widetilde{\boldsymbol{y}}_t - \boldsymbol{H}_t\boldsymbol{x}_t,$$

where $\widetilde{\boldsymbol{y}}_t \triangleq \boldsymbol{y}_t - h(\boldsymbol{m}_{t|T}) + \boldsymbol{H}_t\boldsymbol{m}_{t|T}$. Then the complete log-likelihood function of (3.20) may be approximated by:

$$\begin{aligned} L(\boldsymbol{\theta}; \boldsymbol{x}_{0:T}, \boldsymbol{y}_{1:T}) \approx \widetilde{L}(\boldsymbol{\theta}; \boldsymbol{x}_{0:T}, \boldsymbol{y}_{1:T}) \triangleq & -\frac{1}{2}\log|\boldsymbol{\Sigma}_0| - \frac{1}{2}(\boldsymbol{x}_0 - \boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{x}_0 - \boldsymbol{\mu}_0) \\ & - \frac{T}{2}\log|\boldsymbol{Q}| - \frac{1}{2}\sum_{t=1}^{T}(\boldsymbol{x}_t - \boldsymbol{F}\boldsymbol{x}_{t-1})^T\boldsymbol{Q}^{-1}(\boldsymbol{x}_t - \boldsymbol{F}\boldsymbol{x}_{t-1}) \\ & - \frac{T}{2}\log|\boldsymbol{R}| - \frac{1}{2}\sum_{t=1}^{T}(\widetilde{\boldsymbol{y}}_t - \boldsymbol{H}_t\boldsymbol{x}_t)^T\boldsymbol{R}^{-1}(\widetilde{\boldsymbol{y}}_t - \boldsymbol{H}_t\boldsymbol{x}_t). \end{aligned} \tag{3.23}$$

In turn, we may approximate $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$ by substituting (3.23) into (3.21) yielding

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) \approx \widetilde{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) \triangleq \mathbb{E}_{p(\boldsymbol{x}_{0:T}|\boldsymbol{y}_{1:T};\boldsymbol{\theta}^{(m)})}\left(\widetilde{L}(\boldsymbol{\theta}; \boldsymbol{x}_{0:T}, \boldsymbol{y}_{1:T})\right). \tag{3.24}$$

Using the Taylor expansion of (3.16) to approximate the complete log likelihood in this manner, allows us to evaluate $\widetilde{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$; using a little bit of algebra, it may be shown that it is given by:

$$\begin{aligned} \widetilde{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) = & -\frac{1}{2}\log|\boldsymbol{\Sigma}_0| - \frac{T}{2}\log|\boldsymbol{Q}| - \frac{T}{2}\log|\boldsymbol{R}| \\ & - \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}_0^{-1}\left(\boldsymbol{P}_{0|T} + (\boldsymbol{m}_{0|T} - \boldsymbol{\mu}_0)(\boldsymbol{m}_{0|T} - \boldsymbol{\mu}_0)^T\right)\right) \\ & - \frac{1}{2}\operatorname{tr}\left(\boldsymbol{Q}^{-1}\left(\boldsymbol{C} - \boldsymbol{B}\boldsymbol{F}^T - \boldsymbol{F}\boldsymbol{B}^T + \boldsymbol{F}\boldsymbol{A}\boldsymbol{F}^T\right)\right) \\ & - \frac{1}{2}\operatorname{tr}\left(\boldsymbol{R}^{-1}\sum_{t=1}^{T}\left(\left(\widetilde{\boldsymbol{y}}_t - \boldsymbol{H}_t\boldsymbol{m}_{t|T}\right)\left(\widetilde{\boldsymbol{y}}_t - \boldsymbol{H}_t\boldsymbol{m}_{t|T}\right)^T + \boldsymbol{H}_t\boldsymbol{P}_{t|T}\boldsymbol{H}_t^T\right)\right), \end{aligned} \tag{3.25}$$

where the matrices $\boldsymbol{A}, \boldsymbol{B}$ and $\boldsymbol{C}$ are defined according to:

$$\begin{aligned} \boldsymbol{A} &\triangleq \sum_{t=1}^{T}\mathbb{E}\left(\boldsymbol{x}_{t-1|T}\boldsymbol{x}_{t-1,T}^T; \boldsymbol{\theta}^{(m)}\right) = \sum_{t=1}^{T}\left(\boldsymbol{P}_{t-1|T} + \boldsymbol{m}_{t-1|T}\boldsymbol{m}_{t-1|T}^T\right) \\ \boldsymbol{B} &\triangleq \sum_{t=1}^{T}\mathbb{E}\left(\boldsymbol{x}_{t|T}\boldsymbol{x}_{t-1|T}^T; \boldsymbol{\theta}^{(m)}\right) = \sum_{t=1}^{T}\left(\boldsymbol{P}_{t,t-1|T} + \boldsymbol{m}_{t|T}\boldsymbol{m}_{t-1|T}^T\right) \\ \boldsymbol{C} &\triangleq \sum_{t=1}^{T}\mathbb{E}\left(\boldsymbol{x}_{t|T}\boldsymbol{x}_{t|T}^T; \boldsymbol{\theta}^{(m)}\right) = \sum_{t=1}^{T}\left(\boldsymbol{P}_{t|T} + \boldsymbol{m}_{t|T}\boldsymbol{m}_{t|T}^T\right), \end{aligned}$$

and can be obtained by applying the EKS of Algorithm 3.3.

Thus, the MMSE estimate of $\boldsymbol{x}_{0:T}$ conditioned on $\boldsymbol{y}_{1:T}$ and $\boldsymbol{\theta}^{(m)}$ will be produced at the mth iteration of the EM algorithm during its E-step. The $(m+1)$th estimates of $\boldsymbol{\theta}$

are obtained next, by setting the derivative of $\widetilde{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$ with respect to $\boldsymbol{\theta}$ equal to zero and solving. This leads to the following estimates:

$$\begin{aligned}
\widehat{\boldsymbol{F}}^{(m+1)} &= \boldsymbol{B}\boldsymbol{A}^{-1} \qquad \widehat{\boldsymbol{Q}}^{(m+1)} = \frac{1}{T}\left(\boldsymbol{C} - \boldsymbol{B}\boldsymbol{A}^{\mathbf{-1}}\boldsymbol{B}^T\right) \\
\widehat{\boldsymbol{R}}^{(m+1)} &= \sum_{t=1}^{T}\left(\widetilde{\boldsymbol{y}}_t\widetilde{\boldsymbol{y}}_t^T - \boldsymbol{H}_t\boldsymbol{m}_{t|T}\widetilde{\boldsymbol{y}}_t^T - \widetilde{\boldsymbol{y}}_t(\boldsymbol{H}_t\boldsymbol{m}_{t|T})^T + \boldsymbol{H}_t\left(\boldsymbol{P}_{t|T} + \boldsymbol{m}_{t|T}\boldsymbol{m}_{t|T}^T\right)\boldsymbol{H}_t^T\right) \qquad (3.26) \\
\widehat{\boldsymbol{\mu}_{\mathbf{0}}}^{(m+1)} &= \boldsymbol{m}_{0|T} \qquad \widehat{\boldsymbol{\Sigma}_{\mathbf{0}}}^{(m+1)} = \boldsymbol{P}_{0|T},
\end{aligned}$$

The estimators in (3.26) are unconstrained, but it may be useful to derive constrained estimators in this framework. Recall that in our formant-tracking setting, the dimensionality of the observation vector $\boldsymbol{y}_t$ at frame time index t is equal to the number of cepstral coefficients N, which is often on the order of $12 < N < 20$ in many cases. Consequently, the dimensionality of the covariance matrix $\boldsymbol{R}$ is large relative to the number of frames (e.g., $\boldsymbol{R} \in \mathbb{R}^{15\times 15}$ and $T = 100$ frames). Thus, it is important to constrain the structure of $\boldsymbol{R}$ and regularize the resultant estimator.

Let $\widehat{\boldsymbol{R}}_{\text{unc}}$ denote the unconstrained estimator of $\boldsymbol{R}$ in (3.26), and suppose we now wish to derive an estimator for $\boldsymbol{R}$ under the assumption that it is a constant multiple of a known covariance matrix $\boldsymbol{R}_0$ as in:

$$\boldsymbol{R}_{\text{con}} = c\boldsymbol{R}_0.$$

Natural choices for $\boldsymbol{R}_0$ include the identity matrix and the diagonal matrix whose nth entry is equal $1/n$; both will be used in the experiments below. It is not too hard to show that

$$\widehat{\boldsymbol{R}_{\text{con}}} = N^{-1}\,\text{tr}\left(\widehat{\boldsymbol{R}}_{\text{unc}}\boldsymbol{R}_0^{-1}\right)\boldsymbol{R}_0 \qquad (3.27)$$

Similar constraints may be imposed on the matrices $\boldsymbol{Q}$ and $\boldsymbol{F}$. For instance, an estimator for $\boldsymbol{F}$ under a set of linear constraints on its entries is described in [89].

3.6 Speech Activity Detection: Censoring the Likelihood

The state-space model of (3.5) does not explicitly take into account the uncertainty of speech presence. Hence, approaches based on this model as it stands (along with others previously cited) are liable to suffer significant performance degradations—not only during pauses in utterances, but also whenever formant peaks cease to be observable in the waveform (i.e., are censored in the likelihood of (3.5)). Though many authors discuss the problem of formant disappearance/observability—with some even making the distinction that a formant is an "observed" vocal tract resonance [65]—this effect is most strongly pronounced in examining the trajectories in the recently compiled database of vocal tract resonances [78], in which much of the hand-labeling effort was devoted to correcting mistakes made by automated procedures during non-speech regions.

To address this problem, we censor the likelihood equation in (3.5) by augmenting the state vector $\boldsymbol{x}_t$ with a binary indicator variable for each formant and anti-formant

frequency/bandwidth pair. We model these indicators as statistically independent from frame to frame, and we assume that in each frame, they are estimated by a voice activity detector. The MMSE state estimate may then be obtained in closed form by modifying the update equation in Algorithm 3.1 with the addition of a left multiplication by the matrix $\boldsymbol{M}_t \in \mathbb{R}^{(p+q)\times(p+q)}$—a diagonal mask matrix with zeros in entries associated to the unobservable variables and ones in all other entries. The update equation becomes:

$$\boldsymbol{m}_{t|t} = \boldsymbol{m}_{t|t-1} + \boldsymbol{M}_t \boldsymbol{K}_t (\boldsymbol{y}_t - h(\boldsymbol{m}_{t|t-1})).$$

As an example, suppose that we are tracking three resonances ($p = 6$, $q = 0$), and that a speech activity detector estimates that in frame t_0 there is no energy in a subband typically containing the second formant (e.g., $1 - 2.5$ kHz). In this case, we may simply want to censor the likelihood in this band and "coast" the parameters of the second formant, with the main diagonal of the $\boldsymbol{M}_{t_0}$ given by $\begin{pmatrix} 1 & 0 & 1 & 1 & 0 & 1 \end{pmatrix}$. On the other hand, if there were no speech present in any band, all the entries of the associated mask matrix would be set to zero.

This approach avoids the ill-posed problem of estimating VTR parameters in non-speech regions of the time-frequency plane, which is not currently standard in many existing algorithms. Propagating the VTR trajectories in this manner, which is akin to "coasting" the state estimate, allows for principled estimation not only when speech is absent, but also when it resumes. Moreover, the error variance grows (as it should) during silent regions, thereby indication increasing uncertainty regarding VTR locations.

3.7 Experiments

In this section we describe a number of experiments designed to test various aspects of the state-space model of (3.2). First in Sections 3.7.1 and 3.7.2 we illustrate the performance of the EM algorithm of Section 3.5.2 for system identification using synthetic waveforms and study the efficacy of bandwidth estimation. Next, in Section 3.7.3, we use a recently introduced database of hand-corrected speech utterances to evaluate the proposed formant tracking approaches, and compare their performance to that of WaveSurfer—a popular formant-tracking tool [49] among practitioners. Finally, in Section 3.7.4 we use synthetic and real data examples in order to illustrate simultaneous tracking of formants and anti-formants.

3.7.1 Synthetic Data Experiments

We first validated the inferential procedures of Section 3.4 using synthetic data generated according to the state-space model of (3.5). In the first example, we illustrate the performance of the online system identification approach of Algorithm 3.3 using a 100-sample synthetic waveform. Four formants ($p = 8$) and $N = 15$ cepstral coefficients were used; the diagonal matrices $\boldsymbol{F}$ and $\boldsymbol{Q}$ were set according to $\boldsymbol{F} = \boldsymbol{I}_p$, $\boldsymbol{Q} = 900\boldsymbol{I}_p$, and the observation covariance was set via $\boldsymbol{R} = .0525\boldsymbol{I}_N$ so that the resultant signal-to-noise ratio was 15 dB. The extended Kalman smoother of Algorithm 3.1 was applied to the data using the exact values of $\boldsymbol{F}$ and $\boldsymbol{Q}$, but the observation covariance was purposely set to $10\boldsymbol{R}$

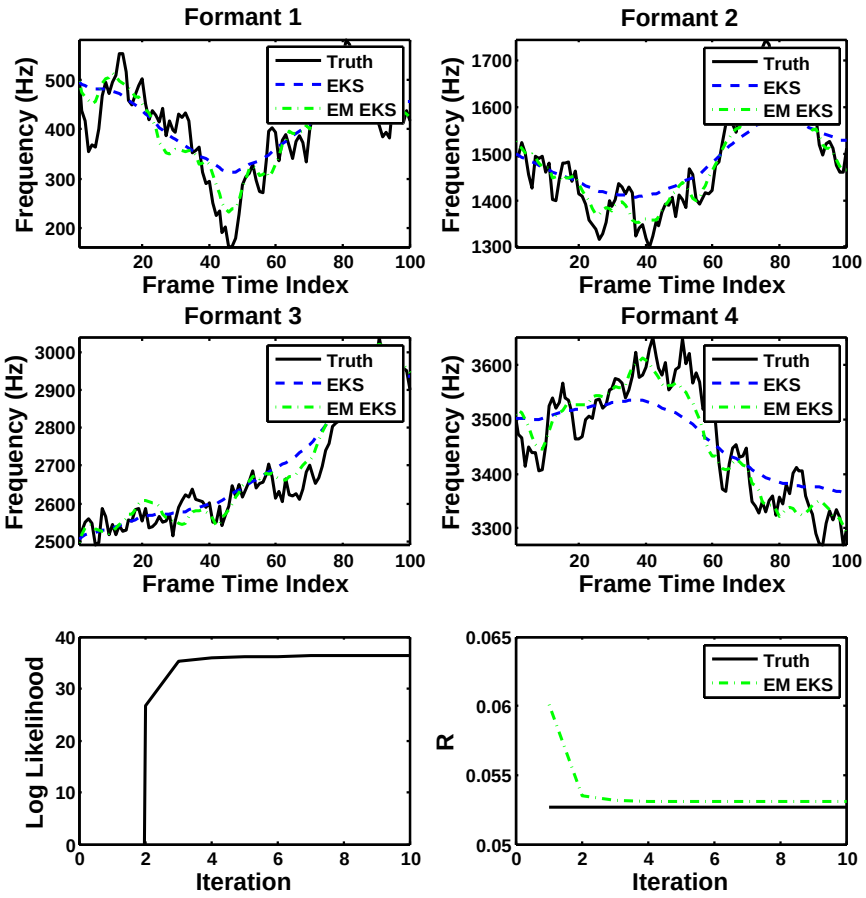


Figure 3.4: Formant tracking using a synthetic waveform with system identification via the EM algorithm. True formant frequencies (solid black lines) are overlaid with estimates obtained by the EKS of Algorithm 3.1 using an inflated $\boldsymbol{R}$ (dashed blue lines) and by the EM-based EKS of Algorithm 3.3 (dash-dotted green lines). The increasing log-likelihood of the model parameters and the improving estimates of $\boldsymbol{R}$ are shown as a function of ten EM iterations in the bottom- left and right panels, respectively.

in order to induce model mismatch. In this setting, we consider the ability of the online adaptation approach in algorithm Algorithm 3.3 to recover from the incorrect initialization. The results of applying the Kalman smoother to the synthetic waveform, with and without EM iterations, are shown in Figure 3.4.

When the inflated observation covariance matrix $\boldsymbol{R}$ is used, the resultant formant trajectory estimates are clearly oversmoothed and essentially track a long-term average of the formant frequency values. On the other hand, after a few iterations the EM algorithm converges to an improved estimate of the observation covariance matrix $\boldsymbol{R}$ resulting in higher-quality formant tracks. Indeed, the log-likelihood of the model parameters increases from -810 (not shown) to 25 (first point shown in the bottom-left panel of Figure 3.4) after the first iteration, and increases monotonically over subsequent iterations showing the efficacy of the EM approach.

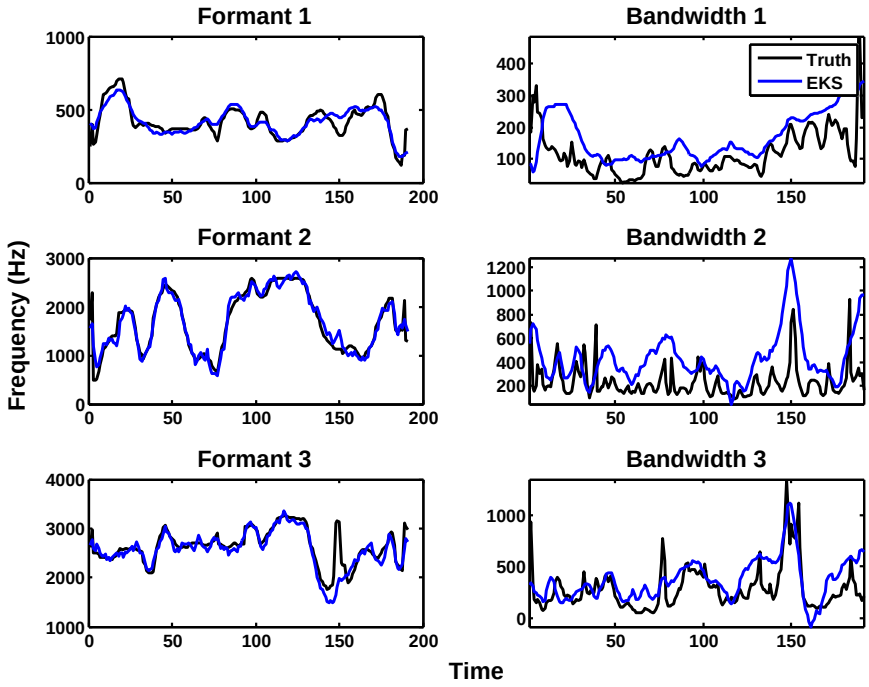


Figure 3.5: Formant frequency and bandwidth tracking using synthetic data generated according to the state-space model (3.5). True formant frequencies for the first three formants (left panels, black) are overlaid with frequencies estimated by the extended Kalman smoother (blue). The associated bandwidth values (right panels, black) are also overlaid with the corresponding estimates.)

Gradually moving away from purely synthetic data toward a more realistic setting, we obtained formant-frequency and bandwidth estimates from an all-voiced speech utterance "Why were you away a year Roy" using WaveSurfer [49], and used these values to synthesize LPC cepstra via (3.4). After adding noise to these observations, we applied the extended Kalman smoother of Algorithm (3.1) to retrack the formant frequencies and bandwidths. The results are shown in Figure 3.5 and illustrate excellent tracking of the formant frequencies. However, the tracking of the associated bandwidth values is slightly worse—we explore this phenomenon in more depth next.

3.7.2 Bandwidth Estimation

In the formant tracking example of Figure 3.5 we have observed the bandwidth estimates to be less accurate that the associated frequency estimates. In the speech analysis literature, it has been similarly observed that obtaining accurate formant bandwidth estimates is difficult [84], and therefore practitioners sometimes use *apriori* fixed bandwidth values in lieu of their estimates [85]. In this section, we quantify and confirm the relative difficulty of estimating formant bandwidths, but at the same time reaffirm the importance of their estimation in the context of formant tracking.

The perceived difficulty of bandwidth estimation stems from the fact that the variance of the bandwidth estimator is significantly larger than that of the formant frequency

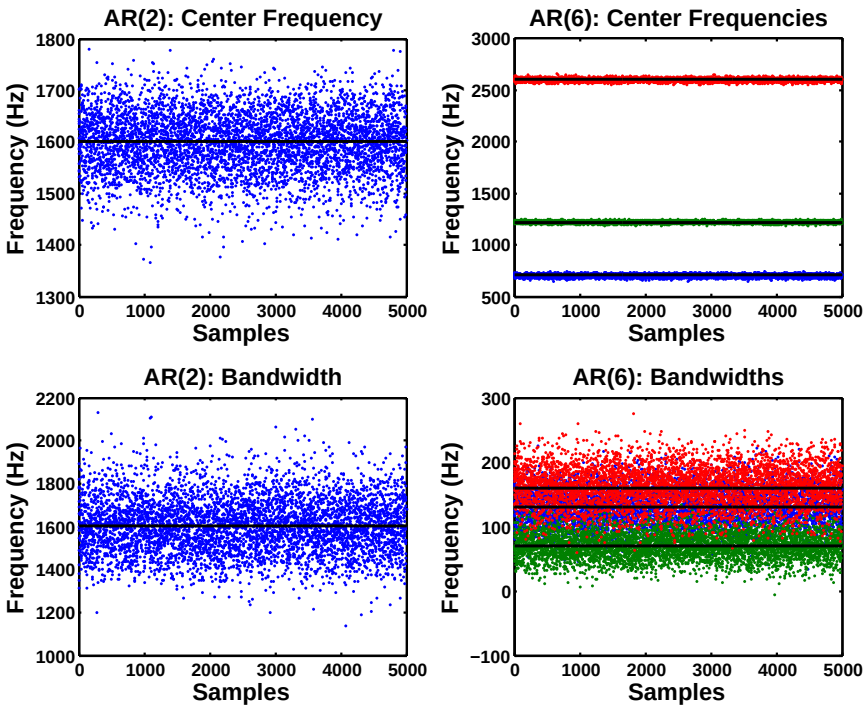


Figure 3.6: Variance of formant frequency and bandwidth estimators for AR(2) and AR(6) models. Scatter plots of frequency/bandwidth estimates for 5000 Monte Carlo instantiations of AR(2) (left) and AR(6) (right) processes are overlaid with true values (black lines), and show that the variance of bandwidth estimates is higher than that of the associated frequency estimates.

estimator. To illustrate this, we estimated the parameters of an 1000-sample formant-like AR(2) waveform obtained by shaping white Gaussian noise by a second-order digital resonator whose center frequency and bandwidth are both equal to 1.6 kHz. For each waveform so obtained, the AR parameters were estimated using the covariance method of linear prediction and the resultant coefficients were transformed to frequency/bandwidth pairs. The scatter plot of the resultant estimates for 5000 Monte Carlo trials is shown together with the true parameter values in the left panels of Figure 3.6. The standard deviations for frequency and bandwidth estimates are 57.66 Hz and 126.99 Hz, respectively, confirming the "difficulty" of estimating bandwidth values as accurately as formant-frequency values. We observed similar findings for a broad range of synthesized and real signals including a 1000-sample AR(6) waveform as shown in the right panels of Figure 3.6—standard deviations for frequency and bandwidth estimates over 5000 Monte Carlo trials are (13.24, 9.70,14.51) Hz and (27.16, 19.63, 29.68) Hz, respectively. These results are not surprising since the Cramer-Rao bounds for AR parameter estimation [4] indicate that the lower bounds on estimator variances are different for each coefficient. These differences are amplified further by the nonlinear transformation of the AR coefficients into frequency/bandwidth pairs.

Even though bandwidth estimates tend to have larger variance and consequently larger mean-squared error than the associated frequency estimates, they nonetheless play an important role in interpreting formant tracker output. Consider, for instance, the voiced

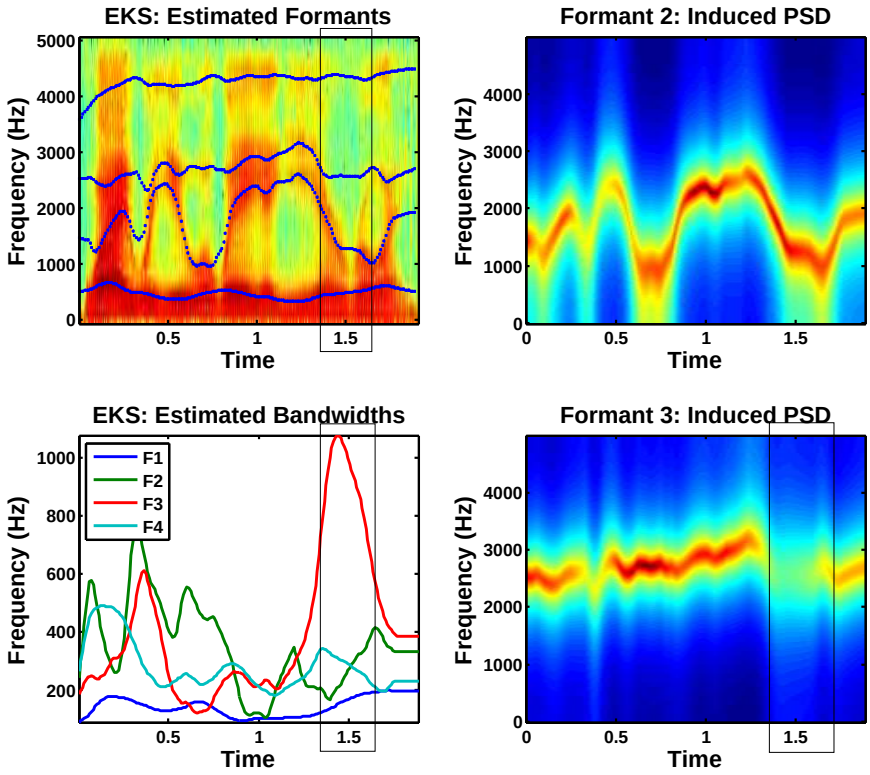


Figure 3.7: Estimated formant tracks for the all-voiced sentence: "Why were you away a year Roy?" Formant frequency estimates obtained via Algorithm 3.1 are overlaid on the spectrogram (16 ms Hamming windows, 50% overlap) of the utterance (top-left panel); the associated bandwidth estimates are also shown (bottom-left panel). Power spectral densities of second-order resonators induced by the second and third frequency/bandwidth pairs are shown in the top- and bottom- right panels, respectively.

waveform "Why were you away a year Roy" whose spectrogram is shown in the top-left panel of Figure 3.7 overlaid with formant frequency estimates obtained by Algorithm 3.1 using the offline initialization procedure described in Section 3.5.1; the associated bandwidth estimates are shown in the bottom-left panel of Figure 3.7. Note the area highlighted by the black rectangle: the estimated third formant frequency covers a low energy region in the time-frequency plane, and it seems that the formant tracker has made an error. However, observe that the value of the third formant bandwidth in the same region spikes, consequently the associated formant peak is lower and has less energy. This effect is illustrated in the bottom-right panel of Figure 3.7 where the power spectrum associated to the sequence of frequency and bandwidth values of the third formant is shown. In contrast, no such behavior is observed in the power spectrum induced by the second formant track as shown in Figure 3.7 for reference.

3.7.3 VTR Database Experiments

In this section, we evaluate the performance of the proposed formant tracking approaches using the recently introduced VTR database of [78], which contains a representative subset of the TIMIT speech corpus [48] consisting of diverse, phonetically-balanced utterances collated across a range of gender, individual speakers, dialects, and phonetic contexts. The VTR database contains state information consisting of four formants and their bandwidths for each frame of analyzed speech; only the first three formants were manually corrected.

Our experimental setup can be described in three parts as follows. First, we extracted all 516 VTR utterances[6] and resampled the corresponding TIMIT waveform data from 16 kHz to 7 kHz in order to track the *first three* formants. Second, we performed careful pre-processing to evaluate the WaveSurfer formant tracking algorithm [49] performance against the VTR "ground truth." All analysis parameters were matched to those used to generate the VTR database: 20 ms Hamming windows with an overlap of 50% were employed, left-aligned with the first sample of each TIMIT utterance; an LPC model order of 12 was used in conjunction with a pre-emphasis coefficient of 0.7, and the first three formants were tracked.

Three formants ($p = 6$) and zero anti-formants ($q = 0$) were used in the state-space model of (3.5). In each 20 ms short-time speech segment 12 AR coefficients were estimated using the autocorrelation method of linear prediction and, subsequently, transformed to $N = 12$ LPC cepstral coefficients via the recursion of (3.13). The state-space model parameters $\boldsymbol{F}, \boldsymbol{Q}, \boldsymbol{R}$ were estimated using either the offline or online (EM) approaches described in Section 3.5. In the latter case, three EM iterations were used—more iterations did not appreciably change the reported results. The initial state estimates were set to $\boldsymbol{x}_0 = \begin{pmatrix} 500 & 1500 & 2500 \end{pmatrix}^T$ to match the processing done for the VTR database, and $\boldsymbol{\Sigma}_0$ was set to $\boldsymbol{Q}$. Due to the fact that no "ground truth" of VTR bandwidths exists, we set the bandwidth values to the average values as reported by WaveSurfer for the duration of the utterance when using the offline parameter estimation approach. When the EM algorithm is used for parameter estimation, the bandwidths were all set to 250 Hz. Finally, for each utterance, a simple energy-based voice activity detector was used to identify non-speech frames as those whose energy falls within the lower 15th percentile.

We begin by examining two specific cases in detail; subsequently, we summarize results for the Kalman filtering approach via root-mean-squared error (RMSE) in Hertz per formant, averaged over all VTR database utterances. Figures 3.8 and 3.9 show two waveforms, selected after paging through approximately 5% of the VTR database entries according to a random sequence. For each utterance we report RMSE, averaged over the first three formants with and without conditioning on speech presence; regions of speech absence are denoted by darkened segments of the estimated trajectories, which are shown in white.

Figure 3.8 indicates a marked improvement in tracking capability relative to the performance of WaveSurfer, both graphically and in terms of absolute RMSE, regardless

[6] Reference [78] lists 538 entries, but 22 of these are text-file annotations, leaving a total of 516 TIMIT utterances.

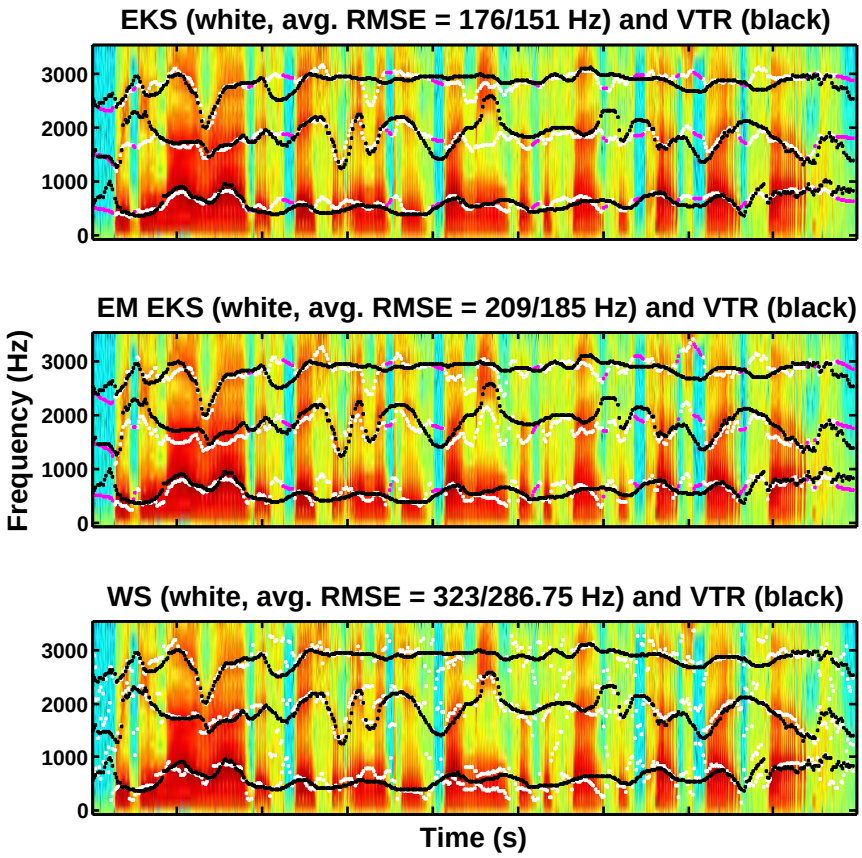


Figure 3.8: Estimated formant tracks for utterance 23: "We do not arrive at spatial images by means of the sense of touch by itself." The VTR trajectories are shown in black, with the output of the extended Kalman smoother using offline (top) and online (middle) system identification methods, and WaveSurfer (bottom), all shown in white. Regions of speech absence are denoted by darkened segments of the white estimated trajectories. Reported RMSE is averaged over 3 formants: entire utterance / conditioned on speech presence.

of whether or not the error is calculated only over regions of speech presence. Figure 3.9 shows a second example that, while ostensibly demonstrating the improvement in tracking performance relative to WaveSurfer, in fact, alerts us that the "ground truth" formant frequency values in the VTR database should be interpreted with caution. In particular, it is obvious from the first portion of the spectrogram that the formant tracks do not always cross high-energy regions—in fact the sentence has not yet begun, but "formant" values are reported for the aspiration energy at the beginning of the utterance. Generally, since the values in the VTR database were generated with significant manual intervention, there are inevitably a variety of labeling errors throughout.

Despite the presence of various labeling errors in the VTR database, it is still useful to obtain a rough idea of performance by evaluating different formant tracking algorithms across all 516 labeled utterances. Table 3.1 summarizes the average RMSE reduction per formant obtained by the proposed approaches relative to WaveSurfer. We report results

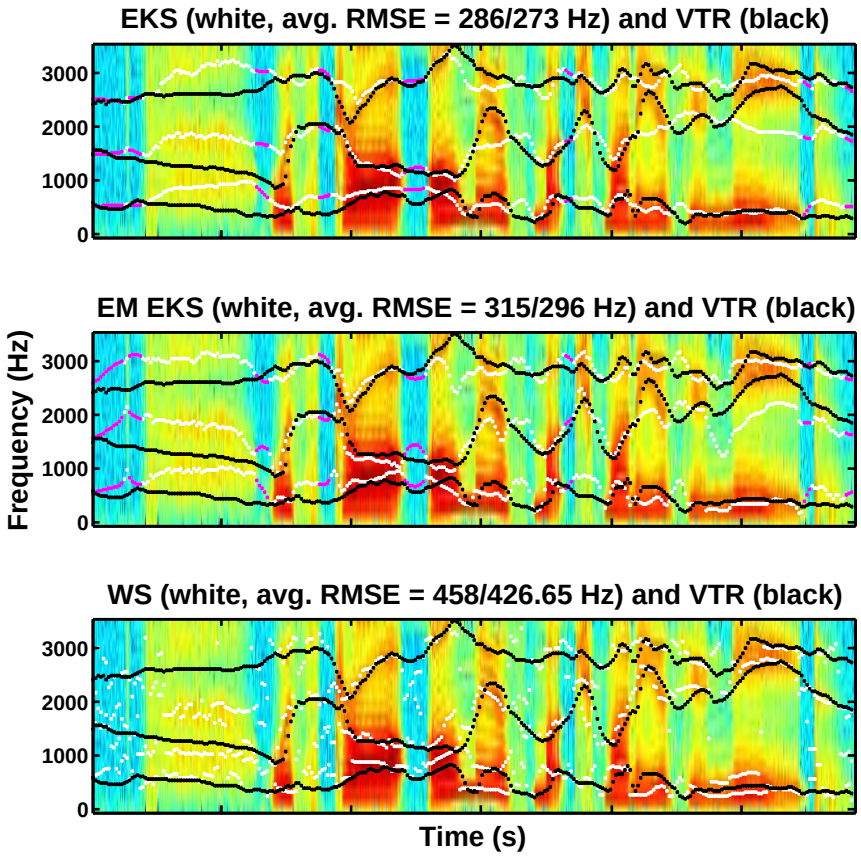


Figure 3.9: Estimated formant tracks for TIMIT utterance 200: "Withdraw only as much money as you need." The VTR trajectories are shown in black, with the output of the extended Kalman smoother using offline (top) and online (middle) system identification methods, and WaveSurfer (bottom), all shown in white. Regions of speech absence are denoted by darkened segments of the white estimated trajectories. Reported RMSE is averaged over 3 formants: entire utterance / conditioned on speech presence.

when offline and online system identification methods described in Section 3.5 are used in conjunction with the approach to speech activity detection presented in Section 3.6.

We can see that the approaches based on the state-space model of (3.5) result in significantly improved tracking of the first and second formants, and slightly improved tracking of the third formant. The value of censoring the likelihood during regions of speech absence is particularly evident. Both parameter estimation methods yield performance gains, with the offline approach yielding slightly better results.

We have also found that the capability of automated methods, including our approaches and WaveSurfer, to track the third formant strongly depends on the frequency to which the waveforms were resampled. This variable controls the amount of energy in the spectrum at higher frequencies. Thus, if the signal were resampled to 10kHz, for instance, the third formant may begin to track energy typically ascribed to the fourth formant leading to large errors when compared to manually-labeled VTR data. To demonstrate this

	Root Mean Square Error Reduction			
Formant	Offline System ID		Online System ID	
Number	All Frames	Speech Frames	All Frames	Speech Frames
1	23.74% (43.69 Hz)	20.81% (31.58 Hz)	16.18% (30.20 Hz)	11.79% (17.88 Hz)
2	28.92% (90.47 Hz)	23.54% (63.38 Hz)	22.64% (70.80 Hz)	17.43% (46.94 Hz)
3	12.27% (41.85 Hz)	4.38% (12.83 Hz)	7.83% (26.69 Hz)	-0.05% (-0.16 Hz)

Table 3.1: Reduction in root-mean-squared error relative to WaveSurfer, taking the VTR database [78] as ground truth. Errors are computed both for the offline (left) and online (right) system identification approaches and both for the entire utterance (1st and 3rd columns) and conditioned on speech presence (2nd and 4th columns).

	Root Mean Square Error Reduction			
Formant	Offline System ID		Online System ID	
Number	All Frames	Speech Frames	All Frames	Speech Frames
1	25.46% (47.52 Hz)	23.86% (35.11 Hz)	18.76% (35.02 Hz)	15.10% (22.85 Hz)
2	31.41% (98.25 Hz)	27.30% (73.47 Hz)	22.04% (68.95 Hz)	17.84% (48.00 Hz)
3	35.74% (89.21 Hz)	31.77% (93.08 Hz)	26.84% (91.52 Hz)	23.80% (69.72 Hz)

Table 3.2: Reduction in root-mean-squared error relative to WaveSurfer, with adaptive waveform resampling, taking the VTR database [78] as ground truth. Errors are computed both for the offline (left) and online (right) system identification approaches and both for the entire utterance (1st and 3rd columns) and conditioned on speech presence (2nd and 4th columns).

dependence, we have repeated the experiment used to construct Table 3.2, but instead of resampling each waveform to 7 kHz, we used an adaptive resampling scheme *for the proposed methods only*. For each utterance we looked up the maximum value that the third formant takes on in the VTR database, rounded it up to the nearest multiple of 250Hz, and then resampled the waveform to twice this value providing, in essence, an "oracle" value for the cutoff frequency. The resultant average RMSE reductions relative to WaveSurfer are reported in Table 3.2, and show a significant reduction in tracking error of the third formant. On the other hand, the tracking performance for the first two formants is essentially unchanged. These results suggest that incorporating adaptive resampling strategies into automated formant trackers may lead to improved performance.

3.7.4 Formant and Anti-Formant Tracking

In the last set of experiments, we illustrate how to use the state-space model of (3.5) in order to simultaneously track vocal tract resonances and anti-resonances (i.e,. formants and anti-formants). In the first example a 500 ms synthetic waveform was obtained by shaping white Gaussian noise with an ARMA$(4,2)$ filter defined by two complex-conjugate pole pairs and one complex-conjugate zero pair—the associated formant (bandwidth) and anti-formant (bandwidth) frequencies were $500(100), 1500(100)$ and $2500(100)$, respectively. An AR(4) model was fitted to the data using the covariance method of linear prediction and an ARMA$(4,2)$ model was fitted using the iterative procedure described in [91, Section 10.2] and implemented in the MATLAB System Identification toolbox as the function "ar-

max". The resultant estimates of the all-pole and all-zero spectra are shown in the top-left and top-right panels of Figure 3.11, respectively, and are overlaid with the true (induced) spectrum (black lines). This highlights not only the obvious ability of the ARMA estimator to capture a spectral valley due to the introduction of spectral zeros, but also to improve the associated pole estimates.

Next, we induced a sequence of short-time segments by tiling the waveform by 20 ms Hamming windows with 50% overlap. In each segment, an ARMA$(4, 2)$ model was fitted using the iterative method of [91, Section 10.2] and 10 cepstral coefficients were obtained via the recursion of (3.2). The extended Kalman smoother of Algorithm 3.1 was used to track two vocal tract resonances ($p = 4$) and one antiresonance ($q = 2$); the bandwidths were all fixed to their true values, the matrices $\boldsymbol{F}, \boldsymbol{Q}$ were set to $\boldsymbol{F} = \boldsymbol{I}_{(p+q)/2}, \boldsymbol{Q} = 1e6\boldsymbol{I}_{(p+q)/2}$ and the observation covariance matrix $\boldsymbol{R}$ was set to be diagonal with the nth diagonal entry equal to $1/n$. Finally, the formant and anti-formant frequencies were initialized to $(800, 2000)$ Hz and 2300 Hz, respectively. The tracking results are shown in the bottom panel of Figure 3.10 and indicate that, despite the initial state bias, Algorithm 3.1 is able to accurately track the formant/anti-formant frequencies.

In order to demonstrate simultaneous tracking of formants and anti-formants using real speech, recall that an *all-pole* transfer function describes articulatory configurations that comprise exactly one direct acoustic path between the source and output positions. Any configuration comprising multiple acoustic paths results in the possibility of both poles and zeros in the transfer function [3, 64], an effect quite noticeable in nasals (e.g., [m], [n], [ŋ] as "bam," "ban," and "bang") as the coupling with the nasal cavity introduces two new paths to the outside air via the nostrils.

Consequently, in our second example, we use the above methodology to analyze a sustained 500 ms alveolar nasal [n] (as in "ban") produced by an adult male speaker, recorded at 16 kHz, and downsampled to 8 kHz prior to subsequent processing; a short segment of the recording is shown in the top-left panel of Figure 3.11. Three estimates of the waveform power spectrum, based on the periodogram, AR(16) and ARMA$(16, 4)$ models, are shown in the top-right panel of Figure 3.11. It is clear that by introducing a zero at ~ 2950 Hz, dramatically improves the spectral estimate relative to the all-pole spectral estimator.

Next, as in the case of the synthetic example, we induced a sequence of short-time segments by tiling the waveform by 20 ms Hamming windows with 50% overlap. In each segment, an ARMA$(16, 4)$ model was fitted using the iterative method of [91, Section 10.2] and 20 cepstral coefficients were obtained via the recursion of (3.2). Algorithm 3.1 was used to track four vocal tract resonances ($p = 8$) and one antiresonance ($q = 2$); the bandwidths were all fixed at 200 Hz, the matrices $\boldsymbol{F}, \boldsymbol{Q}$ were set to $\boldsymbol{F} = \boldsymbol{I}_{(p+q)/2}, \boldsymbol{Q} = 1e6\boldsymbol{I}_{(p+q)/2}$ and the observation covariance matrix $\boldsymbol{R}$ was set to be diagonal with the nth diagonal entry equal to $1/n$. Finally, the formant and anti-formant frequencies were initialized to $(500, 1000, 2500, 3500)$ Hz and 3500 Hz, respectively. The tracking results are shown in the bottom panel of Figure 3.11 and show that, despite the bias in the initial state, Algorithm 3.1 is able to accurately track the frequencies of the resonances and antiresonance in the nasal.

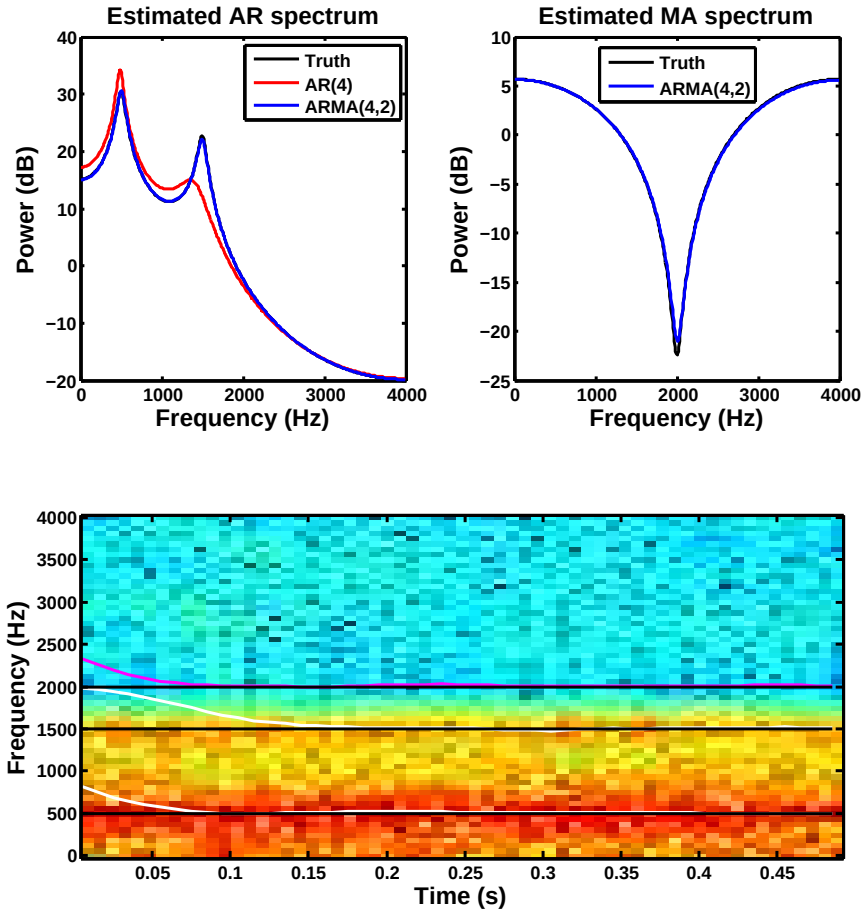


Figure 3.10: Tracking two vocal tract resonances and one antiresonance in a synthetic utterance obtained by shaping white Gaussian noise using an ARMA$(4,2)$ filter. The induced AR(4) power spectrum is shown overlaid with two estimates obtained using fitting procedures based on the AR(4) and ARMA$(4,2)$ models (top-left panel). The spectral estimate of the moving average component is overlaid onto the induced MA(2) power spectrum (top-right panel). Estimated formant (white) and anti-formant (magenta) tracks are overlaid on a spectrogram (bottom panel) computed using 8ms Hamming windows with 50% overlap; true center-frequency values are also shown (black).

3.8 Summary

In this chapter, we have considered the problem of estimating vocal tract resonance and antiresonance trajectories across short-time segments of an observed acoustic waveform using a state-space modeling framework. Our approach extends earlier contributions of Deng et. al., [1, 2] who were first to use the formant-to-cepstrum mapping in the context of formant tracking. We have extended the framework of [1, 2] to enable the estimation of vocal tract anti-resonance trajectories, and proposed offline and online system identification algorithms that allow for explicit modeling of correlation structure across formants and anti-formants. In addition, the proposed model accounts for the absence of waveform energy during pauses and silences, by way of a censored likelihood formulation,

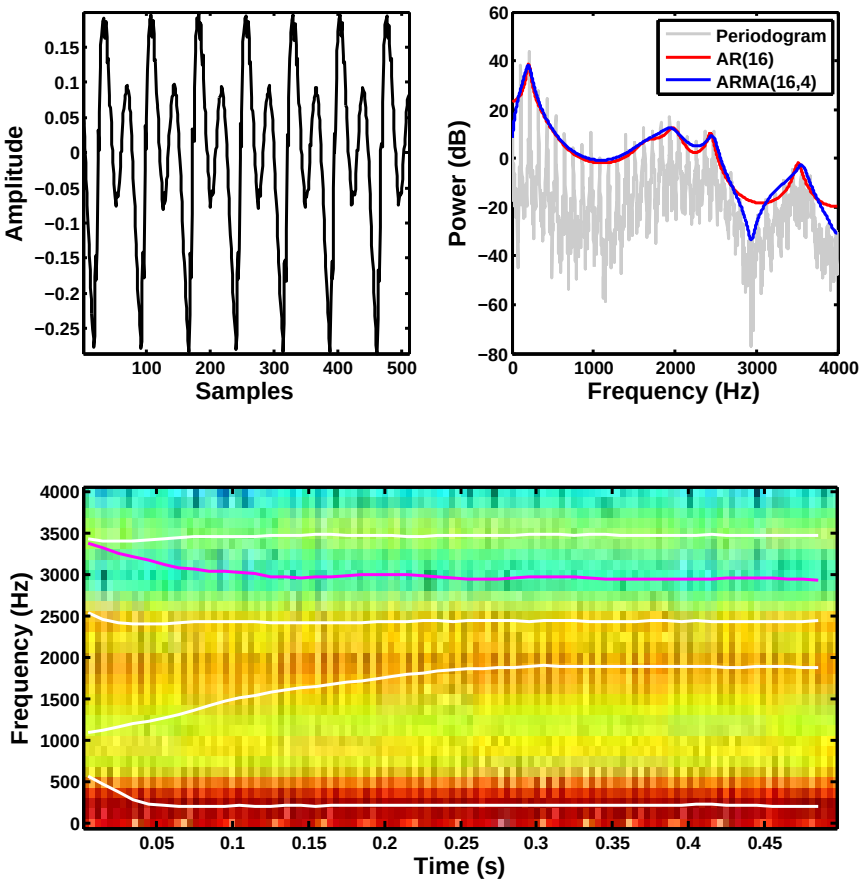


Figure 3.11: Tracking four formants and one anti-formant in the alveolar nasal [n] as is "bane". A waveform segment is shown (top-left panel) together with estimates of its power spectrum (top right) obtained via the periodogram, AR(16), and ARMA(16, 4) estimators. Estimated formant (white) and anti-formant (magenta) tracks are overlaid on a spectrogram (bottom) computed using 8ms Hamming windows with 50% overlap.

adopted a Taylor-based linearization of the formant-to-cepstrum map, and evaluated the resultant EKS-based tracking algorithm using particle filtering. Among a number of illustrations using synthetic and natural speech, we quantified the well-known difficulty of estimating formant bandwidths, evaluated tracking performance using a public database of hand-labeled formant trajectories, and provided first-of-a kind examples for simultaneous formant and anti-formant tracking within a single utterance.

Chapter 4

Time-Varying Autoregressive Models

Our development of formant tracking algorithms in Chapter 3 relied on the standard assumption that the speech signal is piecewise-stationary at a fixed (15–30 ms) time scale. The goal of the current and the next two chapters is to generalize and develop, in the absence of this constraint, two traditional speech processing methods: autoregressive modeling and short-time Fourier analysis. We show that the resultant models lead to more accurate signal representations and improved algorithms for a variety of applications.

In this chapter, we begin by further developing the theory of time-varying linear prediction by studying time-varying autoregressive (TVAR) processes. The application of TVAR models for representing temporal variation of the vocal tract is natural and has a long history in speech analysis [58, 92–102], since it is easily motivated as representing the time-varying digital filter (vocal tract) in the source-filter view of speech production [3]. However, TVAR models have also been applied to a wide variety of problems outside the speech community including blind source separation [101, 103–106], radar signal processing [107, 108], biomedical signal analysis [109–111] and instantaneous frequency estimation [112–114], and the methodological contributions of this chapter are highly relevant in these domains as well.

We begin in Section 4.1 by reviewing the well-known class of TVAR models whose coefficient trajectories are modeled using flexible basis function expansions. In Section 4.2 we derive a closed-form expression for the covariance structure of a general TVAR process, show how to apply the results to computing Cramér-Rao lower bounds, and obtain a new way of visualizing time-frequency content. In Section 4.3, we derive a new estimator of the TVAR coefficients in the case when only noisy observations are available. In Section 4.4, we develop a time-varying lattice formulation that enables us to constrain the temporal evolution of the parameters so that the "instantaneous" poles of the estimated process remain inside the unit circle, which leads to frozen-time stable inverse-prediction-error autoregressive filters. Finally, two new approaches for modeling TVAR coefficient trajectories are introduced. The first, which blends the functional-expansion approach with a stochastic evolution model, is discussed in Section 4.5, while a geometric approach based on viewing each TVAR process as a path on the manifold of AR processes, with estimators realized via

convex optimization, is developed in Section 4.6.

4.1 TVAR Modeling: Function-Expansion Approach

4.1.1 Model Specification

Here we study time-varying autoregressive (TVAR) processes defined according to the following discrete-time difference equation with time-varying coefficients for $n \in \mathbb{Z}$:

$$\text{TVAR}(p): \quad x[n] = \sum_{i=1}^{p} a_i[n]x[n-i] + \sigma w[n], \tag{4.1}$$

which generalizes the classical autoregressive (AR) process, and where $w[n]$ is a zero-mean white Gaussian sequence with unit variance scaled by a gain parameter $\sigma > 0$. The time-dependence of the coefficients $a_i[n]$ implies that (4.1) is a nonstationary stochastic process. The process of (4.1) is defined over a doubly-infinite set of time indices (i.e., for all $n \in \mathbb{Z}$), however, when considering estimation problems from a finite-sequence of observations it will often be necessary to incorporate a boundary condition at the origin.

Note that the time indices of the TVAR coefficients in (4.1) are aligned with the output and are shifted away from the indices of the process samples which they multiply. In this case, the TVAR process of (4.1) is said to be in *shifted* form [57]. An alternate approach is instead to align the time index of each TVAR coefficient with that of its associated lag term leading to the *synchronous* form [57]:

$$x[n] = \sum_{i=1}^{p} a_i[n-i]x[n-i] + \sigma w[n]. \tag{4.2}$$

Since the differences between the shifted and synchronous forms are minor in practice, we will work primarily with the shifted form of (4.1). However, we will use the synchronous form of (4.2) in order generalize the autocorrelation method of linear prediction to the time-varying setting in Section 4.3.2.

In order to complete the specification of the TVAR model of (4.1) and its variants, it is necessary to detail precisely how the linear prediction coefficients evolve in time. Unless otherwise specified, we choose to expand them in a set of $q+1$ basis functions $f_j[n]$ weighted by coefficients α_{ij} as follows:

$$a_i[n] = \sum_{j=0}^{q} \alpha_{ij} f_j[n], \quad \text{for all } 1 \le i \le p. \tag{4.3}$$

We assume throughout that the constant function $f_0[n] = 1$ is included, so that the classical AR(p) model is recovered as $a_i \equiv \alpha_{i0} \cdot 1$ whenever $\alpha_{ij} = 0$ for all $j > 0$. Many choices are possible for the functions $f_j[n]$—Legendre [92] and Fourier [94] polynomials, discrete prolate spheroidal functions [58], and even wavelets [115] have been used in speech applications.

According to the functional expansion of (4.3), the TVAR(p) model of (4.1) is fully described by $p(q+1)$ expansion coefficients α_{ij} and the gain term σ. For convenience

we group the coefficients α_{ij} into $q+1$ vectors $\boldsymbol{\alpha}_j \triangleq \begin{pmatrix}\alpha_{1j} & \alpha_{2j} & \cdots & \alpha_{pj}\end{pmatrix}^T$, $0 \leq j \leq q$, which induces a partition of the expansion coefficients according to:

$$\boldsymbol{\alpha} \triangleq \begin{pmatrix}\boldsymbol{\alpha}_0^T & | & \boldsymbol{\alpha}_1^T & \boldsymbol{\alpha}_2^T & \cdots & \boldsymbol{\alpha}_q^T\end{pmatrix}^T. \tag{4.4}$$

4.1.2 Example

As an illustration of how the TVAR model of (4.1) may be applied to time-frequency modeling of real data, consider the speech waveform shown in the top-left panel of Figure 4.1 comprising the vowel [ɑ] (as in "father") followed by a diphthong [aɪ] (as in "my"), and whispered during recording so that the turbulent source waveform may be modeled using an aperiodic noisy process, which is consistent with the form of (4.1). The formant values are relatively constant during the production of the steady vowel [ɑ], but as the jaw begins to close during the production of the diphthong [aɪ], the first and second formants begin to move down and up, respectively, as can be seen in the spectrogram shown in the bottom-right panel of Figure 4.1.

Next, an AR(8) and a TVAR(8) model were fitted to the *entire* data segment by using the covariance method in the case of the AR(8) model, and by using its generalization to the time-varying case as described in Section 4.3.1, in the case of the TVAR(8) model. In each case, the values of the AR coefficients at time index n were used to calculate an all-pole spectrum according to:

$$S(n, e^{j\omega}) \triangleq \left| \frac{1}{1 - \sum_{k=1}^{p} \widehat{a}_k[n] \exp(-j\omega k)} \right|^2. \tag{4.5}$$

The resultant sequences of power spectra are shown in top- and bottom-right panels of Figure 4.1 for the AR(8) and TVAR(8) models, respectively. In the case of the AR(8) model these values are the same for every value of n and reflect the *average* frequency content of the waveform. For instance, the energy around 1.8 kHz in the frozen-time AR(8) spectrum in the first half of the signal is a result of averaging; the spectrogram, in the bottom-left panel of Figure 4.1 clearly shows there should be no spectral peaks at that frequency until the second half of the signal. On the other hand, the temporal movement of the spectral peaks is clearly visible in the sequence of frozen-time spectra resulting from the fitted TVAR(8) model.

4.1.3 Alternative Formulations

The functional expansion approach to modeling the temporal evolution of the AR coefficient according to (4.3) was first studied in [92,116], and subsequently applied to speech and audio signal processing analysis by [57, 58, 93–99, 102, 117] among others. This model has also been widely applied to other areas including blind source separation [101, 103–106], radar signal processing [107], biomedical signal analysis [109–111] and instantaneous frequency estimation [112–114].

A popular alternative to the functional expansion approach to model each coefficient trajectory as a sample path of a suitably-chosen stochastic process [100, 118–121]. For

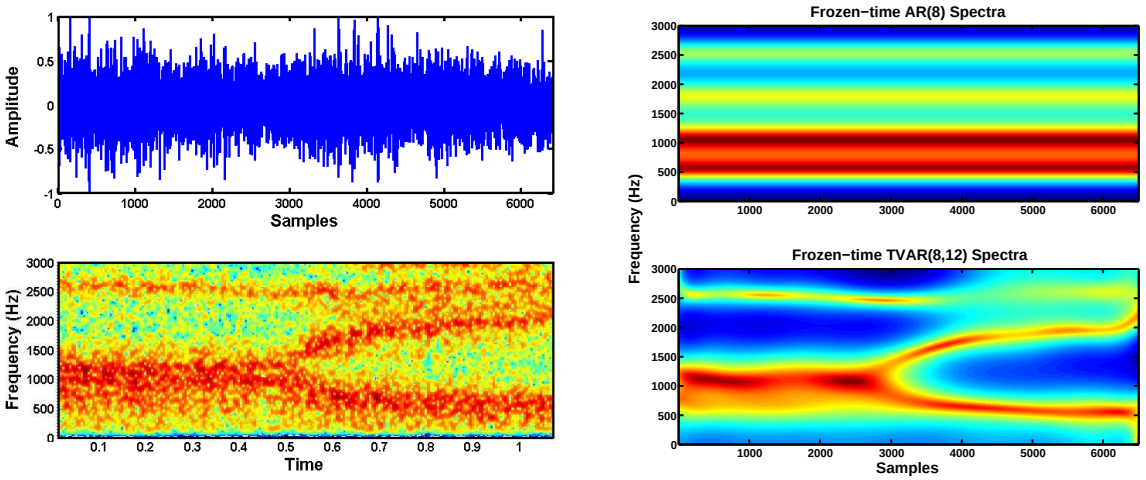


Figure 4.1: Analyzing the time-frequency structure of the whispered waveform [ɑaɪ] via short-time analysis (left) and AR/TVAR modeling (right). Left: Recorded waveform (top) shown together with a spectrogram computed using 10 ms Hamming windows with 50% overlap (bottom). Right: Sequences of power spectra defined by (4.5) with AR coefficients obtained by fitting an AR(8) model (top) and a TVAR(8) model with $q = 12$ Legendre polynomials to the waveform data.

example, the temporal evolution of each AR coefficient could modeled as a random walk:

$$a_i[n] = a_i[n-i] + \gamma_i v[n], \tag{4.6}$$

where $v[n]$ is a zero-mean, unit-variance white Gaussian sequence scaled by the gain parameter $\gamma_i > 0$. In this setting, however, estimation typically requires stochastic filtering [100] or iterative methods [121] in contrast to the least-squares estimators available for the model of (4.3), which are described in Section 4.3 below. Additional alternatives include adaptive filtering (see, e.g., the monograph of [122] and citations therein) or, in the case of multiple observations, the maximum-entropy covariance extension methods of [108, 123].

Finally, as mentioned in the introduction to this Chapter, two additional approaches for modeling TVAR coefficient trajectories are introduced in this thesis. An approach that blends the functional expansion approach with the random walk model of (4.6) is discussed in Section 4.5, while an approach based on viewing each TVAR process as a path on the manifold of AR processes is developed in Section 4.6.

4.2 Covariance Structure and Cramér-Rao Lower Bounds

Here we study the covariance structure of time-varying AR processes, show how to apply the results to computing Cramér-Rao lower bounds, and obtain a new way of visualizing time-frequency content.

4.2.1 Calculating Covariance Structure

Consider the subsequence of the TVAR process of (4.1) corresponding to the vector of random variables $\boldsymbol{x}$ defined by $\boldsymbol{x} \triangleq \begin{pmatrix} x[0] & x[1] & \cdots & x[N-1] \end{pmatrix}^T \in \mathbb{R}^{N \times 1}$. We are interested

in computing the $N \times N$ covariance matrix $\mathbb{E}(\boldsymbol{x}\boldsymbol{x}^T)$, under the assumption that $x[n] = 0$ for all $n < 0$. To this end, define the vector of innovations $\boldsymbol{w} \triangleq \begin{pmatrix} w[0] & w[1] & \cdots & w[N-1] \end{pmatrix}^T \in \mathbb{R}^{N \times 1}$, the matrix $\boldsymbol{A} \in \mathbb{R}^{N \times N}$ by

$$\boldsymbol{A} \triangleq \left(\begin{array}{ccc|cccccc} 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ -a_1[1] & 1 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \hline -a_p[p] & \cdots & -a_1[p] & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & -a_p[p+1] & \cdots & -a_1[p+1] & 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \cdots & 0 & 0 & -a_p[N-3] & \cdots & -a_1[N-3] & 1 & 0 & 0 \\ 0 & \cdots & 0 & 0 & -a_p[N-2] & \cdots & -a_1[N-2] & 1 & 0 \\ 0 & 0 & \cdots & 0 & 0 & -a_p[N-1] & \cdots & -a_1[N-1] & 1 \end{array}\right), \tag{4.7}$$

and observe that

$$\boldsymbol{w} = \boldsymbol{A}\boldsymbol{x} \tag{4.8}$$

is consistent with the difference equation of (4.1) for all discrete-time indices $n \in \{p, \ldots, N-1\}$. In addition, (4.8) corresponds to the following system of p equations associated to time indices $n \in \{0, \ldots, p-1\}$:

$$\begin{aligned} w[0] &= x[0], \\ w[1] &= -a_1[1]x[0] + x[1], \\ &\vdots \\ w[p-1] &= -a_{p-1}[p-1]x[0] - \ldots - a_1[p-1]x[p-2] + x[p-1], \end{aligned}$$

which is consistent with the assumption that $x[n]$ is one-sided with $x[n] = 0$ for all $n < 0$.

Next, observe that by using (4.8) we obtain

$$\mathbb{E}\left(\boldsymbol{w}\boldsymbol{w}^T\right) = \mathbb{E}\left(\boldsymbol{A}\boldsymbol{x}(\boldsymbol{A}\boldsymbol{x})^T\right) = \boldsymbol{A}\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}^T\right)\boldsymbol{A}^T = \boldsymbol{A}\boldsymbol{R}\boldsymbol{A}^T. \tag{4.9}$$

Since $\boldsymbol{A}$ is a lower-triangular matrix with ones along the main diagonal, it is invertible, and together with (4.9) this yields the following lower-diagonal-upper (LDU) decomposition of the covariance matrix $\boldsymbol{R}$ as:

$$\boldsymbol{R} = \boldsymbol{A}^{-1}\boldsymbol{\Sigma}(\boldsymbol{A}^T)^{-1} = \sigma^2\left(\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}, \tag{4.10}$$

with $\boldsymbol{\Sigma} \triangleq \sigma^2\boldsymbol{I}_N$ and $\boldsymbol{I}_N$ is the $N \times N$ identity matrix.

The LDU decomposition of (4.10) can be used to estimate the covariance structure of a TVAR process from the TVAR coefficients; similarly, an estimate of $\boldsymbol{R}$ can be obtained via (4.10) after the coefficient trajectories have been estimated from observed data. As an example, we synthesized $5,000$ instances of a 400-sample TVAR(2) process with piecewise-constant coefficient trajectories that undergo a single change exactly halfway at time index $n = 200$. Then the sample covariance was compared to the matrix $\boldsymbol{R}$ computed from the *known* TVAR coefficient trajectories via (4.10)—elements along the diagonals of $\boldsymbol{R}$ and $\widehat{\boldsymbol{R}}$ are plotted as time-series in Figure 4.2 with the kth diagonal containing $400 - k$ entries and

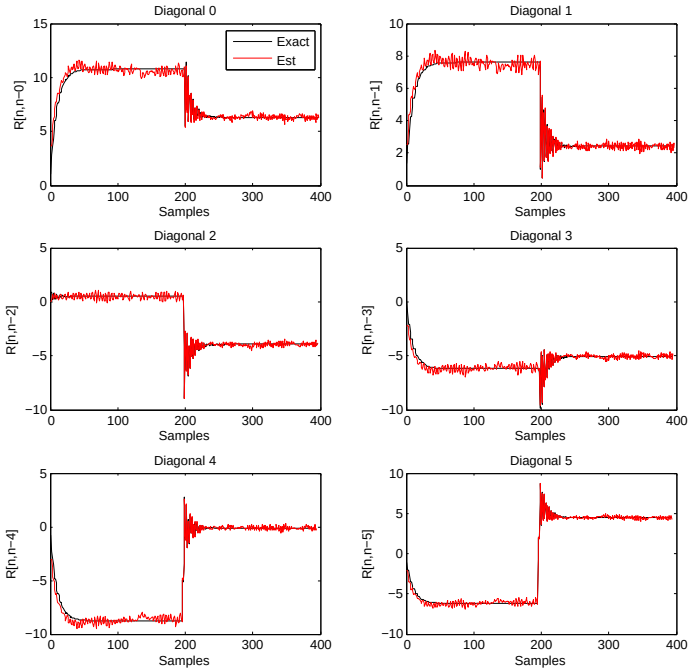


Figure 4.2: Computing the covariance matrix $\boldsymbol{R}$ of a 400-sample TVAR(2) process. Six diagonals of the sample covariance matrix are plotted as time series (red lines) and are overlaid with estimates obtained via the LDU decomposition of (4.10) (black lines).

the main diagonal corresponding to $k = 0$. The entries of the sample covariance matrix are seen to converge to those computed via (4.10), confirming the calculations.

The $\boldsymbol{A}$ is fundamentally related to the likelihood function of a TVAR process. Define the matrix $\widetilde{\boldsymbol{A}} \in \mathbb{R}^{(N-p)\times N}$ to be the lower block of $\boldsymbol{A}$ in (4.7). Gaussianity of $w[n]$ then implies that given N observations of $x[n]$, partitioned according to

$$\boldsymbol{x} = (\boldsymbol{x}_p \,|\, \boldsymbol{x}_{N-p})^T \triangleq \big(x[0] \cdots x[p-1] \;\mid\; x[p] \cdots x[N-1]\big)^T, \tag{4.11}$$

the conditional probability density function of $\boldsymbol{\alpha}, \sigma^2$ is given by:

$$\begin{aligned} p(\boldsymbol{x}_{N-p} \,|\, \boldsymbol{x}_p \,;\, \boldsymbol{\alpha}, \sigma^2) &= \frac{1}{(2\pi\sigma^2)^{(N-p)/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{n=p}^{N-1} \left(x[n] - \sum_{i=1}^{p} a_i[n] x[n-i] \right)^2 \right) \\ &= \frac{1}{(2\pi\sigma^2)^{(N-p)/2}} \exp\left(-\frac{1}{2\sigma^2} \boldsymbol{x}^T \widetilde{\boldsymbol{A}}^T \widetilde{\boldsymbol{A}} \boldsymbol{x} \right). \end{aligned} \tag{4.12}$$

The conditional likelihood of (4.12) is quadratic in $\boldsymbol{x}$; its restriction to the time-invariant setting is well known (see, e.g., [124, Section 4.3]), and proven useful in missing data interpolation [124, Chapter 5] and blind source separation problems [103]. Note that the $N \times N$ matrix $\widetilde{\boldsymbol{A}}^T \widetilde{\boldsymbol{A}}$ is not invertible since it has rank $N-p$. Thus to obtain the *conditional* covariance matrix of $\boldsymbol{x}_{N-p}$, the first p columns of $\widetilde{A}$, which correspond to the conditioning variables $\boldsymbol{x}_p$, need to be dropped. The resultant $(N-p)\times(N-p)$ matrix, corresponding to the lower-right-hand block of (4.7), is then inverted to obtain the desired covariance matrix.

4.2.2 Cramér-Rao Lower Bounds

The calculations in the previous section may be used in order to calculate the Cramér-Rao lower bounds for parameter estimators associated to the TVAR model of (4.1). The following second derivatives are necessary:

$$\frac{\partial \ln p(\boldsymbol{x}_{N-p} \,|\, \boldsymbol{x}_p; \boldsymbol{\alpha}, \sigma^2)}{\partial \alpha_{ij} \partial \alpha_{kl}} = -\frac{1}{\sigma^2} \sum_{n=p}^{N-1} \left(\frac{\partial e[n]}{\partial \alpha_{kl}} \right) f_j[n] x[n-i] = -\frac{1}{\sigma^2} \sum_{n=p}^{N-1} f_j[n] f_l[n] x[n-i] x[n-k]$$

$$\frac{\partial \ln p(\boldsymbol{x}_{N-p} \,|\, \boldsymbol{x}_p; \boldsymbol{\alpha}, \sigma^2)}{\partial \alpha_{ij} \partial \sigma^2} = \frac{\partial}{\partial \sigma^2} \left(-\frac{1}{\sigma^2} \sum_{n=p}^{N-1} e[n] f_j[n] x[n-i] \right) = \frac{1}{\sigma^4} \sum_{n=p}^{N-1} e[n] f_j[n] x[n-i]$$

$$\frac{\partial \ln p(\boldsymbol{x}_{N-p} \,|\, \boldsymbol{x}_p; \boldsymbol{\alpha}, \sigma^2)}{\partial \sigma^2 \partial \sigma^2} = \frac{N-p}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{n=p}^{N-1} e^2[n].$$

The expectations required to compute the Fisher information matrix $\boldsymbol{I}$ are then given by:

$$-\mathbb{E}\left(\frac{\ln p(\boldsymbol{x}_{N-p} \,|\, \boldsymbol{x}_p; \boldsymbol{\alpha}, \sigma^2)}{\partial \alpha_{ij} \partial \alpha_{kl}} \right) = \frac{1}{\sigma^2} \sum_{n=p}^{N-1} f_j[n] f_l[n] \mathbb{E}(x[n-i]x[n-k]) \tag{4.13}$$

$$-\mathbb{E}\left(\frac{\ln p(\boldsymbol{x}_{N-p} \,|\, \boldsymbol{x}_p; \boldsymbol{\alpha}, \sigma^2)}{\partial \alpha_{ij} \partial \sigma^2} \right) = -\frac{1}{\sigma^4} \sum_{n=p}^{N-1} f_j[n] \mathbb{E}(e[n]x[n-i]) = 0$$

$$-\mathbb{E}\left(\frac{\ln p(\boldsymbol{x}_{N-p} \,|\, \boldsymbol{x}_p; \boldsymbol{\alpha}, \sigma^2)}{\partial \sigma^2 \partial \sigma^2} \right) = -\frac{N-p}{2\sigma^4} + \frac{1}{\sigma^6} \sum_{n=p}^{N-1} \mathbb{E}\big(e^2[n]\big) = \frac{N-p}{2\sigma^4}.$$

Let $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\sigma^2}$ be unbiased estimators of $\boldsymbol{\alpha}$ and σ^2, respectively. Then, the Cramér-Rao lower bounds on the variance of $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\sigma^2}$ are given by:

$$\operatorname{Cov} \begin{pmatrix} \widehat{\boldsymbol{\alpha}} \\ \widehat{\sigma^2} \end{pmatrix} \geq \boldsymbol{I}^{-1},$$

with the Fisher information matrix $\boldsymbol{I}$ defined according to:

$$\boldsymbol{I} \triangleq \begin{pmatrix} \boldsymbol{I}_{\boldsymbol{\alpha}\boldsymbol{\alpha}} & \boldsymbol{I}_{\boldsymbol{\alpha}\sigma^2} \\ \boldsymbol{I}_{\sigma^2\boldsymbol{\alpha}} & \boldsymbol{I}_{\sigma^2\sigma^2} \end{pmatrix} = \begin{pmatrix} \boldsymbol{I}_{\boldsymbol{\alpha}\boldsymbol{\alpha}} & \boldsymbol{0}_{p(q+1)\times 1} \\ \boldsymbol{0}_{1\times p(q+1)} & \frac{N-p}{2\sigma^4} \end{pmatrix},$$

where entries of $\boldsymbol{I}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}$ are computed via (4.13) and depend on the covariance matrix $\boldsymbol{R}$ that can be computed via (4.10).

4.2.3 Visualizing Time-Frequency Content

Since one of the main goals of nonstationary signal modeling is to capture the time-varying spectral content of real-world signals, it is natural to find a way to visualize the time-frequency content captured by a TVAR process. We propose a new time-frequency display and compare it to an existing approach; both may be useful in practice.

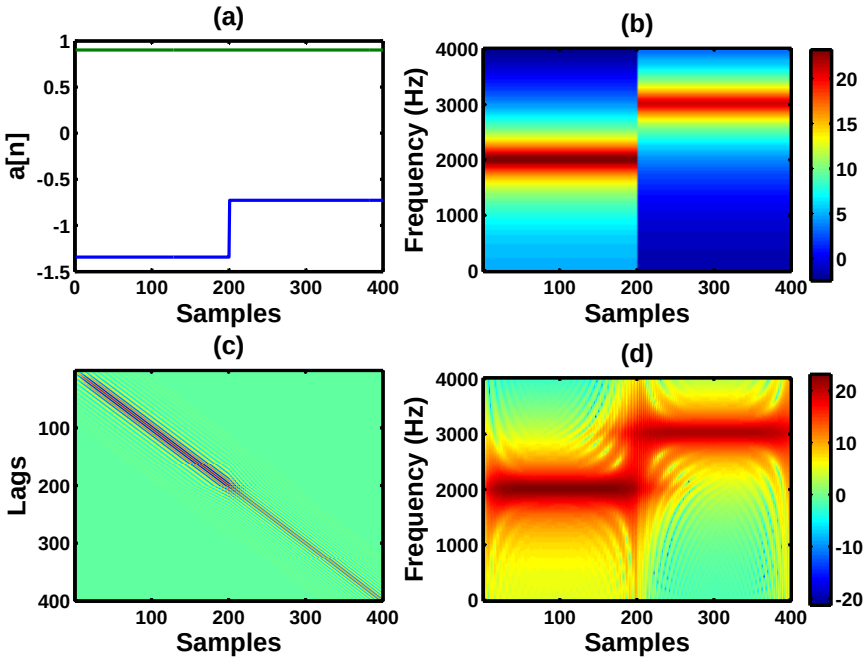


Figure 4.3: Visualizing the time-frequency content of a TVAR(2) process. Time-varying AR coefficient trajectories (a) are shown along with the associated sequence of frozen-time power spectra computed via (4.5)(b) and the covariance matrix computed via (4.10) (c). The time-frequency content obtained from the covariance matrix via a DFT (d) exhibits different structure from the sequence frozen-time spectra in (b).

Given TVAR coefficient trajectories $\{a_1[n], \ldots, a_p[n] \,|\, 0 \leq n \leq N-1\}$, which are either known apriori or estimated from data, we may construct an image of time-frequency content by calculating a *frozen-time* power spectrum for each time index n via (4.5). Then $S(n, e^{j\omega})$ could serve as a spectrographic display as shown, for instance, in the bottom-right panel of Figure 4.1. This method is quite popular in the extant literature, and likely predates its appearance in [57]. In contrast, we propose an alternative approach based computing the covariance matrix $\boldsymbol{R}$ from the TVAR coefficient trajectories via (4.10) and calculating a one-dimensional discrete Fourier transform (DFT) of its rows.

The differences between these two approaches are easily illustrated using a 400-sample piecewise-constant TVAR(2) process synthesized by filtering white Gaussian noise via an autoregressive filter whose time-varying coefficients are shown in the top-left panel of Figure 4.3. The AR coefficients in the first and second halves of the signal correspond to second-order digital resonators with equal bandwidths and center frequencies of 2000 Hz and 3000 Hz, respectively. The associated sequence of frozen-time spectra computed via (4.5) is shown in top-right panel of Figure 4.3, while the covariance matrix $\boldsymbol{R}$ obtained via (4.10) and the associated DFT are shown in the bottom- left and right panels of Fig 4.3, respectively.

These two visualizations essentially agree in regions where the signal is "stationary." Their differences are quite pronounced, however, in the vicinity of the discontinuity centered at time index $n = 200$. The frozen-time representation based on (4.5) shows an in-

stantaneous change in the time-frequency content, whereas the covariance-based approach shows a more gradual transition indicative of the memory in an autoregressive process. This difference can be exacerbated by shrinking the bandwidth of the underlying resonator toward 0, which has the effect of moving the instantaneous "poles" toward the unit circle.

Note that, from a theoretical point of view, neither of these two approaches is more appropriate than the other. Indeed, we have not avoided the fact that generalizing equivalent parameterizations of stationary processes to the nonstationary setting does not yield equivalent classes of nonstationary processes as discussed in Section 2.6.2.3. The frozen-time approach is implicit in the second method, too, since the columns of the covariance matrix $\boldsymbol{R}$ are transformed independently from one another using a one-dimensional DFT. The primary difference lies in whether the frozen-time assumption is made with respect to the TVAR coefficients or the associated covariance structure—both calculations can be useful tools. Above all else, this example illustrates that care must be taken when manipulating time-domain and "frequency"-domain information in the nonstationary setting.

4.3 Parameter Estimation

In this section we consider estimation of the TVAR model parameters $\boldsymbol{\alpha}$ and σ^2. Following [94], we discuss the generalization of the covariance and autocorrelation methods of linear prediction, described earlier in Section 2.4, to the time-varying setting in Sections 4.3.1 and 4.3.2, respectively. Both of these methods will be used in Chapter 5. Next, in Section 4.3.3, we propose a novel estimator of the TVAR coefficients when only noisy observations of the process are available. Lattice-based estimation is treated later during our discussion of frozen-time stability in Section 4.4.1.3.

4.3.1 Conditional Maximum Likelihood Estimation

Given N observations of a TVAR(p) process, partitioned according to (4.11), the joint probability density function of $\boldsymbol{\alpha}, \sigma^2$ is given by:

$$p(\boldsymbol{x}\,;\,\boldsymbol{\alpha},\sigma^2) = p(\boldsymbol{x}_{N-p}\,|\,\boldsymbol{x}_p\,;\,\boldsymbol{\alpha},\sigma^2)p(\boldsymbol{x}_p\,;\,\boldsymbol{\alpha},\sigma^2). \tag{4.14}$$

As is standard practice, we approximate the *unconditional* data likelihood of (4.14) by the *conditional* likelihood $p(\boldsymbol{x}_{N-p}\,|\,\boldsymbol{x}_p\,;\,\boldsymbol{\alpha},\sigma^2)$, whose maximization yields an estimator that converges to the exact (unconditional) ML estimator as $N \to \infty$.

Gaussianity of $w[n]$ implies the conditional likelihood

$$\begin{aligned} p(\boldsymbol{x}_{N-p}\,|\,\boldsymbol{x}_p;\boldsymbol{\alpha},\sigma^2) &= \frac{1}{(2\pi\sigma^2)^{(N-p)/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{n=p}^{N-1}\left(x[n]-\sum_{i=1}^{p}\sum_{j=0}^{q}\alpha_{ij}f_j[n]x[n-i]\right)^2\right), \\ &= \frac{1}{(2\pi\sigma^2)^{(N-p)/2}} \exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{x}_{N-p}-\boldsymbol{H}_{\boldsymbol{x}}\boldsymbol{\alpha}\right)^T\left(\boldsymbol{x}_{N-p}-\boldsymbol{H}_{\boldsymbol{x}}\boldsymbol{\alpha}\right)\right) \end{aligned}$$

where the $(n-p+1)$th row of the matrix $\boldsymbol{H}_{\boldsymbol{x}} \in \mathbb{R}^{(N-p)\times p(q+1)}$ is given by the Kronecker product $(x[n-1]\ \cdots\ x[n-p]) \otimes (f_0[n]\ f_1[n]\ \cdots\ f_q[n])$ for any $p \le n \le N-1$. The

conditional log-likelihood function is therefore given by

$$\ln p(\boldsymbol{x}_{N-p} \,|\, \boldsymbol{x}_p; \boldsymbol{\alpha}, \sigma^2) = -\frac{N-p}{2} \ln(2\pi\sigma^2) - \frac{(\boldsymbol{x}_{N-p} - \boldsymbol{H}_{\boldsymbol{x}}\boldsymbol{\alpha})^T (\boldsymbol{x}_{N-p} - \boldsymbol{H}_{\boldsymbol{x}}\boldsymbol{\alpha})}{2\sigma^2}. \quad (4.15)$$

Maximizing (4.15) with respect to $\boldsymbol{\alpha}$ yields the least-squares solution of the following linear regression problem:

$$\boldsymbol{x}_{N-p} = \boldsymbol{H}_{\boldsymbol{x}}\boldsymbol{\alpha} + \sigma \boldsymbol{w}_{N-p}, \quad (4.16)$$

where $\boldsymbol{w}_{N-p} \triangleq \begin{pmatrix} w[p] & w[p+1] & \cdots & w[N-1] \end{pmatrix}^T$. Consequently, the conditional ML estimate of $\boldsymbol{\alpha}$ follows from (4.15) and (4.16) as

$$\widehat{\boldsymbol{\alpha}} = \left(\boldsymbol{H}_{\boldsymbol{x}}^T \boldsymbol{H}_{\boldsymbol{x}}\right)^{-1} \boldsymbol{H}_{\boldsymbol{x}}^T \boldsymbol{x}_{N-p}. \quad (4.17)$$

The estimator of (4.17) corresponds to a generalization of the *covariance method* of linear prediction, which we described earlier in Section 2.4.1, and to which it exactly reduces when the number q of non-constant basis functions employed is set to zero.

The conditional ML estimate of σ^2 is obtained by substituting (4.17) into (4.15) and maximizing with respect to σ^2, yielding

$$\widehat{\sigma^2} = \frac{1}{N-p} \sum_{n=p}^{N-1} \left(x[n]x[n] - \sum_{i=1}^{p} \sum_{j=0}^{q} \widehat{\alpha_{ij}} f_j[n] x[n] x[n-i] \right). \quad (4.18)$$

When q is set to 0, (4.18) reduces to the familiar estimator of the variance: $\widehat{\sigma^2} = \widehat{r}_{xx}[0] - \sum_{i=1}^{p} \alpha_{i0} \widehat{r}_{xx}[i]$, where $r_{xx}[\tau]$ is the autocorrelation function of $x[n]$ at lag τ.

4.3.2 Method-of-Moments Estimation

The autocorrelation method of linear prediction can also be generalized to the time-varying case. The derivation is based on the *synchronous* TVAR model of (4.2) in lieu of the *shifted* model of (4.1). Grouping the coefficients α_{ij} into p vectors $\widetilde{\boldsymbol{\alpha}}_i \triangleq (\alpha_{i0} \; \alpha_{i1} \; \cdots \; \alpha_{iq})^T, 1 \leq i \leq p$, induces a partition of the expansion coefficients given by:

$$\widetilde{\boldsymbol{\alpha}} \triangleq (\widetilde{\boldsymbol{\alpha}}_1^T \; \widetilde{\boldsymbol{\alpha}}_2^T \; \cdots \; \widetilde{\boldsymbol{\alpha}}_p^T)^T$$

—a permutation of elements of $\boldsymbol{\alpha}$ in (4.4). The autocorrelation estimator of $\widetilde{\boldsymbol{\alpha}}$ is then obtained by minimizing the prediction error over all $n \in \mathbb{Z}$, while assuming that $x[n] = 0$ for all $n \notin [0, \ldots, N-1]$, and is equivalent to the least-squares solution of:

$$\boldsymbol{x} = \widetilde{\boldsymbol{H}}_{\boldsymbol{x}} \widetilde{\boldsymbol{\alpha}} + \sigma \boldsymbol{w}, \quad (4.19)$$

where the nth row of $\widetilde{\boldsymbol{H}}_{\boldsymbol{x}} \in \mathbb{R}^{N \times p(q+1)}$ is given by $(f_0[n-1]x[n-1] \; \cdots \; f_0[n-p]x[n-p] \; \cdots \; f_q[n-1]x[n-1] \; \cdots \; f_q[n-p]x[n-p])$. The *autocorrelation* estimate of $\widetilde{\boldsymbol{\alpha}}$ then follows from (4.19) as:

$$\widehat{\widetilde{\boldsymbol{\alpha}}} = (\widetilde{\boldsymbol{H}}_{\boldsymbol{x}}^T \widetilde{\boldsymbol{H}}_{\boldsymbol{x}})^{-1} \widetilde{\boldsymbol{H}}_{\boldsymbol{x}}^T \boldsymbol{x}. \quad (4.20)$$

An alternate way to arrive at the estimator of (4.20), which parallels the development of the covariance method in [94], is to compute the subblocks of the matrices $\widetilde{\boldsymbol{H}}_{\boldsymbol{x}}^T \widetilde{\boldsymbol{H}}_{\boldsymbol{x}}$

and $\widetilde{\boldsymbol{H}}_{\boldsymbol{x}}^T\boldsymbol{x}$ explicitly. To this end, note that the solution to (4.19) must satisfy the following set of $p(q+1)$ *generalized* Yule-Walker equations for $1 \leq i,k \leq p$ and $0 \leq j,l \leq q$:

$$\sum_{i=1}^{p}\sum_{j=0}^{q}\alpha_{ij}c_{j,l}(i,k) = -c_{0,l}(k,0), \tag{4.21}$$

with the function $c_{j,l}(i,k)$ given by: $c_{j,l}(i,k) = \sum_{n=0}^{N-1} f_j[n-i]f_l[n-k]x[n-i]x[n-k]$. By defining $\boldsymbol{\psi}_i \triangleq (c_{00}(i,0)\ c_{01}(i,0)\ \cdots\ c_{0q}(i,0))^T$, the $(q+1)\times(q+1)$ matrix $\boldsymbol{\Phi}_{ik}$ as:

$$\boldsymbol{\Phi}_{ik} = \begin{pmatrix} c_{00}(i,k) & c_{10}(i,k) & \ldots & c_{q0}(i,k) \\ c_{01}(i,k) & c_{11}(i,k) & \ldots & c_{1q}(i,k) \\ \vdots & \vdots & \ddots & \vdots \\ c_{0q}(i,k) & c_{1q}(i,k) & \ldots & c_{qq}(i,k) \end{pmatrix}, \tag{4.22}$$

and the augmented matrices $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ by:

$$\boldsymbol{\Phi} = \begin{pmatrix} \boldsymbol{\Phi}_{11} & \boldsymbol{\Phi}_{12} & \ldots & \boldsymbol{\Phi}_{1p} \\ \boldsymbol{\Phi}_{21} & \boldsymbol{\Phi}_{22} & \ldots & \boldsymbol{\Phi}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Phi}_{p1} & \boldsymbol{\Phi}_{p2} & \ldots & \boldsymbol{\Phi}_{pp} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Psi} = \begin{pmatrix} \boldsymbol{\psi}_1 \\ \boldsymbol{\psi}_2 \\ \vdots \\ \boldsymbol{\psi}_p \end{pmatrix}, \tag{4.23}$$

we may rewrite (4.21) as $\boldsymbol{\Phi}\widetilde{\boldsymbol{\alpha}} = -\boldsymbol{\Psi}$, yielding the following least-squares estimator of $\widetilde{\boldsymbol{\alpha}}$:

$$\widehat{\widetilde{\boldsymbol{\alpha}}} = -\boldsymbol{\Phi}^{-1}\boldsymbol{\Psi}. \tag{4.24}$$

Comparing (4.20) to (4.24), we observe that $\boldsymbol{\Phi} = \widetilde{\boldsymbol{H}}_{\boldsymbol{x}}^T\widetilde{\boldsymbol{H}}_{\boldsymbol{x}}$ and $\boldsymbol{\Psi} = \widetilde{\boldsymbol{H}}_{\boldsymbol{x}}^T\boldsymbol{x}$. By inspection of (4.21)–(4.23), it is apparent that $\boldsymbol{\Phi}$ is a *block-Toeplitz* matrix comprising p^2 *symmetric* blocks of size $(q+1)\times(q+1)$—this special structure arises as a direct consequence of the synchronous form of the TVAR trajectories in (4.2), and may be used to invert $\widetilde{\boldsymbol{H}}_{\boldsymbol{x}}^T\widetilde{\boldsymbol{H}}_{\boldsymbol{x}}$ efficiently using $O(p^3(q+1)^2)$ operations [125].

4.3.3 Estimation in Presence of Noise

When only noisy observations of the TVAR process specified by (4.1) and (4.3) are available, it is more difficult to estimate the unknown TVAR coefficients because their likelihood is a *nonlinear* function of the observed data. In the time-invariant case, for example, estimating an AR(p) model from noisy observations is equivalent to estimating an ARMA(p,p) model [4]. Despite these difficulties, a large number of parameter estimation methods have been proposed in the time-invariant case ranging from approaches based on ARMA modeling [4] and noise compensation [126] to those involving subspace methods [127] and approximate maximum likelihood methods [128,129]. In contrast, the problem of using noisy data to estimate TVAR coefficients when they are represented by a functional basis as in (4.3) has not yet been addressed in the literature, to our knowledge. Here, we address this problem by generalizing the EM approach of [128,129] to the time-varying setting.

4.3.3.1 Model Formulation

We assume that a TVAR(p) process is observed in additive white Gaussian noise according to:

$$\begin{aligned} x[n] &= \sum_{i=1}^{p}\sum_{j=0}^{q} \alpha_{ij} f_j[n] x[n-i] + \sigma_w w[n], \\ y[n] &= x[n] + \sigma_v v[n], \end{aligned} \tag{4.25}$$

where $w[n]$ and $v[n]$ are zero-mean, unit-variance white Gaussian noise sequences scaled by $\sigma_w > 0$ and $\sigma_v > 0$, respectively. Rewriting (4.25) in state-space form yields:

$$\begin{aligned} \boldsymbol{x}_n &= \boldsymbol{F}_n \boldsymbol{x}_{n-1} + \boldsymbol{Q} v[n] \\ y[n] &= \boldsymbol{H}\boldsymbol{x}_n + \boldsymbol{R} w[n], \end{aligned} \tag{4.26}$$

where the *augmented*[1] state vector is defined by $\boldsymbol{x}_n = \begin{pmatrix} x[n] & x[n-1] & \cdots & x[n-p] \end{pmatrix}^T$, and the matrices $\boldsymbol{F}_n \in \mathbb{R}^{(p+1)\times(p+1)}$, $\boldsymbol{Q} \in \mathbb{R}^{(p+1)\times 1}$, $\boldsymbol{H} \in \mathbb{R}^{1\times(p+1)}$ and $\boldsymbol{R} \in \mathbb{R}$ are given by:

$$\boldsymbol{F}_n = \begin{pmatrix} a_1[n] & a_2[n] & \dots & a_p[n] & 0 \\ 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{pmatrix} \qquad \boldsymbol{Q} = \begin{pmatrix} \sigma_v^2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$\boldsymbol{H} = \begin{pmatrix} 1 & 0 & \cdots & 0 \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{R} = (\sigma_w^2).$$

We complete the model specification by assuming that $\boldsymbol{x}_0 \sim \mathcal{N}(\boldsymbol{x}_0; \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$. Consequently, the state-space model of (4.26) is parameterized by the set of parameters $\boldsymbol{\theta}$ defined by:

$$\boldsymbol{\theta} \triangleq (\boldsymbol{\alpha}, \sigma_w^2, \sigma_v^2, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0).$$

Finally, we set up a useful notational device by noting that the vector $\boldsymbol{a}_n \triangleq \begin{pmatrix} a_1[n] & a_2[n] & \cdots & a_p[n] \end{pmatrix}^T$ can be written according to:

$$\boldsymbol{a}_n = \boldsymbol{B}_n \boldsymbol{\alpha}, \tag{4.27}$$

where $\boldsymbol{\alpha}$ is a $p(q+1) \times 1$ vector of TVAR coefficients defined by (4.4), and the matrix $\boldsymbol{B}_n \in \mathbb{R}^{p\times p(q+1)}$ is defined via:

$$\boldsymbol{B}_n(j,k) = \begin{cases} f_j[n] \text{ if } k \in \{j, q+j, 2q+j, \dots, pq+j\} \\ 0 \quad \text{otherwise.} \end{cases} \tag{4.28}$$

[1]The variable $x[n-p]$ is added to the state vector $\boldsymbol{x}_n$ so that the one-step-ahead covariance for $\begin{pmatrix} x[n] & x[n-1] & \cdot & x[n-p+1] \end{pmatrix}$, needed in the implementation of the EM algorithm, is automatically calculated by the Kalman smoother.

4.3.3.2 Inference via EM Algorithm

We now describe how to obtain an ML estimate of the parameters $\boldsymbol{\theta}$ given a vector of N observations $\boldsymbol{y} \triangleq \begin{pmatrix} y[0] & y[1] & \cdots & y[N-1] \end{pmatrix}^T$. Since the observed data likelihood $L(\boldsymbol{y};\boldsymbol{\theta})$ is difficult to work with, we approach the problem using the expectation-maximization framework. The approach is similar to the case of the EM algorithm for estimating parameters in the formant tracking setting of Section 3.5.2, but the details of the calculations are different.

Viewing the unobserved variable $\boldsymbol{x} \triangleq \begin{pmatrix} x[0] & x[1] & \cdots & x[N-1] \end{pmatrix}^T$ as "missing data", the complete data log likelihood $L(\boldsymbol{\theta};\boldsymbol{x},\boldsymbol{y})$ follows from (4.26) as

$$\begin{aligned} L(\boldsymbol{\theta};\boldsymbol{x},\boldsymbol{y}) \triangleq \log p(\boldsymbol{x},\boldsymbol{y}\,;\,\boldsymbol{\theta}) = &-\frac{1}{2}\log|\boldsymbol{\Sigma}_0| - \frac{1}{2}(\boldsymbol{x}_0-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{x}_0-\boldsymbol{\mu}_0) \\ &- \frac{N-1}{2}\log|\boldsymbol{Q}| - \frac{1}{2}\sum_{n=1}^{N-1}(\boldsymbol{x}_n-\boldsymbol{F}_n\boldsymbol{x}_{n-1})^T\boldsymbol{Q}^{-1}(\boldsymbol{x}_n-\boldsymbol{F}_n\boldsymbol{x}_{n-1}) \\ &- \frac{N-1}{2}\log|\boldsymbol{R}| - \frac{1}{2}\sum_{n=1}^{N-1}(y[n]-\boldsymbol{H}\boldsymbol{x}_n)^T\boldsymbol{R}^{-1}(y[n]-\boldsymbol{H}\boldsymbol{x}_n). \end{aligned} \tag{4.29}$$

Substituting the values of $\boldsymbol{F}_n, \boldsymbol{Q}, \boldsymbol{H}, \boldsymbol{R}$ into (4.29) yields:

$$\begin{aligned} L(\boldsymbol{\theta};\boldsymbol{x},\boldsymbol{y}) = &-\frac{1}{2}\log|\boldsymbol{\Sigma}_0| - \frac{1}{2}(\boldsymbol{x}_0-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{x}_0-\boldsymbol{\mu}_0) - \frac{N-1}{2}\log\sigma_w^2 - \frac{1}{2\sigma_w^2}\sum_{n=1}^{N-1}(x[n]-\boldsymbol{a}_n^T\boldsymbol{x}_{n-1})^2 \\ &- \frac{N-1}{2}\log\sigma_v^2 - \frac{1}{2\sigma_v^2}\sum_{n=1}^{N-1}(y[n]-x[n])^2. \end{aligned}$$

When $N >> p$, the complete log likelihood can be approximated via

$$L(\boldsymbol{\theta};\boldsymbol{x},\boldsymbol{y}) \approx -\frac{N-1}{2}\left(\log\sigma_v^2+\log\sigma_w^2\right) - \frac{1}{2\sigma_w^2}\sum_{n=1}^{N-1}(x[n]-\boldsymbol{a}_n^T\boldsymbol{x}_{n-1})^2 - \frac{1}{2\sigma_v^2}\sum_{n=1}^{N-1}(y[n]-x[n])^2. \tag{4.30}$$

In addition to the complete log likelihood, we also need a way to compute $p(\boldsymbol{x}\,|\,\boldsymbol{y};\boldsymbol{\theta})$—the conditional probability of the missing data given the observations and unknown parameters. Since the state-space model of (4.26) is linear and Gaussian, $p(\boldsymbol{x}_n\,|\,\boldsymbol{y};\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{x}_n;\boldsymbol{m}_{n|N},\boldsymbol{P}_{n|N})$ where the quantities $\boldsymbol{m}_{n|N}$ and $\boldsymbol{P}_{n|N}$ are defined according to:

$$\boldsymbol{m}_{n|N} \triangleq \mathbb{E}(\boldsymbol{x}_n|\boldsymbol{y};\boldsymbol{\theta}) \quad \text{and} \quad P_{n|N} \triangleq \mathbb{E}\big((\boldsymbol{x}_n-\boldsymbol{m}_{n|N})(\boldsymbol{x}_n-\boldsymbol{m}_{n|N})^T|\boldsymbol{y};\boldsymbol{\theta}\big)\,,$$

and may be recursively computed using the Kalman smoother summarized in Algorithm 4.1.

Given expressions for the complete log likelihood and the conditional distribution of the missing data $p(\boldsymbol{x}_n\,|\,\boldsymbol{y};\boldsymbol{\theta})$, define the function $Q(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}})$ according to:

$$Q(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}) = \mathbb{E}_{p(\boldsymbol{x}_n\,|\,\boldsymbol{y};\widehat{\boldsymbol{\theta}})}L(\boldsymbol{\theta};\boldsymbol{x},\boldsymbol{y}) = \int_{\boldsymbol{x}} p(\boldsymbol{x}_n\,|\,\boldsymbol{y};\widehat{\boldsymbol{\theta}})L(\boldsymbol{\theta};\boldsymbol{x},\boldsymbol{y})d\boldsymbol{x}. \tag{4.31}$$

Algorithm 4.1 TVAR Estimation: Kalman Smoother

- Initialize: $\boldsymbol{m}_{-1|-1} = \mathbf{0}_{p\times 1},\ \boldsymbol{P}_{-1|-1} = \mathbf{0}_{p\times p}$
- Filtering: Repeat for $n = 0, \ldots, N-1$

$$\begin{aligned}
\boldsymbol{m}_{n|n-1} &= \boldsymbol{F}_n \boldsymbol{m}_{n-1|n-1} \\
\boldsymbol{P}_{n|n-1} &= \boldsymbol{F}_n \boldsymbol{P}_{n-1|n-1} \boldsymbol{F}_n^T + \boldsymbol{Q} \\
\boldsymbol{S}_n &= \boldsymbol{H} \boldsymbol{P}_{n|n-1} \boldsymbol{H}^T + \boldsymbol{R} \\
\boldsymbol{m}_{n|n} &= \boldsymbol{m}_{n|n-1} + \boldsymbol{P}_{n|n-1} \boldsymbol{H}^T \boldsymbol{S}_n^{-1} (\boldsymbol{y}_n - \boldsymbol{H} \boldsymbol{m}_{n|n-1}) \\
\boldsymbol{P}_{n|n} &= \boldsymbol{P}_{n|n-1} - \boldsymbol{P}_{n|n-1} \boldsymbol{H}^T \boldsymbol{S}_n^{-1} \boldsymbol{H} \boldsymbol{P}_{n|n-1}
\end{aligned}$$

- Smoothing: Repeat for $n = N-1, \ldots, 1$

$$\begin{aligned}
\boldsymbol{S}_{n-1} &= \boldsymbol{P}_{n-1|n-1} \boldsymbol{F}_n^T \boldsymbol{P}_{n|n-1}^{-1} \\
\boldsymbol{m}_{n-1|N} &= \boldsymbol{m}_{n-1|n-1} + \boldsymbol{S}_{n-1} \left(\boldsymbol{\mu}_{n|N} - \boldsymbol{F}_n \boldsymbol{m}_{n-1|n-1} \right) \\
\boldsymbol{P}_{n-1|N} &= \boldsymbol{P}_{n-1|n-1} + \boldsymbol{S}_{n-1} \left(\boldsymbol{P}_{n|N} - \boldsymbol{P}_{n-1|n-1} \right) \boldsymbol{S}_{n-1}^T
\end{aligned}$$

The EM algorithm then proceeds by alternating between computing $Q(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}})$ for some fixed $\widehat{\boldsymbol{\theta}}$ and maximizing it with respect to $\boldsymbol{\theta}$ in order to obtain a new estimate of the unknown parameters. The entire procedure is summarized in Algorithm 4.2.

The exact computations involved in the mth E-step and M-steps of Algorithm 4.2 are obtained as follows. From (4.31) and (4.30), we obtain the following expression for $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$:

$$\begin{aligned}
Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) = &-\frac{N-1}{2} \log \sigma_v^2 - \frac{N-1}{2} \log \sigma_w^2 - \frac{1}{2\sigma_v^2} \sum_{n=1}^{N-1} \left\{ y^2[n] - 2y[n] \mathbb{E}\left(x[n] | \boldsymbol{y}; \boldsymbol{\theta}^m \right) + \mathbb{E}\left(x^2[n] | \boldsymbol{y}; \boldsymbol{\theta}^{(m)} \right) \right\} \\
&- \frac{1}{2\sigma_w^2} \sum_{n=1}^{N-1} \left\{ E\left(x^2[n] | \boldsymbol{y}; \boldsymbol{\theta}^{(m)} \right) - 2\mathbb{E}\left(\boldsymbol{a}_n^T \boldsymbol{x}_{n-1} x[n] | \boldsymbol{y}; \boldsymbol{\theta}^{(m)} \right) + \boldsymbol{\alpha}^T \boldsymbol{B}_n^T \mathbb{E}\left(\boldsymbol{x}_{n-1} \boldsymbol{x}_{n-1}^T | \boldsymbol{y}; \boldsymbol{\theta}^{(m)} \right) \boldsymbol{B}_n \boldsymbol{\alpha} \right\}.
\end{aligned}$$

Using (4.27), we substitute $\boldsymbol{B}_n \boldsymbol{\alpha}$ for $\boldsymbol{a}_n$ to obtain:

$$\begin{aligned}
Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) = &-\frac{N-1}{2} \log \sigma_w^2 - \frac{N-1}{2} \log \sigma_v^2 - \frac{1}{2\sigma_v^2} \sum_{n=1}^{N-1} \left\{ y^2[n] - 2y[n] \mathbb{E}\left(x[n] | \boldsymbol{y}; \boldsymbol{\theta}^{(m)} \right) + \mathbb{E}\left(x^2[n] | \boldsymbol{y}; \boldsymbol{\theta}^{(m)} \right) \right\} \\
&- \frac{1}{2\sigma_w^2} \sum_{n=1}^{N-1} \left\{ \mathbb{E}\left(x^2[n] | \boldsymbol{y}; \boldsymbol{\theta}^{(m)} \right) - 2\boldsymbol{\alpha}^T \boldsymbol{B}_n^T \mathbb{E}\left(\boldsymbol{x}_{n-1} x[n] | \boldsymbol{y}; \boldsymbol{\theta}^{(m)} \right) \boldsymbol{\alpha}^T \boldsymbol{B}_n^T \mathbb{E}\left(\boldsymbol{x}_{n-1} \boldsymbol{x}_{n-1}^T | \boldsymbol{y}; \boldsymbol{\theta}^{(m)} \right) \boldsymbol{B}_n \boldsymbol{\alpha} \right\}.
\end{aligned} \tag{4.32}$$

The necessary expectations required in (4.32) can be obtained from the Kalman smoother of Algorithm 4.1 according to:

$$\mathbb{E}\left(\boldsymbol{x}_n | \boldsymbol{y}; \boldsymbol{\theta}^i \right) = \boldsymbol{m}_{n|N} \qquad \text{and} \qquad \mathbb{E}\left(\boldsymbol{x}_n \boldsymbol{x}_n^T | \boldsymbol{y}; \boldsymbol{\theta}^i \right) = \boldsymbol{P}_{n|N} + \boldsymbol{m}_{n|N} \boldsymbol{m}_{n|N}^T.$$

The M-step is obtained by maximizing (4.32) with respect to $\boldsymbol{\theta}$. The resultant

Algorithm 4.2 TVAR Estimation: Online System Identification via EM Algorithm

- Initialize $\widehat{\boldsymbol{\theta}}^{(1)} = (\boldsymbol{\alpha}^{(1)}, \sigma_w^{2\,(1)}, \sigma_v^{2(1)})$
- For $m = 1, \ldots, M$
  - E-Step: Compute $Q(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(m)})$
  - M-Step: Obtain the next estimate of $\boldsymbol{\theta}$ by using the estimators in (4.34) via:

$$\widehat{\boldsymbol{\theta}}^{(m+1)} = \sup_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(m)}) \tag{4.33}$$

- Return $\widehat{\boldsymbol{\theta}}^{(M)}$

parameter estimates are given by:

$$\begin{aligned}
\boldsymbol{\alpha}^{(m+1)} &= -\left(\sum_{n=1}^{N-1} \boldsymbol{B}_n^T \mathbb{E}\left(\boldsymbol{x}_{n-1}\boldsymbol{x}_{n-1}^T | \boldsymbol{y}; \boldsymbol{\theta}^{(m)}\right) \boldsymbol{B}_n\right)^{-1} \sum_{n=1}^{N-1} \boldsymbol{B}_n^T \mathbb{E}\left(\boldsymbol{x}_{n-1} | \boldsymbol{y}; \boldsymbol{\theta}^{(m)}\right) \\
\sigma_w^{2\,(m+1)} &= \frac{1}{N-1} \sum_{n=1}^{N-1} \Big(\mathbb{E}\left(x^2[n] | \boldsymbol{y}; \boldsymbol{\theta}^{(m)}\right) - 2\boldsymbol{\alpha}^{(m+1)^T} \boldsymbol{B}_n^T \mathbb{E}\left(\boldsymbol{x}_{n-1} x[n] | \boldsymbol{y}; \boldsymbol{\theta}^{(m)}\right) \\
&\qquad + \boldsymbol{\alpha}^{(m+1)^T} \boldsymbol{B}_n^T \mathbb{E}\left(\boldsymbol{x}_{n-1}\boldsymbol{x}_{n-1}^T | \boldsymbol{y}; \boldsymbol{\theta}^{(m)}\right) \boldsymbol{B}_n \boldsymbol{\alpha}^{(m+1)}\Big) \\
\sigma_v^{2(m+1)} &= \frac{1}{N-1} \sum_{n=1}^{N-1} \left(y^2[n] - 2y[n]\mathbb{E}\left(x[n] | \boldsymbol{y}; \boldsymbol{\theta}^{(m)}\right) + \mathbb{E}\left(x^2[n] | \boldsymbol{y}; \boldsymbol{\theta}^{(m)}\right)\right).
\end{aligned} \tag{4.34}$$

Compare the equation for $\boldsymbol{\alpha}^{(m+1)}$ with the generalized Yule-Walker estimate:

$$\widehat{\boldsymbol{\alpha}} \triangleq -\left(\sum_{n=1}^{N-1} \boldsymbol{B}_n^T \boldsymbol{x}_{n-1}\boldsymbol{x}_{n-1}^T \boldsymbol{B}_n\right)^{-1} \sum_{n=1}^{N-1} \boldsymbol{B}_n^T \boldsymbol{x}_{n-1} x[n], \tag{4.35}$$

for the TVAR coefficients obtained from $\boldsymbol{x}$ if it were available. The primary difference is that in lieu of $\boldsymbol{x}_{n-1}\boldsymbol{x}_{n-1}^T$ its posterior mean is used, since the time series $x[n]$ is unobserved.

In the time-variant AR setting, we have $\boldsymbol{a}_n = \boldsymbol{a}$ and can therefore rewrite the estimators of (4.34) according to:

$$\begin{aligned}
\boldsymbol{a}^{(m+1)} &= -\left(\sum_{n=1}^{N-1} \mathbb{E}\left(\boldsymbol{x}_{n-1}\boldsymbol{x}_{n-1}^T | \boldsymbol{y}; \boldsymbol{\theta}^{(m)}\right)\right)^{-1} \sum_{n=1}^{N-1} \mathbb{E}\left(\boldsymbol{x}_{n-1} x[n] | \boldsymbol{y}; \boldsymbol{\theta}^{(m)}\right) \\
\sigma_w^{(m+1)} &= \frac{1}{N-1} \sum_{n=1}^{N-1} \Big\{\mathbb{E}\left(x^2[n] | \boldsymbol{y}; \boldsymbol{\theta}^{(m)}\right) - 2\boldsymbol{a}^{(m+1)^T} \mathbb{E}\left(\boldsymbol{x}_{n-1} x[n] | \boldsymbol{y}; \boldsymbol{\theta}^{(m)}\right) \\
&\qquad + \boldsymbol{a}^{(m+1)^T} \mathbb{E}\left(\boldsymbol{x}_{n-1}\boldsymbol{x}_{n-1}^T | \boldsymbol{y}; \boldsymbol{\theta}^{(m)}\right) \boldsymbol{a}^{(m+1)}\Big\} \\
\sigma_v^{(i+1)} &= \frac{1}{N-1} \sum_{n=1}^{N-1} \left(y^2[n] - 2y[n]\mathbb{E}\left(x[n] | \boldsymbol{y}; \boldsymbol{\theta}^{(m)}\right) + \mathbb{E}\left(x^2[n] | \boldsymbol{y}; \boldsymbol{\theta}^{(m)}\right)\right).
\end{aligned}$$

These are the same estimators as those derived by [128, 129].

4.4 Stability

In this section, we discuss another TVAR estimation problem that arises in a variety of applications—that of estimating TVAR models which lead to "frozen-time stable" inverse-prediction filters. While it is well known that a *time-invariant* autoregressive process is stable if and only if the roots of its predictor polynomial lie inside the unit circle—a constraint implicitly enforced by many standard estimators—most estimation procedures for *time-varying* autoregressions yield processes whose instantaneous "poles" temporarily exit the unit circle. Here we consider a time-varying lattice formulation that enables us to appropriately constrain the temporal evolution of the associated reflection coefficients. The resultant estimators are given by the solution of a sequence of convex optimization problems and are illustrated using synthetic examples and speech signals.

Expanding the TVAR coefficient trajectories of (4.1) in a set of known basis functions allows one to obtain maximum likelihood estimates of the resultant expansion coefficients from a single time series of observations [58]. However, as is well known, the frozen-time "poles" of the estimated coefficient trajectories (i.e., roots of the prediction polynomials $1 + \sum_{i=1}^{p} \widehat{a}_i[n] z^{-i}$) may temporarily lie outside the unit circle. In this case, using an inverse-prediction-error filter for signal synthesis leads to large local amplitude fluctuations [95], which in the context of speech synthesis greatly reduces the quality of the synthesized waveform [95, 102]. Other applications relying on this "pointwise" stability include EEG analysis [111] and audio content retrieval [130].

To address this problem, we formulate a convex optimization approach to estimating instantaneously-stable inverse-prediction-error filters from time-series data. Our approach is to parameterize the TVAR model in lattice form, expanding the associated reflection coefficients in a functional basis, and appropriately constraining the expansion coefficients in the parameter estimation process. Evaluating the resultant estimators requires solving constrained convex optimization problems, which can be done efficiently using widely available software such as CVX [131]. In contrast, some existing approaches, such as reflecting any estimated poles of modulus greater than one about the unit circle to their reciprocal locations, do not preserve any optimality properties of the original estimator [37, 132, 133]. Others involve successive iterative linearization of nonlinear constraints [111] or maximizing a nonlinear, non-convex objective function [95], and fail to guarantee the existence of a global optimum.

First, in Section 4.4.1, we relate the model of (4.1) to time-varying lattice filters (TVLF) and review generalizations of lattice parameter estimation algorithms to the time-varying setting. In Section 4.4.2, we formalize the notion of frozen-time stability, present a novel algorithm for estimating frozen-time stable inverse-prediction-error lattice filters, relate our approach to an alternative method of [95], and discuss the case of unequal forward/backward reflection coefficients. We conclude with a number of illustrative examples in Section 4.4.3.

4.4.1 Time-Varying Lattice Filters

4.4.1.1 Orthogonal Realizations of Autoregressive Processes

We begin by relating time-varying lattice filters (see, e.g., Figure 4.4) to the TVAR process of (4.1). To this end, fix a positive integer j and define $\{a_{i,j}[n] \,|\, 1 \leq i \leq j\}$ and $\{b_{i,j}[n] \,|\, 1 \leq i \leq j\}$ to be the time-varying forward and backward linear prediction coefficients that minimize the squared errors of predicting $x[n]$ and $x[n-j]$, respectively:

$$e_j^{\mathrm{f}}[n] \triangleq x[n] - \sum_{i=1}^{j} a_{i,j}[n]x[n-i], \tag{4.36}$$

$$e_j^{\mathrm{b}}[n] \triangleq x[n-j] - \sum_{i=1}^{j} b_{i,j}[n]x[n-j+i]. \tag{4.37}$$

It is clear from (4.36) that $e_j^{\mathrm{f}}[n]$ is the error of approximating $x[n]$ by its projection onto the space spanned by $\{x[n-1], x[n-2], \ldots, x[n-j]\}$. However, this expansion (i.e., $\widehat{x}[n] \triangleq \sum_{i=1}^{j} a_{i,j}[n]x[n-i]$) is not orthogonal since $x[n]$ is an autoregressive process, which means that the variables $\{x[n-1], x[n-2], \ldots, x[n-j]\}$ in general fail to be independent.

An alternative realization of $\widehat{x}[n]$, using an *orthogonal* set of vectors, may be obtained by applying the Gram-Schmidt procedure to the variables $\{x[n-1], x[n-2], \ldots, x[n-j]\}$ in order starting from $x[n-1]$ (see, e.g., [27]). This orthogonalization yields, as a generalization of the Levinson recursion to the time-varying setting, a recursive way of computing the optimal forward and backward linear prediction coefficients of order j from those of order $j-1$ by:

$$a_{i,j}[n] = \begin{cases} a_{i,j-1}[n] + \kappa_j^{\mathrm{f}}[n] b_{j-i,j-1}[n-1] \text{ if } 1 \leq i < j \\ -\kappa_j^{\mathrm{f}}[n] \text{ if } i = j \end{cases} \tag{4.38}$$

$$b_{i,j}[n] = \begin{cases} b_{i,j-1}[n-1] + \kappa_j^{\mathrm{b}}[n] a_{j-i,j-1}[n] \text{ if } 1 \leq i < j \\ -\kappa_j^{\mathrm{b}}[n] \text{ if } i = j. \end{cases} \tag{4.39}$$

Here the forward and backward time-varying lattice coefficients $\kappa_j^{\mathrm{f}}[n]$ and $\kappa_j^{\mathrm{b}}[n]$, often called *reflection coefficients*, are defined via

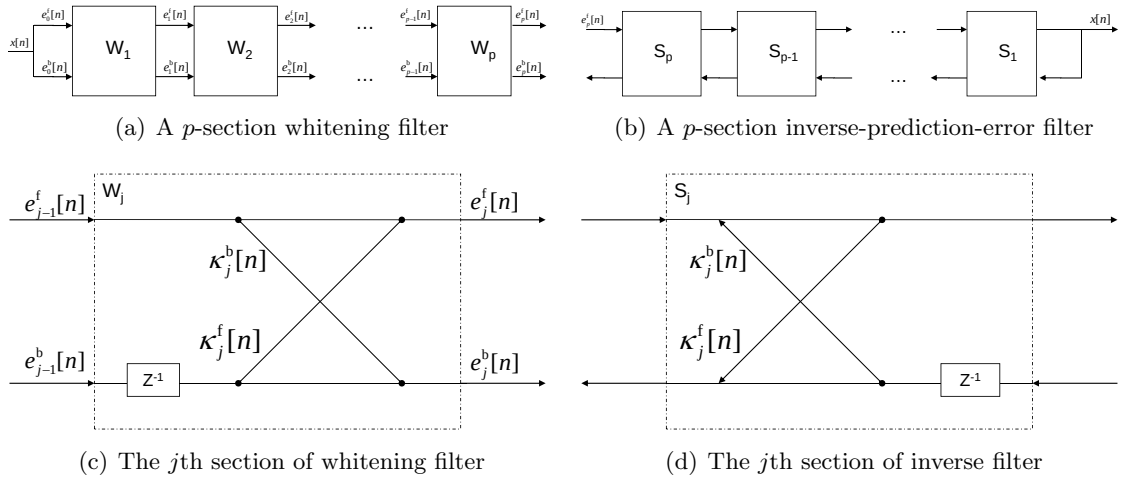
$$\kappa_j^{\mathrm{f}}[n] \triangleq -\frac{\langle e_{j-1}^{\mathrm{f}}[n], e_{j-1}^{\mathrm{b}}[n-1] \rangle}{\rho_{j-1}^{\mathrm{b}}[n-1]} \quad \text{and} \quad \kappa_j^{\mathrm{b}}[n] \triangleq -\frac{\langle e_{j-1}^{\mathrm{f}}[n], e_{j-1}^{\mathrm{b}}[n-1] \rangle}{\rho_{j-1}^{\mathrm{f}}[n]}, \tag{4.40}$$

with the error norms $\rho_j^{\mathrm{f}}[n]$, $\rho_j^{\mathrm{b}}[n]$ defined by:

$$\rho_j^{\mathrm{f}}[n] \triangleq \langle e_j^{\mathrm{f}}[n], e_j^{\mathrm{f}}[n] \rangle \qquad \text{and} \qquad \rho_j^{\mathrm{b}}[n] \triangleq \langle e_j^{\mathrm{b}}[n], e_j^{\mathrm{b}}[n] \rangle. \tag{4.41}$$

By substituting (4.38) and (4.39) into (4.36) and (4.37), respectively, we obtain the following familiar recursive lattice structure:

$$\begin{pmatrix} e_j^{\mathrm{f}}[n] \\ e_j^{\mathrm{b}}[n] \end{pmatrix} = \begin{pmatrix} 1 & \kappa_j^{\mathrm{f}}[n] \\ \kappa_j^{\mathrm{b}}[n] & 1 \end{pmatrix} \begin{pmatrix} e_{j-1}^{\mathrm{f}}[n] \\ e_{j-1}^{\mathrm{b}}[n-1] \end{pmatrix}. \tag{4.42}$$



(a) A p-section whitening filter

(b) A p-section inverse-prediction-error filter

(c) The jth section of whitening filter

(d) The jth section of inverse filter

Figure 4.4: Diagrams of whitening and inverse-prediction-error lattice filters with time-dependent forward and backward lattice coefficients $\kappa_j^{\mathrm{f}}[n]$ and $\kappa_j^{\mathrm{b}}[n]$, respectively; z^{-1} denotes a unit delay.

Indeed, together with the natural initial condition of $e_0^{\mathrm{f}}[n] = e_0^{\mathrm{b}}[n] = x[n]$, the recursion of (4.42) corresponds to the time-varying *whitening* lattice filter shown in Figure 4.4a. The original signal $x[n]$ may be recovered from the residual $e_p^{\mathrm{f}}[n]$ and the lattice coefficients using the *inverse-prediction-error* filter shown in Figure 4.4b.

By substituting (4.42) into (4.41), a recursive relationship may also be obtained for the error norms $\rho_j^{\mathrm{f}}[n]$ and $\rho_j^{\mathrm{b}}[n]$:

$$\rho_j^{\mathrm{f}}[n] = \rho_{j-1}^{\mathrm{f}}[n] - \left(\kappa_j^{\mathrm{f}}[n]\right)^2 \rho_{j-1}^{\mathrm{b}}[n-1] \quad \text{and} \quad \rho_j^{\mathrm{b}}[n] = \rho_{j-1}^{\mathrm{b}}[n-1] - \left(\kappa_j^{\mathrm{b}}[n]\right)^2 \rho_{j-1}^{\mathrm{f}}[n]. \tag{4.43}$$

Observe that in contrast to the stationary case, the forward and backward reflection coefficients are not necessarily equal to one another. If they were—implying that $\rho_j^{\mathrm{f}}[n] = \rho_j^{\mathrm{b}}[n-1]$—then the recursions of (4.43) would reduce to $\rho_j[n] = \rho_{j-1}[n](1-\kappa_j^2[n])$, and to the familiar $\rho_j = \rho_{j-1}(1-\kappa_j^2)$, in the time-invariant case.

4.4.1.2 Modeling Time Variation

We are interested in estimating lattice coefficients from a single time series of N observations of a TVAR(p) process $x[n]$. However, since the problem of estimating the $2pN$ lattice coefficients from only N observations is ill-posed, the temporal evolution of the lattice coefficients needs to be constrained. To this end, we follow [58, 95, 102, 132] and constrain each TVLF by assuming that each latticed coefficient trajectory can be written as a linear combination of $q+1$ functions $f_k[n]$ weighted by coefficients κ_{jk}^{f} and κ_{jk}^{b}:

$$\kappa_j^{\mathrm{f}}[n] = \sum_{k=0}^{q} \kappa_{jk}^{\mathrm{f}} f_k[n] \qquad \kappa_j^{\mathrm{b}}[n] = \sum_{k=0}^{q} \kappa_{jk}^{\mathrm{b}} f_k[n]. \tag{4.44}$$

In addition, we let the "constant" function $f_0[n] = 1$ be among the chosen set of functions, so that the classical time-invariant lattice filter is recovered whenever $\kappa_{jk} = 0$

for all $k > 0$. Define the vectors $\boldsymbol{\kappa}_j^{\mathrm{f}} \triangleq \begin{pmatrix} \kappa_{j0}^{\mathrm{f}} & \kappa_{j1}^{\mathrm{f}} & \cdots & \kappa_{jq}^{\mathrm{f}} \end{pmatrix}^T$ and $\boldsymbol{\kappa}_j^{\mathrm{b}} \triangleq \begin{pmatrix} \kappa_{j0}^{\mathrm{b}} & \kappa_{j1}^{\mathrm{b}} & \cdots & \kappa_{jq}^{\mathrm{b}} \end{pmatrix}^T$, then (4.44) implies that an order j whitening lattice filter is represented by $2j(q+1)$ parameters which we group via

$$\boldsymbol{\theta}_j \triangleq \left(\boldsymbol{\theta}_j^{\mathrm{f}} \quad | \quad \boldsymbol{\theta}_j^{\mathrm{b}}\right) = \left(\boldsymbol{\kappa}_1^{\mathrm{f}\,T} \;\; \boldsymbol{\kappa}_2^{\mathrm{f}\,T} \;\; \cdots \;\; \boldsymbol{\kappa}_j^{\mathrm{f}\,T} \quad | \quad \boldsymbol{\kappa}_1^{\mathrm{b}\,T} \;\; \boldsymbol{\kappa}_2^{\mathrm{b}\,T} \;\; \cdots \;\; \boldsymbol{\kappa}_j^{\mathrm{b}\,T}\right).$$

As we have seen earlier in this chapter, the functional expansion approach has also been used to model time-varying AR coefficients (see, e.g., [92,94,117]). However, note that the recursion of (4.38) implies that the model resulting from the functional expansions of lattice coefficients as in (4.44) is not the same as expanding TVAR coefficients via (4.3) in the same functional basis.

4.4.1.3 Unconstrained Estimation

In the absence of further constraints, time-varying reflection coefficients can be estimated directly from the data using a generalization of Burg's method [58]. This approach is structurally similar to the time-invariant case in that $\boldsymbol{\kappa}_j^{\mathrm{f}}$ and $\boldsymbol{\kappa}_j^{\mathrm{b}}$ can be obtained only after $\boldsymbol{\theta}_{j-1}$ is estimated. Suppose that $\boldsymbol{\theta}_{j-1}$, and consequently $e_{j-1}^{\mathrm{f}}[n]$ and $e_{j-1}^{\mathrm{b}}[n]$ are known, then we may estimate $\boldsymbol{\kappa}_j^{\mathrm{f}}$ and $\boldsymbol{\kappa}_j^{\mathrm{b}}$ as the minimizers of the sum of squared forward and backward prediction errors given by:

$$\widehat{\boldsymbol{\kappa}_j^{\mathrm{f}}} = \underset{\boldsymbol{\kappa}_j^{\mathrm{f}}}{\operatorname{argmin}}\; L^{\mathrm{f}}(\boldsymbol{\kappa}_j^{\mathrm{f}}; \boldsymbol{\theta}_{j-1}^{\mathrm{f}}) \triangleq \sum_{n=j}^{N-1} \|e_j^{\mathrm{f}}[n]\|^2 \tag{4.45}$$

$$\widehat{\boldsymbol{\kappa}_j^{\mathrm{b}}} = \underset{\boldsymbol{\kappa}_j^{\mathrm{b}}}{\operatorname{argmin}}\; L^{\mathrm{b}}(\boldsymbol{\kappa}_j^{\mathrm{b}}; \boldsymbol{\theta}_{j-1}^{\mathrm{b}}) \triangleq \sum_{n=j}^{N-1} \|e_j^{\mathrm{b}}[n]\|^2. \tag{4.46}$$

Note that $\boldsymbol{\kappa}_j^{\mathrm{f}}$ and $\boldsymbol{\kappa}_j^{\mathrm{b}}$ can be estimated independently of one another.

Define $\boldsymbol{f}[n] \triangleq \begin{pmatrix} f_0[n] & f_1[n] & \cdots & f_q[n] \end{pmatrix}^T$ and the auxiliary vectors $\boldsymbol{e}_j^{\mathrm{f}}$ and $\boldsymbol{e}_j^{\mathrm{b}}$ as

$$\boldsymbol{e}_j^{\mathrm{f}} \triangleq \begin{pmatrix} e_j^{\mathrm{f}}[j] & e_j^{\mathrm{f}}[j+1] & \cdots & e_j^{\mathrm{f}}[N-1] \end{pmatrix}^T \qquad \boldsymbol{e}_j^{\mathrm{b}} \triangleq \begin{pmatrix} e_j^{\mathrm{b}}[j-1] & e_j^{\mathrm{b}}[j] & \cdots & e_j^{\mathrm{b}}[N-2] \end{pmatrix}^T.$$

In addition, defining the auxiliary matrices $\boldsymbol{F}_j$ and $\boldsymbol{B}_j$ as

$$\boldsymbol{F}_j \triangleq \begin{pmatrix} e_j^{\mathrm{f}}[j] \otimes \boldsymbol{f}[j] \\ e_j^{\mathrm{f}}[j+1] \otimes \boldsymbol{f}[j+1] \\ \vdots \\ e_j^{\mathrm{f}}[N-1] \otimes \boldsymbol{f}[N-1] \end{pmatrix} \quad \text{and} \quad \boldsymbol{B}_j \triangleq \begin{pmatrix} e_j^{\mathrm{b}}[j-1] \otimes \boldsymbol{f}[j-1] \\ e_j^{\mathrm{b}}[j] \otimes \boldsymbol{f}[j] \\ \vdots \\ e_j^{\mathrm{b}}[N-2] \otimes \boldsymbol{f}[N-2] \end{pmatrix},$$

where $\otimes$ denotes the Kronecker product, allows us to rewrite $L^{\mathrm{f}}(\boldsymbol{\kappa}_j^{\mathrm{f}}; \boldsymbol{\theta}_{j-1}^{\mathrm{f}})$ and $L^{\mathrm{b}}(\boldsymbol{\kappa}_j^{\mathrm{b}}; \boldsymbol{\theta}_{j-1}^{\mathrm{b}})$ as follows:

$$\begin{aligned} L^{\mathrm{f}}(\boldsymbol{\kappa}_j^{\mathrm{f}}; \boldsymbol{\theta}_{j-1}^{\mathrm{f}}) &= \left(\boldsymbol{e}_{j-1}^{\mathrm{f}} + \boldsymbol{B}_{j-1}\boldsymbol{\kappa}_j^{\mathrm{f}}\right)^T \left(\boldsymbol{e}_{j-1}^{\mathrm{f}} + \boldsymbol{B}_{j-1}\boldsymbol{\kappa}_j^{\mathrm{f}}\right), \\ L^{\mathrm{b}}(\boldsymbol{\kappa}_j^{\mathrm{b}}; \boldsymbol{\theta}_{j-1}^{\mathrm{b}}) &= \left(\boldsymbol{e}_{j-1}^{\mathrm{b}} + \boldsymbol{F}_{j-1}\boldsymbol{\kappa}_j^{\mathrm{b}}\right)^T \left(\boldsymbol{e}_{j-1}^{\mathrm{b}} + \boldsymbol{F}_{j-1}\boldsymbol{\kappa}_j^{\mathrm{b}}\right). \end{aligned} \tag{4.47}$$

Minimizing (4.47) with respect to $\boldsymbol{\kappa}_j^{\mathrm{f}}$ and $\boldsymbol{\kappa}_j^{\mathrm{b}}$ then yields:

$$\widehat{\boldsymbol{\kappa}_j^{\mathrm{f}}} = \left(\boldsymbol{F}_{j-1}^T \boldsymbol{F}_{j-1}\right)^{-1} \boldsymbol{F}_{j-1}^T \boldsymbol{e}_{j-1}^{\mathrm{b}} \qquad \text{and} \qquad \widehat{\boldsymbol{\kappa}_j^{\mathrm{b}}} = \left(\boldsymbol{B}_{j-1}^T \boldsymbol{B}_{j-1}\right)^{-1} \boldsymbol{B}_{j-1}^T \boldsymbol{e}_{j-1}^{\mathrm{f}}. \tag{4.48}$$

Since $\boldsymbol{\theta}_{j-1}$ is unknown in practice, its estimate is used instead and $L^{\mathrm{f}}(\boldsymbol{\kappa}_j^{\mathrm{f}}; \widehat{\boldsymbol{\theta}}_{j-1}^{\mathrm{f}})$ and $L^{\mathrm{b}}(\boldsymbol{\kappa}_j^{\mathrm{b}}; \widehat{\boldsymbol{\theta}}_{j-1}^{\mathrm{b}})$ are minimized to obtain $\widehat{\boldsymbol{\kappa}_j^{\mathrm{f}}}$ and $\widehat{\boldsymbol{\kappa}_j^{\mathrm{b}}}$, respectively.

4.4.2 Stability-Constrained Estimation

Here, we first discuss a notion of stability—termed *frozen-time* stability—suitable to our time-varying setting. In particular, we constrain the estimators of (4.48) by requiring that the forward and backward reflection coefficients are equal ($\kappa_j^{\mathrm{f}}[n] = \kappa_j^{\mathrm{b}}[n]$), and bounded in magnitude by unity ($|\kappa_j[n]| < 1$). In addition, we relate our approach to that of [95], and discuss what happens in the case that the reflection coefficients are unequal ($\kappa_j^{\mathrm{f}}[n] \neq \kappa_j^{\mathrm{b}}[n]$).

4.4.2.1 Frozen-Time Stability

In the time-invariant setting, if the magnitude of the lattice coefficients is bounded from above by unity, then the associated inverse-prediction-error lattice filter (Figure 4.4b) is bounded-input bounded-output (BIBO) stable [37]. A natural generalization of this to the time-varying case is to require that the magnitudes of the reflection coefficients are bounded from above by unity at every time instant. In particular, note that (4.40) and the Cauchy-Schwartz inequality imply that $\kappa_j^{\mathrm{f}}[n]$ and $\kappa_j^{\mathrm{b}}[n]$ must have the same sign with their magnitudes constrained according to:

$$0 \leq \kappa_j^{\mathrm{f}}[n]\kappa_j^{\mathrm{b}}[n] = \frac{|\langle e_{j-1}^{\mathrm{f}}[n], e_{j-1}^{\mathrm{b}}[n-1]\rangle|^2}{\|e_{j-1}^{\mathrm{f}}[n]\|^2 \|e_{j-1}^{\mathrm{b}}[n-1]\|^2} < 1, \tag{4.49}$$

for all $1 \leq j \leq p$ and $0 \leq n \leq N-1$. Under the assumption that the forward and backward reflection coefficients are equal ($\kappa_j^{\mathrm{f}}[n] = \kappa_j^{\mathrm{b}}[n]$), the constraint of (4.49) reduces to $0 \leq \kappa_j^2[n] < 1$. This leads to the following natural definition of *frozen-time* stability.

Definition 1 (Frozen-Time Stability)**.** *An inverse-prediction-error lattice filter (Figure 4.4b) is called* frozen-time stable *if:*

$$|\kappa_j[n]| < 1. \tag{4.50}$$

for all $1 \leq j \leq p$ *and* $0 \leq n \leq N-1$.

The frozen-time stability condition of (4.50) has been widely used in practice [95, 110, 111, 134], and it reduces to the familiar constraint that all the lattice coefficients lie in the unit hypercube, in the stationary case. It is also possible to define frozen-time stability using (4.49) in lieu of (4.50) by relaxing the constraint that $\kappa_j^{\mathrm{f}}[n] = \kappa_j^{\mathrm{b}}[n]$, we discuss this in Section 4.4.2.4.

It is helpful to think of frozen-time stability as guaranteeing BIBO stability for each of the N time-invariant lattice filters induced by the trajectories of the reflection coefficients, rather than as a constraint on the asymptotic stability of the underlying TVAR

process $x[n]$. Indeed, the local constraints of (4.50) *do not* imply BIBO stability of the process. It can be shown, for instance, that the rapidly-varying TVAR process $x[n] = (-1)^n x[n-1] + .25x[n-2] + 1$ satisfies (4.50), but is not BIBO stable. Upper bounds on the speed of variation are, in fact, required to guarantee that frozen-time stability implies BIBO stability [135].

In this work, however, we are not interested in developing sufficient conditions under which local stability implies asymptotic stability. The signals motivating the development of our methods are produced by physical time-varying systems (e.g., speech or EEG), and the underlying system does not typically exhibit such exotic behavior. Moreover, in our *offline* time-varying setting, notions of asymptotic stability are not entirely appropriate, since the basis function expansion is only defined on the set of indices $\{0, 1, \ldots, N-1\}$. This is in contrast to *online* adaptive estimation approaches such as those described in [136], where it is sensible to study the asymptotic behavior of $x[n]$.

Further intuition about the constraints of (4.49) and (4.50) may be gained by considering what happens if $\kappa_j^{\mathrm{f}}[n] = \kappa_j^{\mathrm{b}}[n] = 1$ for some n. Define $\boldsymbol{\alpha}_{n,j} \triangleq \begin{pmatrix} 1 & \alpha_{1,j}[n] & \cdots & \alpha_{j,j}[n] \end{pmatrix}^T$ and $\boldsymbol{\beta}_{n,j} \triangleq \begin{pmatrix} 1 & \beta_{1,j}[n] & \cdots & \beta_{j,j}[n] \end{pmatrix}^T$, then (4.36) and (4.37) imply that (4.41) may be written according to:

$$\rho_j^{\mathrm{f}}[n] = \boldsymbol{\alpha}_{n,j}^T \boldsymbol{R}_{n,j} \boldsymbol{\alpha}_{n,j} \quad \text{and} \quad \rho_j^{\mathrm{b}}[n] = \boldsymbol{\beta}_{n,j}^T \boldsymbol{R}_{n,j} \boldsymbol{\beta}_{n,j}, \tag{4.51}$$

where $r[n,m] \triangleq \mathbb{E}(x[n]x[n-m])$ and $\boldsymbol{R}_{n,j}$ is the $(j+1) \times (j+1)$ autocorrelation matrix given by:

$$\boldsymbol{R}_{n,j} \triangleq \begin{pmatrix} r[n,0] & r[n,1] & \cdots & r[n,j] \\ r[n,1] & r[n-1,0] & \cdots & r[n-1,j-1] \\ \vdots & \vdots & \ddots & \vdots \\ r[n,j] & r[n-1,j-1] & \cdots & r[n-j,0] \end{pmatrix}.$$

It may be shown [136, Property 1] that if $\kappa_j^{\mathrm{f}}[n] = \kappa_j^{\mathrm{b}}[n] = 1$ then $\rho_j^{\mathrm{f}}[n] = \rho_j^{\mathrm{b}}[n] = 0$. This coupled with (4.51) implies that if $\kappa_j^{\mathrm{f}}[n] = \kappa_j^{\mathrm{b}}[n] = 1$, then the covariance matrix $\boldsymbol{R}_{n,j}$ is singular. Consequently, the *time-invariant* AR process parameterized by the coefficients $\{a_1[n], \ldots, a_p[n]\}$ is not BIBO stable [4].

4.4.2.2 Constrained Estimation

We now turn to the question of estimating the lattice coefficients under the frozen-time stability constraints of (4.50). Define $\boldsymbol{\kappa}_j \triangleq \boldsymbol{\kappa}_j^{\mathrm{f}} = \boldsymbol{\kappa}^{\mathrm{b}}$. Then the objective function of interest is given by

$$\begin{aligned} L(\boldsymbol{\kappa}_j \,;\, \boldsymbol{\theta}_{j-1}) = &\left(\boldsymbol{e}_{j-1}^{\mathrm{f}} + \boldsymbol{B}_{j-1}\boldsymbol{\kappa}_j\right)^T \left(\boldsymbol{e}_{j-1}^{\mathrm{f}} + \boldsymbol{B}_{j-1}\boldsymbol{\kappa}_j\right) \\ &+ \left(\boldsymbol{e}_{j-1}^{\mathrm{b}} + \boldsymbol{F}_{j-1}\boldsymbol{\kappa}_j\right)^T \left(\boldsymbol{e}_{j-1}^{\mathrm{b}} + \boldsymbol{F}_{j-1}\boldsymbol{\kappa}_j\right) \end{aligned}, \tag{4.52}$$

in lieu of the two separate objective functions in (4.47).

If $\kappa_{jk} = 0$ for all $k > 0$, then the estimator obtained by minimizing the objective function of (4.52) reduces to the Burg method, which guarantees $|\widehat{\kappa}_{j0}| < 1$, and in turn the BIBO stability of the associated inverse-prediction-error filter [37]. But in the general

Algorithm 4.3 Estimation of Stable Shaping Time-Varying Lattice Filter

1. Initialization: input waveform data $x[n]$
   - Set $e_0^{\mathrm{f}}[n] = e_0^{\mathrm{b}}[n] = x[n]$ for all $0 \leq n \leq N-1$
   - Fix $p, q > 0$, $\eta \in (0, 1]$ and select a set of functions $\{f_0[n], f_1[n], \ldots, f_q[n]\}$
2. For $j = 1 \ldots p$
   - Estimate $\boldsymbol{\kappa}_j$ via (4.53) and compute $\widehat{\kappa}_j[n]$ via (4.3)
   - Compute $e_j^{\mathrm{f}}[n]$ and $e_j^{\mathrm{b}}[n]$ via (4.42)
3. Return the coefficients $\{\widehat{\boldsymbol{\kappa}}_1, \widehat{\boldsymbol{\kappa}}_2, \ldots, \widehat{\boldsymbol{\kappa}}_p\}$

time-varying case, estimated lattice coefficient trajectories obtained by minimizing (4.52) do not automatically satisfy the frozen-time stability constraint of (4.50).

A key observation, however, is that the objective function of (4.52) is quadratic and the frozen-time stability constraints in (4.50) are linear in the coefficients $\boldsymbol{\kappa}_j$. Minimizing (4.52) subject to (4.50) for all $0 \leq n \leq N-1$ is, therefore, a *convex* optimization problem that has a unique global minimum (as long as $L(\boldsymbol{\kappa}_j; \boldsymbol{\theta}_{j-1})$ is bounded from below, and a feasible solution exists).

Suppose $\widehat{\boldsymbol{\theta}}_{j-1}$ is obtained so that $|\widehat{\kappa}_i[n]| < \eta \leq 1$ for all $1 \leq i < j$, all $0 \leq n \leq N-1$, and a constant $\eta \in (0, 1]$. Define $\boldsymbol{f}_k \triangleq \begin{pmatrix} f_k[0] & f_k[1] & \cdots & f_k[N-1] \end{pmatrix}^T$ and consider the following *constrained* estimator of $\boldsymbol{\kappa}_j$:

$$\begin{aligned} \widehat{\boldsymbol{\kappa}}_j = & \underset{\boldsymbol{\kappa}_j}{\operatorname{argmin}}\ L(\boldsymbol{\kappa}_j; \widehat{\boldsymbol{\theta}}_{j-1}), \\ \text{subject to} \quad & \begin{pmatrix} \boldsymbol{f}_0 & \boldsymbol{f}_1 & \cdots & \boldsymbol{f}_q \end{pmatrix} \boldsymbol{\kappa}_j < \boldsymbol{\eta}, \\ \text{and} \quad & -\begin{pmatrix} \boldsymbol{f}_0 & \boldsymbol{f}_1 & \cdots & \boldsymbol{f}_q \end{pmatrix} \boldsymbol{\kappa}_j < \boldsymbol{\eta}, \end{aligned} \tag{4.53}$$

where $\boldsymbol{\eta}$ is an $N \times 1$ vector with all entries equal to η.

The inequality constraints in (4.53) guarantee that $|\widehat{\boldsymbol{\kappa}_j}[n]| < \eta \leq 1$ for all $0 \leq n \leq N-1$. By setting $\eta = 1$ we enforce the frozen-time stability constraint of (4.50), whereas by setting $\eta < 1$ we enforce a more stringent constraint—that the estimated time-varying lattice coefficient trajectories are constrained to be within an origin-centered hypercube with sides of length η. This is sometimes referred to as η-hyperstability [111, 134], and is useful when applying time-varying lattice filtering to coding, since quantization errors increase as the magnitudes of the frozen-time reflection coefficients approach unity. All the lattice coefficients may be estimated iteratively by solving (4.53); the entire procedure is summarized in Algorithm 4.3.

The estimator of Algorithm 4.3 is obtained as a solution to a sequence of p *convex* optimization problems—each with a quadratic objective function subject to $2N$ linear constraints. While there is no general closed-form solution, software packages that include quadratic program solvers may be employed. Below we use the interior-point solver in CVX [131], initialized with the solution to the unconstrained problem given by (4.48).

4.4.2.3 Frozen-Time Stability via Log-Area Ratios: Comparison to the Approach of [95]

An alternative approach to estimating a frozen-time stable inverse-prediction-error filter is to first estimate time-varying log-area ratios, and to subsequently transform them into time-varying lattice coefficients [95]. In particular, the log-area ratios are related to the time-varying lattice coefficients by the one-to-one *nonlinear* mapping

$$\gamma_j[n] = \log\left(\frac{1+\kappa_j[n]}{1-\kappa_j[n]}\right) = -2\tanh^{-1}\left(-\kappa_j[n]\right). \tag{4.54}$$

Lattice coefficients can then be obtained from estimates of the log-area ratios via $\widehat{\kappa}_j[n] = -\tanh\left(\widehat{\gamma}_j[n]/2\right)$, which guarantees that the frozen-time stability constraint of (4.50) is satisfied.

Estimates of log-area ratios are obtained by regularizing their temporal trajectories via a functional expansion similar to that of (4.3) according to $\gamma_j[n] \triangleq \sum_{k=0}^{q} \gamma_{jk} f_k[n]$. Next, the parameters $\boldsymbol{\gamma}_j \triangleq \begin{pmatrix}\gamma_{j0} & \gamma_{j1} & \cdots & \gamma_{jq}\end{pmatrix}^T$ are estimated by solving a sequence of p *nonlinear* least-squares problems given by:

$$\widehat{\boldsymbol{\gamma}}_j = \underset{\boldsymbol{\gamma}_j}{\operatorname{argmin}} \sum_{n=j}^{N-1} \left(\|e_j^f[n]\|^2 + \|e_j^b[n]\|^2\right) \quad 1 \le j \le p. \tag{4.55}$$

The nonlinear dependence of (4.55) on $\boldsymbol{\kappa}_j$ is transparent when (4.42) and (4.54) are substituted into (4.55); consequently a Newton-Raphson procedure was employed in [95] to minimize (4.55).

In light of the nonlinearity, it is perhaps not surprising that the objective function of (4.55) is *nonconvex* as shown in Figure 4.5, in contrast to the convex formulations of (4.53) and (4.56). This may lead solvers to get stuck in local minima. Moreover, it is not easy to enforce η-hyperstability in this setting—the solution to (4.55) only guarantees frozen-time stability ($\eta = 1$). Finally, if the solver for (4.53) were initialized with the unconstrained estimate of (4.48), which happens to satisfy the frozen-time stability constraints, then this initial point is returned with no further (interior point) iterations required. On the other hand, the LAR estimator of (4.55) always requires multiple iterations since the error criterion is nonlinear.

4.4.2.4 Unequal Reflection Coefficients ($\kappa_j^{\mathrm{f}}[n] \neq \kappa_j^{\mathrm{b}}[n]$)

An enticing possibility is to use the constraint of (4.49), in place of (4.50), in order to define frozen-time stability without assuming that the forward and backward lattice coefficients are equal (i.e., that $\kappa^f[n] = \kappa^b[n]$). One attempt in this direction was made by Grenier [95], who estimated forward and backward time-varying LAR coefficients separately, requiring that the induced $\kappa^f[n]$ and $\kappa^b[n]$ are each less than 1 in magnitude, but this leaves open the possibility that they would have different signs, which (4.49) prohibits.

Clearly, enforcing (4.49) in lieu of (4.50) leads to a weaker definition of local stability. As we show in Section 4.4.3 below, one can construct (somewhat exotic) examples

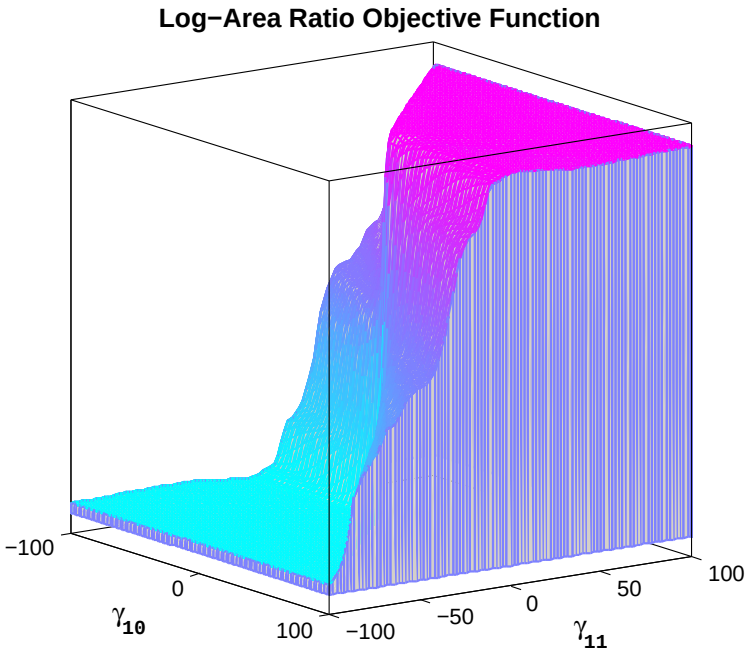


Figure 4.5: The objective function of (4.55) is nonconvex as shown via its pointwise evaluation over a range of values for γ_{10} and γ_{11} in an example with $p = 2$ and $q = 2$ Legendre polynomials.

to exploit this extra degree of freedom in order to create temporary instabilities. However, in practice when $\kappa^f[n]$ and $\kappa^b[n]$ are slowly time-varying using (4.49) can be quite sensible.

The salient point we wish to make, however, is that our framework can also be employed in this case, even though at first glance the problem appears non-convex due to the *hyperbolic* constraints of (4.49). Recall that according to (4.48) , $\widehat{\boldsymbol{\kappa}_j^{\mathrm{f}}}$ and $\widehat{\boldsymbol{\kappa}_j^{\mathrm{b}}}$ do not depend on one another. Thus, we may first estimate $\boldsymbol{\kappa}_j^{\mathrm{f}}$ using the unconstrained estimator of (4.48), and then estimate $\boldsymbol{\kappa}_j^{\mathrm{b}}$ subject to the constraint derived by substituting $\widehat{\boldsymbol{\kappa}_j^{\mathrm{f}}}$ into (4.49). The latter calculation, once again, involves minimizing a quadratic form subject to linear inequality constraints—a convex optimization problem with a unique solution.

To state this optimization problem precisely, we need to define the diagonal matrices $\boldsymbol{I}_j^+, \boldsymbol{I}_j^- \in \{0,1\}^{N\times N}$ whose entries depend on the indices on which $\widehat{\kappa_j^{\mathrm{f}}}[n]$ takes on positive and negative values as follows:

$$\boldsymbol{I}_j^+(m,n) \triangleq \begin{cases} 1 & \text{if } m = n \text{ and } \widehat{\kappa_j^{\mathrm{f}}}[n] \geq 0 \\ 0 & \text{otherwise} \end{cases} \qquad \boldsymbol{I}_j^-(m,n) \triangleq \begin{cases} 1 & \text{if } m = n \text{ and } \widehat{\kappa_j^{\mathrm{f}}}[n] < 0 \\ 0 & \text{otherwise} \end{cases}.$$

Next, fix a constant $\eta \in (0,1]$, define and the vector $\boldsymbol{\eta}_j \in \mathbb{R}^{N\times 1}$ entrywise by $\boldsymbol{\eta}_j[n] = \eta/\widehat{\kappa_j^{\mathrm{f}}}[n]$,

Algorithm 4.4 Shaping TVLF Estimation when $\kappa_j^{\mathrm{f}}[n] \neq \kappa_j^{\mathrm{b}}[n]$

1. Initialization: input waveform data $x[n]$
   - Set $e_0^{\mathrm{f}}[n] = e_0^{\mathrm{b}}[n] = x[n]$ for all $0 \leq n \leq N-1$
   - Fix $p, q > 0$, $\eta \in [0, 1)$ and select a set of functions $\{f_0[n], f_1[n], \ldots, f_q[n]\}$
2. For $j = 1 \ldots p$
   - Estimate $\boldsymbol{\kappa}_j^{\mathrm{f}}$ via (4.48) and compute $\widehat{\kappa}_j^{\mathrm{f}}[n]$ via (4.3)
   - Estimate $\boldsymbol{\kappa}_j^{\mathrm{b}}$ according to (4.56)
   - Compute $e_j^{\mathrm{f}}[n]$ and $e_j^{\mathrm{b}}[n]$ via (4.42)
3. Return the coefficients $\{\widehat{\boldsymbol{\kappa}_1^{\mathrm{f}}}, \widehat{\boldsymbol{\kappa}_2^{\mathrm{f}}}, \ldots, \widehat{\boldsymbol{\kappa}_p^{\mathrm{f}}} \,|\, \widehat{\boldsymbol{\kappa}_1^{\mathrm{b}}}, \widehat{\boldsymbol{\kappa}_2^{\mathrm{b}}}, \ldots, \widehat{\boldsymbol{\kappa}_p^{\mathrm{b}}}\}$

and consider the following convex optimization problem:

$$\begin{aligned} \widehat{\boldsymbol{\kappa}}_j^{\mathrm{b}} = & \underset{\boldsymbol{\kappa}_j^{\mathrm{b}}}{\operatorname{argmin}}\ L^{\mathrm{b}}(\boldsymbol{\kappa}_j^{\mathrm{b}}; \widehat{\boldsymbol{\theta}}_{j-1}^{\mathrm{f}}), \\ \text{subject to} \quad & \boldsymbol{I}_j^{+} \begin{pmatrix} \boldsymbol{f}_0 & \boldsymbol{f}_1 & \cdots & \boldsymbol{f}_q \end{pmatrix} \boldsymbol{\kappa}_j^{\mathrm{b}} < \boldsymbol{I}_j^{+} \boldsymbol{\eta}_j, \\ \text{and} \quad & -\boldsymbol{I}_j^{+} \begin{pmatrix} \boldsymbol{f}_0 & \boldsymbol{f}_1 & \cdots & \boldsymbol{f}_q \end{pmatrix} \boldsymbol{\kappa}_j^{\mathrm{b}} < \boldsymbol{0}, \\ \text{and} \quad & \boldsymbol{I}_j^{-} \begin{pmatrix} \boldsymbol{f}_0 & \boldsymbol{f}_1 & \cdots & \boldsymbol{f}_q \end{pmatrix} \boldsymbol{\kappa}_j^{\mathrm{b}} < \boldsymbol{0}, \\ \text{and} \quad & -\boldsymbol{I}_j^{-} \begin{pmatrix} \boldsymbol{f}_0 & \boldsymbol{f}_1 & \cdots & \boldsymbol{f}_q \end{pmatrix} \boldsymbol{\kappa}_j^{\mathrm{b}} < \boldsymbol{I}_j^{-} \boldsymbol{\eta}_j. \end{aligned} \tag{4.56}$$

The inequality constraints in (4.56) guarantee that $\left|\widehat{\kappa_j^{\mathrm{f}}}[n]\widehat{\kappa_j^{\mathrm{b}}}[n]\right| < \eta$ for all $0 \leq n \leq N-1$. As before, setting $\eta = 1$ enforces the frozen-time stability constraint of (4.49), while setting $\eta < 1$ enforces η-hyperstability. The full solution is obtained by iteratively solving for higher-order lattice section parameters using (4.56) and is summarized in Algorithm 4.4. As in the case of Algorithm 4.3, Algorithm 4.4 consists of solving a sequence of p *convex* optimization problems—the solution to each quadratic program may be obtained by an interior point solver such as CVX [131].

4.4.3 Examples

We illustrate the proposed methods through a number of examples using synthetic and speech waveforms.

4.4.3.1 Example 1: Piecewise-Constant Synthetic Signal

In the first example, similar to one discussed in [95], a piecewise constant TVAR(2) waveform was synthesized by filtering white Gaussian noise through a second-order digital resonator whose bandwidth decreased and center frequency shifted halfway through the duration of the signal. The true trajectory of the second reflection coefficient $\kappa_2[n]$ is shown in Figure 4.6 along with unconstrained estimators obtained via the generalized Yule-Walker

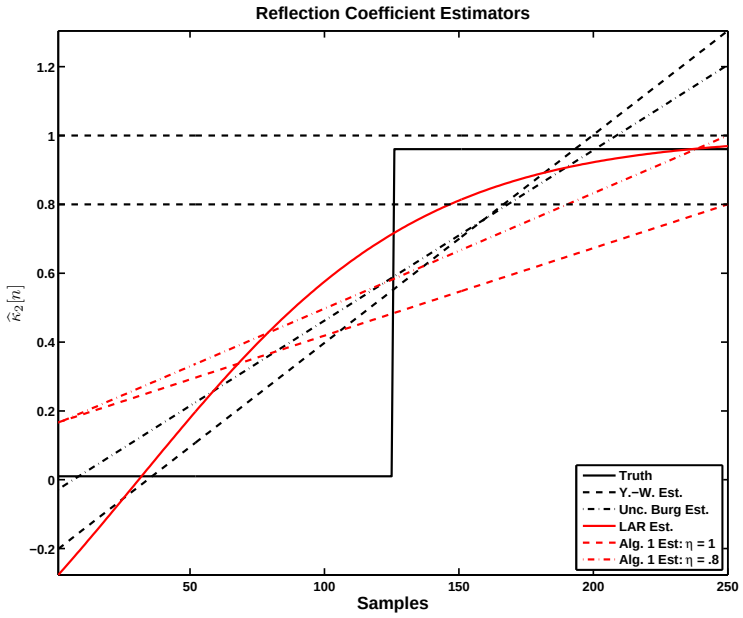


Figure 4.6: Fitting an order $p = 2$ TVLF to synthetic TVAR(2) data using a variety of unconstrained and stability constrained methods, all with two ($q = 1$) Legendre polynomials.

approach of [58], the unconstrained lattice estimator of (4.48), and the constrained estimators obtained by the LAR approach of [95] and Algorithm 4.3. Clearly, the reflection coefficient trajectories estimated by the unconstrained methods violate frozen-time stability, whereas both of the constrained approaches yield the desired result. The ability of Algorithm 4.3 to estimate η-hyperstable lattices is also shown.

4.4.3.2 Example 2: Temporarily-Unstable Process

Our second example highlights the difference between the frozen-time stability constraints of (4.49) and (4.50). Algorithms 4.3 and (4.4) were applied to a temporarily-unstable TVAR(2) process generated by filtering white Gaussian noise through a second-order digital resonator whose minimum-phase poles $(z; z^* \,|\, z = re^{j\theta}\,;\, r = .8, \theta = \pi/4)$ are briefly moved to their *conjugate-reciprocal* locations outside the unit circle resulting in the expected local amplitude spike seen in the top-left panel of Figure 4.7 (black line). The power spectral density of the signal remains unchanged, since reflecting poles about the boundary of the unit circle only changes the phase of the filter.

The bottom-left panel of Figure 4.7 shows the trajectories of the forward and backward lattice coefficients estimated by Algorithm 4.4; their product is also plotted in order to show that the constraint of (4.49) is satisfied. Note that the magnitudes of the forward and backward coefficients exceed unity in various parts of the waveform. It is this flexibility that allows the resynthesized signal, shown (red) in the top-left panel of Figure 4.7 to closely approximate the local instability in the original waveform. In contrast, as shown

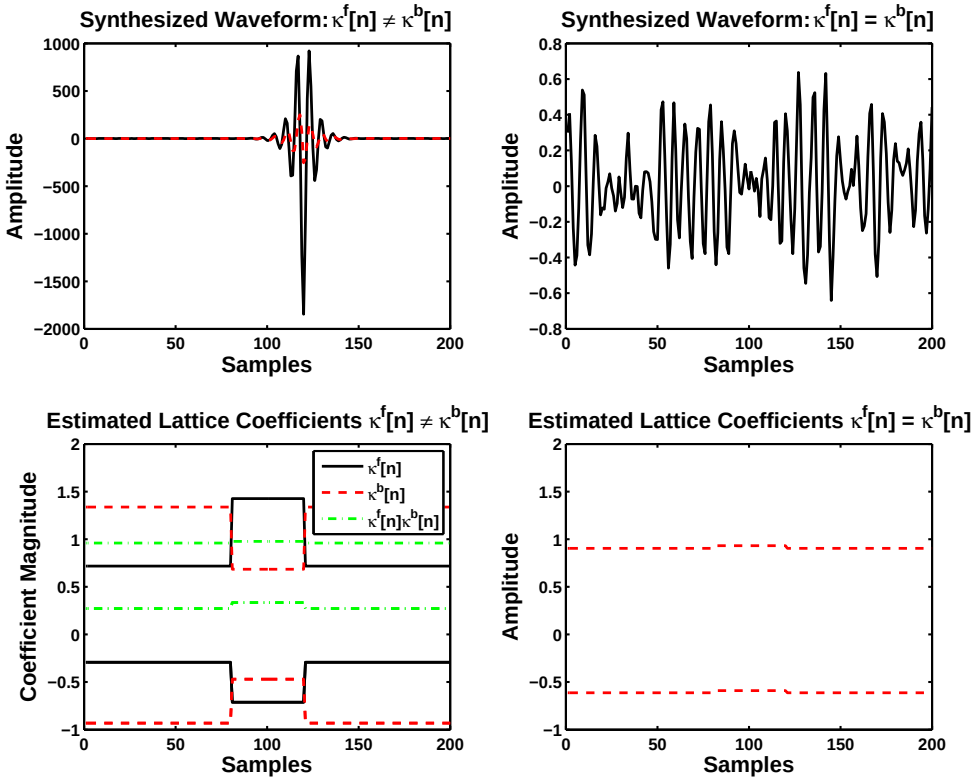


Figure 4.7: Fitting an order $p = 2$ TVLF to a synthetic, temporarily-unstable TVAR(2) process. The time-domain signal (solid, black) showing the temporary instability (top-left) is shown together with a signal (dashed, red) resynthesized using the forward (solid, black) and backward (dashed, red) lattice coefficients estimated using Algorithm 4.4 and shown in the bottom-left panel; the product coefficients $\kappa^{\mathrm{f}}[n] \cdot \kappa^{\mathrm{b}}$ (dashed-dot green) are also shown. The waveform (top-right) resynthesized by using lattice coefficients (bottom-left) estimated via Algorithm 4.3 is shown for comparison.

in the top- and bottom-right panels of Figure 4.7, applying Algorithm 4.3, which is based on the more restrictive condition of (4.50), yields a set of coefficients that do not lead to any local instabilities when used for resynthesis. To wit, the amplitude of the signal resynthesized from unequal forward/backward coefficients is several orders of magnitude larger than that of the signal resynthesized from lattice coefficients constrained to be equal during the estimation process.

4.4.3.3 Example 3: Speech Waveform

In the third example, shown in Figure 4.8, a ten-section ($p = 10$) time-varying lattice filter is fitted using the generalized covariance method of (4.17), and the stability-constrained estimator of Algorithm 4.3, to a portion of the phonetically-balanced TIMIT [48]

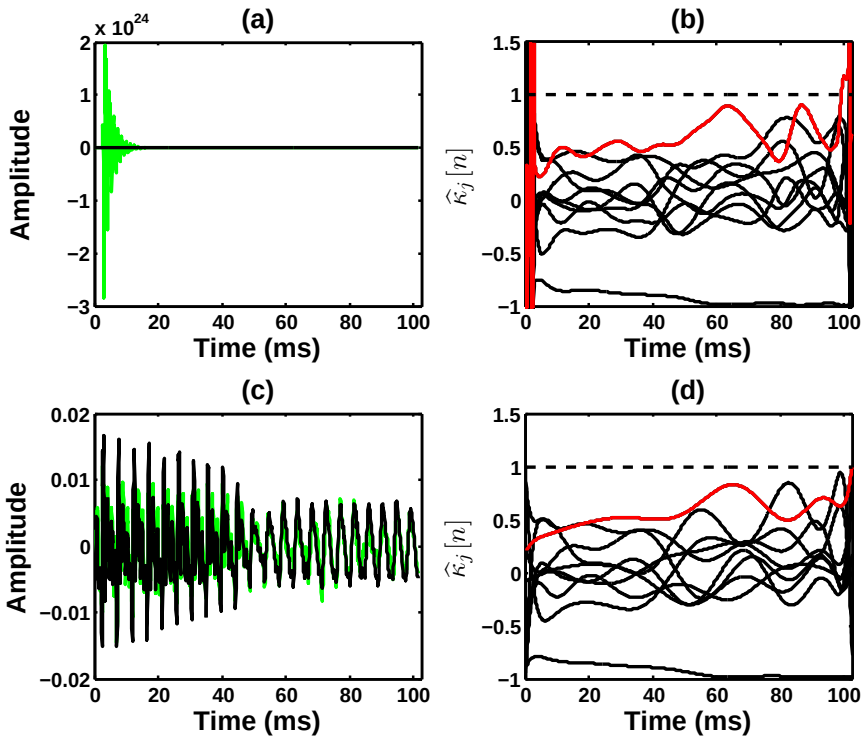


Figure 4.8: Fitting an order $p = 10$ time-varying lattice filter to a TIMIT speech waveform [ɛn] downsampled to 10 kHz. Top left: The original waveform (black) is overlaid with a version resynthesized (green) from lattice coefficients estimated using the generalized covariance method of (4.17) (top-right) using twelve ($q = 11$) Legendre polynomials. Bottom left: The original waveform (black) is overlaid with a version resynthesized (green) from lattice coefficients estimated using Algorithm 4.3 with twelve ($q = 11$) Legendre polynomials. The trajectory of one coefficient is highlighted in red in order to highlight the differences between estimators.

speech waveform "/train/dr1/fcjf0/si1027.wav" corresponding to the phrase "Even th[en] if she took. . . ." The temporary instabilities of the *unconstrained* fit at the left and right edges of the waveform, shown in the top-right panel of Figure 4.8, exemplify typical estimator behavior. Such instabilities tend to arise in frames that localize regions of rapid spectral change in the signal (e.g., during transitions from silence to voicing or from a narrowband to a wideband spectrum), and to occur at the frame boundaries.

Resynthesizing the speech signal using the estimated lattice coefficients results in the local instability shown in the top-left panel of Figure 4.8. On the other hand, as shown in the bottom-left and bottom-right panels of Figure 4.8, applying Algorithm 4.3 yields a frozen-time stable fit and the resynthesized signal closely tracks the original waveform.

4.5 Random Coefficient TVAR Models

As discussed in Section 4.1.3, different classes of time-varying autoregressive (TVAR) models can be obtained depending on how the temporal evolution of the AR coefficients

is modeled. Indeed, the functional basis expansion of (4.3) is only one of many possible approaches: two popular alternatives include modeling each coefficient trajectory as a sample path of an appropriately-chosen stochastic process [100, 118, 119, 121, 137, 138]) and implicitly regularizing each TVAR coefficient trajectory through the application of adaptive filtering methods, e.g., weighted and recursive least squares, for parameter estimation (see e.g., the text of [122] and references therein). In this section and the next, we describe two additional and novel formulations: random coefficient time-varying AR models and regularized AR models. Since neither approach has been previously studied, our focus here is on the formulation of the models, discussion of their properties, and the study of appropriate estimators.

4.5.1 Model Formulation

Here we introduce a new class of time-varying autoregressive models by merging the TVAR modeling approach of Section 4.1 together with the random coefficient autoregressive (RCA) models of [139, 140] to form a class of models we term time-varying random coefficient autoregressive models (TV-RCAR) defined according to the following discrete-time difference equation:

$$x[n] = \sum_{i=1}^{p} (a_i[n] + v_i[n])x[n-i] + w[n], \tag{4.57}$$

where the $a_i[n]$ are the time-varying autoregressive coefficients whose temporal evolution is modeled using a basis function expansion of (4.3), and $w[n]$ is white Gaussian noise with variance σ^2. The variable $\boldsymbol{v}_n \triangleq \begin{pmatrix} v_1[n] & v_2[n] & \cdots & v_p[n] \end{pmatrix}^T \in \mathbb{R}^{p\times 1}$ is a zero-mean Gaussian random vector with covariance matrix $\gamma^2 \boldsymbol{I}_{p\times p} \in \mathbb{R}^{p\times p}$, and is uncorrelated with the innovations sequence $w[n]$.

The model of (4.57) strikes a compromise between the functional expansion approach of (4.1), to which it reduces when $\gamma^2 = 0$, and the stochastic modeling approach where, for instance, each $a_i[n]$ evolves according to a random walk. We will see that parameter estimation for the TV-RCAR model can be realized using computationally-efficient maximum likelihood estimators, unlike the fully stochastic approach in which recursive Bayesian estimators based on the Kalman filter are required. On the other hand, the TV-RCAR model does require the judicious selection of a functional basis.

The model of (4.57) may be written in matrix-vector form according to:

$$x[n] = \boldsymbol{h}_n^T \boldsymbol{\alpha} + \boldsymbol{v}_n^T \boldsymbol{x}_{n-1} + w[n], \tag{4.58}$$

where $\boldsymbol{\alpha}$ is defined via (4.4), $\boldsymbol{x}_n \triangleq \begin{pmatrix} x[n-1] & x[n-2] & \cdots & x[n-p] \end{pmatrix}^T \in \mathbb{R}^{p\times 1}$ and $\boldsymbol{h}_n \triangleq \boldsymbol{x}_n \otimes \begin{pmatrix} f_0[n] & f_1[n] & \cdots & f_q[n] \end{pmatrix}^T \in \mathbb{R}^{(q+1)\times 1}$ with $\otimes$ denoting the Kronecker product. The model is parameterized by a total of $p(q+1)+2$ parameters $\{\boldsymbol{\alpha}, \gamma^2, \sigma^2\}$.

4.5.2 Maximum Likelihood Estimation

To derive an ML estimator for the model parameters, given a vector $\boldsymbol{x} \in \mathbb{R}^{N\times 1}$ of N observations, consider the joint probability density function of $\boldsymbol{\alpha}, \gamma^2$ and σ^2 given by:

$$p(\boldsymbol{x}\,;\,\boldsymbol{\alpha},\gamma^2,\sigma^2) = p(x[p],\ldots,x[N-1]\,|\,\boldsymbol{x}_p\,;\,\boldsymbol{\alpha},\gamma^2,\sigma^2)p(\boldsymbol{x}_p\,;\,\boldsymbol{\alpha},\gamma^2,\sigma^2). \tag{4.59}$$

As before, we approximate the *unconditional* data likelihood of (4.59) by the *conditional* likelihood $p(x[p],\ldots,x[N-1]\,|\,\boldsymbol{x}_p\,;\,\boldsymbol{\alpha},\gamma^2,\sigma^2)$, whose maximization yields an estimator that converges to the exact (unconditional) ML estimator as $N\to\infty$. The conditional likelihood can be factored according to:

$$p(x[p],\ldots,x[N-1]\,|\,\boldsymbol{x}_p\,;\,\boldsymbol{\alpha},\gamma^2,\sigma^2) = \prod_{k=p}^{N-1} p(x[k]\,|\boldsymbol{x}_k\,;\,\boldsymbol{\alpha},\gamma^2,\sigma^2). \tag{4.60}$$

Since $\boldsymbol{v}_n$ and $w[n]$ are uncorrelated zero-mean random variables, it follows that $p(x[k]\,|\,\boldsymbol{x}_k\,;\,\boldsymbol{\alpha},\gamma^2,\sigma^2)$ is a Gaussian density with mean $\boldsymbol{h}_k^T\boldsymbol{\alpha}$ and covariance:

$$\begin{aligned}\text{Var}\,(x[k]|\boldsymbol{x}_k) &= \mathbb{E}\left((x[k]-\boldsymbol{h}_k^T\boldsymbol{\alpha})^T(x[k]-\boldsymbol{h}_k^T\boldsymbol{\alpha})\,|\,\boldsymbol{x}_k\right) = \mathbb{E}\left((\boldsymbol{v}_k^T\boldsymbol{x}_{k-1}+w[k])^T(\boldsymbol{v}_k^T\boldsymbol{x}_{k-1}+w[k])\,|\,\boldsymbol{x}_k\right)\\ &= \mathbb{E}\left(\boldsymbol{x}_{k-1}^T\boldsymbol{v}_k\boldsymbol{v}_k^T\boldsymbol{x}_{k-1}\,|\,\boldsymbol{x}_k\right)+\sigma^2 = \gamma^2\boldsymbol{x}_{k-1}^T\boldsymbol{x}_{k-1}+\sigma^2,\end{aligned}$$

for each $p\le k\le N-1$. To obtain the ML estimators of $\boldsymbol{\alpha},\sigma^2$ and γ^2, we minimize the negative log-likelihood $l_N(\boldsymbol{\alpha},\sigma^2,\gamma^2)\triangleq -2/(N-p)\log p(x[p],\ldots,x[N-1]\,|\,\boldsymbol{x}_p\,;\,\boldsymbol{\alpha},\sigma^2,\gamma^2) - \log(2\pi)$ given by

$$l_N(\boldsymbol{\alpha},\gamma^2,\sigma^2) = \frac{1}{N-p}\sum_{k=p}^{N-1}\log\left(\gamma^2\boldsymbol{x}_{k-1}^T\boldsymbol{x}_{k-1}+\sigma^2\right)+\frac{1}{N-p}\sum_{k=p}^{N-1}\frac{\left(x[k]-\boldsymbol{h}_k^T\boldsymbol{\alpha}\right)^2}{\gamma^2\boldsymbol{x}_{k-1}^T\boldsymbol{x}_{k-1}+\sigma^2}. \tag{4.61}$$

It will be convenient to reparameterize $l_N(\boldsymbol{\alpha},\gamma^2,\sigma^2)$ as $l_N(\boldsymbol{\alpha},\delta^2,\sigma^2)$ where $\delta^2\triangleq\gamma^2/\sigma^2$ leading to:

$$l_N(\boldsymbol{\alpha},\sigma^2,\delta^2) = \log(\sigma^2)+\frac{1}{N-p}\sum_{k=p}^{N-1}\log\left(\gamma^2\boldsymbol{x}_{k-1}^T\boldsymbol{x}_{k-1}+1\right)+\frac{1}{(N-p)\sigma^2}\sum_{k=p}^{N-1}\frac{\left(x[k]-\boldsymbol{h}_k^T\boldsymbol{\alpha}\right)^2}{\gamma^2\boldsymbol{x}_{k-1}^T\boldsymbol{x}_{k-1}+1}. \tag{4.62}$$

Even though it is not possible to obtain simultaneous closed-form solutions for $\boldsymbol{\alpha}$, σ^2 and δ^2 by minimizing (4.62), estimators may be obtained by relying on numerical procedures. We present two such algorithms; both have counterparts in the time-invariant case [139].

One approach is to first minimize the negative log-likelihood of (4.62) conditional on δ^2 with respect to $\boldsymbol{\alpha}$ and σ^2 to obtain the following weighted-least-squares (WLS) estimators:

$$\begin{aligned}\widehat{\boldsymbol{\alpha}}_{\text{WLS}} &= \left(\sum_{k=p}^{N-1}\frac{\boldsymbol{h}_k\boldsymbol{h}_k^T}{\gamma^2\boldsymbol{x}_{k-1}^T\boldsymbol{x}_{k-1}+1}\right)^{-1}\sum_{k=p}^{N-1}\frac{x[k]\boldsymbol{h}_k}{\gamma^2\boldsymbol{x}_{k-1}^T\boldsymbol{x}_{k-1}+1},\\ \widehat{\sigma^2}_{\text{WLS}} &= \sum_{k=p}^{N-1}\frac{(x[k]-\boldsymbol{h}_k^T\boldsymbol{\alpha})^2}{\gamma^2\boldsymbol{x}_{k-1}^T\boldsymbol{x}_{k-1}+1}.\end{aligned} \tag{4.63}$$

These estimates may be plugged into the negative log-likelihood of (4.62) and an ML estimate for δ^2 may be obtained using a one-dimensional line search. Then ML estimates of $\boldsymbol{\alpha}$ and σ^2 may be obtained by substituting $\widehat{\delta^2}_{\text{ML}}$ into (4.63).

A second approach is to minimize the negative log-likelihood of (4.62) conditional on δ^2 and $\boldsymbol{\alpha}$ with respect to σ^2, to substitute the resultant expression (the same WLS estimator as in (4.63)) into (4.62), and to solve numerically for $\boldsymbol{\alpha}$ and δ^2. This approach may seem more computationally demanding than the first method, however, it is easy to compute the gradient of $l_N(\boldsymbol{\alpha}, \sigma^2, \delta^2)$ yielding an efficient search procedure that works better in practice. The necessary derivatives are given by:

$$\begin{aligned}
\frac{\partial l_N(\boldsymbol{\alpha}, \sigma^2, \delta^2)}{\partial \boldsymbol{\alpha}} &= -2 \sum_{k=p}^{N-1} \frac{(x[k] - \boldsymbol{h}_k^T \boldsymbol{\alpha}) \boldsymbol{h}_k}{1 + \gamma^2 \boldsymbol{x}_k^T \boldsymbol{x}_k} \left(\sum_{k=p}^{N-1} \frac{(x[k] - \boldsymbol{h}_k^T \boldsymbol{\alpha})^2}{1 + \gamma^2 \boldsymbol{x}_k^T \boldsymbol{x}_k} \right)^{-1} \\
\frac{\partial l_N(\boldsymbol{\alpha}, \sigma^2, \delta^2)}{\partial \delta^2} &= \sum_{k=p}^{N-1} \frac{\boldsymbol{x}_k^T \boldsymbol{x}_k}{1 + \gamma^2 \boldsymbol{x}_k^T \boldsymbol{x}_k} - \sum_{k=p}^{N-1} \frac{(x[k] - \boldsymbol{h}_k^T \boldsymbol{\alpha})^2}{(1 + \gamma^2 \boldsymbol{x}_k^T \boldsymbol{x}_k)^2} \left(\sum_{k=p}^{N-1} \frac{(x[k] - \boldsymbol{h}_k^T \boldsymbol{\alpha})^2}{(1 + \gamma^2 \boldsymbol{x}_k^T \boldsymbol{x}_k)} \right)^{-1}.
\end{aligned} \tag{4.64}$$

Since the calculation of both ML estimators involves numerical search, it is useful to initialize the estimators using least-squares (LS) estimators of all the parameters. Taking the appropriate derivatives leads to the following least-squares estimators of $\boldsymbol{\alpha}$, γ^2, and σ^2:

$$\begin{aligned}
\widehat{\boldsymbol{\alpha}}_{\text{LS}} &= \left(\sum_{k=p}^{N-1} \boldsymbol{h}_k \boldsymbol{h}_k^T \right)^{-1} \sum_{k=p}^{N-1} x[k] \boldsymbol{h}_k, \\
\widehat{\gamma^2}_{\text{LS}} &= \frac{\sum_{k=p}^{N-1} (x[k] - \boldsymbol{h}_k^T \widehat{\boldsymbol{\alpha}}_{\text{LS}})^2 (\boldsymbol{x}_k^T \boldsymbol{x}_k - z)}{\sum_{k=p}^{N-1} (\boldsymbol{x}_k^T \boldsymbol{x}_k)^2 - \frac{z^2}{N-p}} \\
\widehat{\sigma^2}_{\text{LS}} &= \frac{1}{N-p} \sum_{k=p}^{N-1} (x[k] - \boldsymbol{h}_k^T \widehat{\boldsymbol{\alpha}}_{\text{LS}})^2 - \widehat{\gamma^2}_{\text{LS}} z,
\end{aligned} \tag{4.65}$$

where $z = \frac{1}{N-p} \sum_{k=p}^{N-1} \boldsymbol{x}_k^T \boldsymbol{x}_k$). An initial estimate of δ^2 may be obtained as $\widehat{\gamma^2}_{\text{LS}} / \widehat{\sigma^2}_{\text{LS}}$.

An example of applying both ML estimators developed in this section to two different order 4 TV-RCAR processes synthesized using a sinusoidal basis ($q = 1$) is shown in Figure 4.9. The time-domain instantiations are shown in Figure 4.9(a) and Figure 4.9(d) along with the associated coefficient trajectories. Maximum likelihood estimates of their means obtained by the two approaches described above with the associated optimization procedures, both initialized via the least-squares estimators of (4.65), are also shown. As can be seen in Figure 4.9(c), both estimators do an excellent job estimating the time-varying mean ($a_i[n]$) of each AR coefficient in the first case. The second example, however, contains a temporary instability—the amplitude of the waveform temporarily spikes. In this case, though both algorithms were identically initialized, the approach based on a gradient-descent search using (4.64) is able to converge and obtain reasonable estimates, while the first approach gets stuck in a local minimum as shown in Figure 4.9(c).

We note that the larger the perturbation variance γ^2, the more beneficial it is to use the TV-RCAR models of (4.57) and associated estimators relative to their counterparts in the TVAR case of Section 4.1.

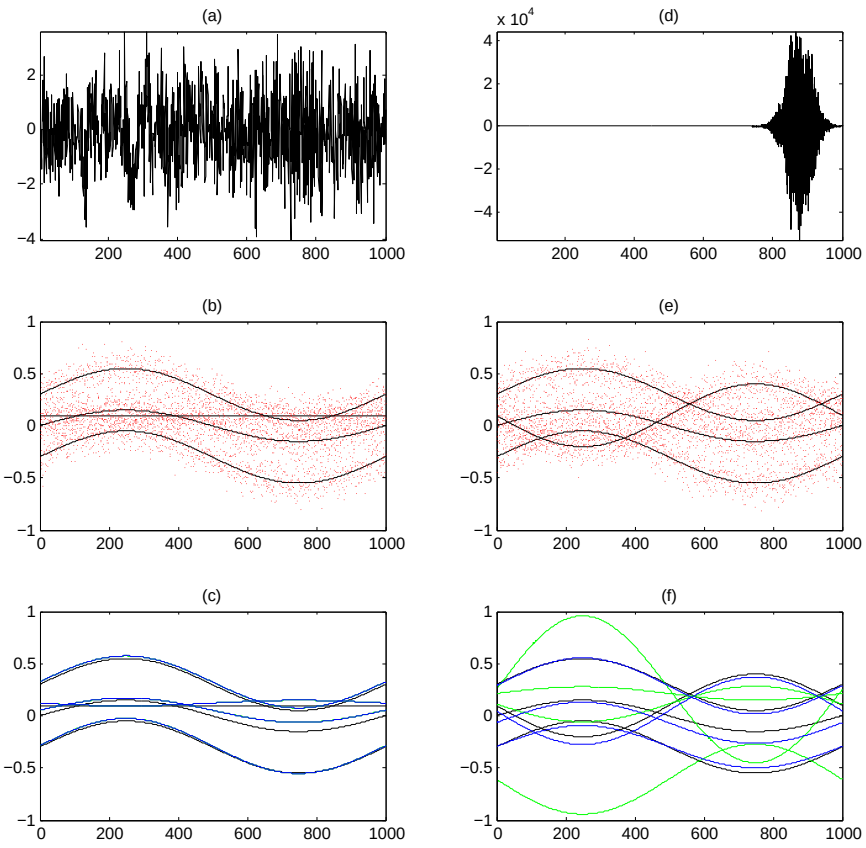


Figure 4.9: Example of two TV-RCAR processes and the performance of the maximum-likelihood estimators. The time-domain waveform is shown in panels (a) and (d), and the time-varying means of ($a_i[n]$, black line) of each AR coefficient are shown together with true AR coefficient trajectories ($a_i[n] + v[n]$, red dots) in panels (b) and (e). The first (green line) and second (blue line) ML estimates of $a_i[n]$ are overlaid on the true values of $a_i[n]$ in panels (c) and (f).

4.6 Regularized TVAR Models

We now consider a very different approach to TVAR modeling, which we motivate from a geometric point of view. The TVAR coefficients at time instant n $\{a_1[n], \ldots, a_p[n], \sigma_n^2\}$ implicitly define a "frozen-time" AR process with power spectral density f_n. In this sense the TVAR coefficient trajectories define a path on the space of AR processes, and the "length" of this path is a distinguishing characteristic of each process.

This observation forms the basis of our approach to modeling the temporal evolution of TVAR coefficients—we constrain the coefficient variation from one timestep to the next by bounding a measure of distance between induced AR models along the TVAR path. We show that this formulation admits efficient estimators realized through convex optimization programs. Aspects of our approach are similar to the recent work of [141] in the context of signal segmentation, however, our presentation is more general touching upon lattice parameterizations, tests for stationarity, and the relationship of our distance

measure to Riemannian metrics intrinsic to the manifold of AR processes.

To begin we consider the models of (4.1) and (4.42), and *do not assume* that either the time-varying coefficients $a_i[n]$ or the time-varying lattice coefficients $\kappa_i^{\mathrm{f}}[n]$ $\kappa_i^{\mathrm{b}}[n]$ have been expanded in a functional basis. After discussing parameter estimation for these unconstrained models in Section 4.6.1, we narrow the set of nonstationary processes under consideration by regularizing the temporal trajectories of the TVAR and lattice coefficients using local and global constraints, and derive the corresponding constrained estimators in Section 4.6.2. Next, we show how to use these estimators to construct hypothesis tests for stationarity in Section 4.6.3. Finally, in Section 4.6.4, we contrast the distance measures employed in our regularization framework with distance measures arising from Riemannian metrics on the space of AR processes.

4.6.1 Unconstrained Estimation

When a single sequence of N observations is available for estimating $(N-p)p$ parameters of either (4.1) or (4.42), the problem is ill-posed without further constraints. We describe appropriately constrained estimators in Section 4.6.2 below; in preparation, we first develop the unconstrained estimators in this section.

4.6.1.1 Time-Varying Autoregressive Coefficients

Given a vector $\boldsymbol{x} \in \mathbb{R}^{N\times 1}$ of N observations of the process $x[n]$, partitioned as

$$\boldsymbol{x} = (\boldsymbol{x}_p \,|\, \boldsymbol{x}_{N-p})^T \triangleq \big(x[0] \cdots x[p-1] \;\mid\; x[p] \cdots x[N-1]\big)^T,$$

we wish to estimate a vector $\boldsymbol{a} \in \mathbb{R}^{(N-p)p\times 1}$ of TVAR coefficients grouped via:

$$\boldsymbol{a} \triangleq \begin{pmatrix} \boldsymbol{a}_p & \boldsymbol{a}_{p+1} & \cdots & \boldsymbol{a}_{N-1} \end{pmatrix}^T,$$

where $\boldsymbol{a}_n \triangleq \begin{pmatrix} a_1[n] & a_2[n] & \cdots & a_p[n] \end{pmatrix}$ for $p \le n \le N-1$. The unconditional likelihood of the TVAR coefficients and σ^2 can be factored according to:

$$p(\boldsymbol{x}\,;\, \boldsymbol{a}_0, \ldots, \boldsymbol{a}_{N-1}, \sigma^2) = p(\boldsymbol{x}_{N-p} \,|\, \boldsymbol{x}_p\,;\, \boldsymbol{a}, \sigma^2) p(\boldsymbol{x}_p; \boldsymbol{a}_0, \ldots, \boldsymbol{a}_{p-1}, \sigma^2).$$

As before, we approximate the above unconditional data likelihood by the conditional likelihood $p(\boldsymbol{x}_{N-p} \,|\, \boldsymbol{x}_p\,;\, \boldsymbol{a}, \sigma^2)$. Gaussianity of $w[n]$ implies that maximizing the conditional likelihood is equivalent to solving the following least-squares problem:

$$\widehat{\boldsymbol{a}} = \underset{\boldsymbol{a}}{\text{minimize}}\; \|\boldsymbol{x}_{N-p} - \boldsymbol{H}_{\boldsymbol{x}}\boldsymbol{a}\|_2^2, \tag{4.66}$$

where $\boldsymbol{H}_{\boldsymbol{x}} \in \mathbb{R}^{(N-p)\times(N-p)p}$ is the appropriate data-dependent block-Hankel matrix. Its solution is readily obtained as $\widehat{\boldsymbol{a}} = \left(\boldsymbol{H}_{\boldsymbol{x}}^T \boldsymbol{H}_{\boldsymbol{x}}\right)^{-1} \boldsymbol{H}_{\boldsymbol{x}}^T \boldsymbol{x}_{N-p}$.

4.6.1.2 Time-Varying Reflection Coefficients

The time-varying reflection coefficients may also be directly estimated from data. Specifically, given N observations, we need to estimate the parameter vectors $\{\boldsymbol{\kappa}_1^{\mathrm{f}}, \boldsymbol{\kappa}_2^{\mathrm{f}}, \ldots, \boldsymbol{\kappa}_p^{\mathrm{f}}\}$ and $\{\boldsymbol{\kappa}_1^{\mathrm{b}}, \boldsymbol{\kappa}_2^{\mathrm{b}}, \ldots, \boldsymbol{\kappa}_p^{\mathrm{b}}\}$ where $\boldsymbol{\kappa}_j^{\mathrm{f}}, \boldsymbol{\kappa}_j^{\mathrm{b}} \in \mathbb{R}^{(N-j)\times 1}$ are defined for all $1 \leq j \leq p$ by

$$\boldsymbol{\kappa}_j^{\mathrm{f}} \triangleq \begin{pmatrix} \kappa_j^{\mathrm{f}}[j] & \kappa_j^{\mathrm{f}}[j+1] & \cdots & \kappa_j^{\mathrm{f}}[N-1] \end{pmatrix}^T, \quad \text{and} \quad \boldsymbol{\kappa}_j^{\mathrm{b}} \triangleq \begin{pmatrix} \kappa_j^{\mathrm{b}}[j] & \kappa_j^{\mathrm{b}}[j+1] & \cdots & \kappa_j^{\mathrm{b}}[N-1] \end{pmatrix}^T.$$

We may group these first j sets of coefficients according to $\boldsymbol{\theta}_j^{\mathrm{f}} \triangleq \{\boldsymbol{\kappa}_1^{\mathrm{f}}, \boldsymbol{\kappa}_2^{\mathrm{f}}, \ldots, \boldsymbol{\kappa}_j^{\mathrm{f}}\}$ and $\boldsymbol{\theta}_j^{\mathrm{b}} \triangleq \{\boldsymbol{\kappa}_1^{\mathrm{b}}, \boldsymbol{\kappa}_2^{\mathrm{b}}, \ldots, \boldsymbol{\kappa}_j^{\mathrm{b}}\}$.

The estimation approach, structurally similar to Burg's algorithm in the time-invariant case, is to estimate $\widehat{\boldsymbol{\kappa}}_j^{\mathrm{f}}$ and $\widehat{\boldsymbol{\kappa}}_j^{\mathrm{b}}$ only after $\widehat{\boldsymbol{\theta}}_{j-1}^{f}$ and $\widehat{\boldsymbol{\theta}}_{j-1}^{b}$ are obtained. In particular, suppose that $\boldsymbol{\theta}_{j-1}$ and consequently $e_{j-1}^{\mathrm{f}}[n]$ and $e_{j-1}^{\mathrm{b}}[n]$ are known. Then, we wish to find estimates of $\boldsymbol{\kappa}_j^{\mathrm{f}}$ and $\boldsymbol{\kappa}_j^{\mathrm{b}}$ that minimize, respectively, the sums of squared forward and backward prediction errors:

$$\widehat{\boldsymbol{\kappa}}_j^{\mathrm{f}} = \underset{\boldsymbol{\kappa}_j^{\mathrm{f}}}{\operatorname{argmin}}\ L(\boldsymbol{\kappa}_j^{\mathrm{f}}; \boldsymbol{\theta}_{j-1}^{\mathrm{f}}) \triangleq \sum_{n=j-1}^{N-1} \|e_j^{\mathrm{f}}[n]\|^2 \qquad \widehat{\boldsymbol{\kappa}}_j^{\mathrm{b}} = \underset{\boldsymbol{\kappa}_j^{\mathrm{b}}}{\operatorname{argmin}}\ L(\boldsymbol{\kappa}_j^{\mathrm{b}}; \boldsymbol{\theta}_{j-1}^{\mathrm{b}}) \triangleq \sum_{n=j-1}^{N-1} \|e_j^{\mathrm{b}}[n]\|^2. \tag{4.67}$$

The objective function of (4.67) may be conveniently rewritten in the form of a linear regression. To this end, define the vectors $\boldsymbol{e}_j^{\mathrm{f}}$ and $\boldsymbol{e}_j^{\mathrm{b}} \in \mathbb{R}^{(N-j)\times 1}$ according to:

$$\boldsymbol{e}_j^{\mathrm{f}} \triangleq \begin{pmatrix} e_j^{\mathrm{f}}[j] & e_j^{\mathrm{f}}[j+1] & \cdots & e_j^{\mathrm{f}}[N-1] \end{pmatrix}^T \quad \text{and} \quad \boldsymbol{e}_j^{\mathrm{b}} \triangleq \begin{pmatrix} e_j^{\mathrm{b}}[j-1] & e_j^{\mathrm{b}}[j] & \cdots & e_j^{\mathrm{b}}[N-2] \end{pmatrix}^T,$$

and the matrices $\boldsymbol{E}_j^{\mathrm{f}}, \boldsymbol{E}_j^{\mathrm{b}} \in \mathbb{R}^{(N-j)\times(N-j)}$ specified entrywise by: $\boldsymbol{E}^{\mathrm{f}}(m,n) \triangleq e_j^{\mathrm{f}}[m]\delta[m-n]$ and $\boldsymbol{E}^{\mathrm{b}}(m,n) \triangleq e_j^{\mathrm{b}}[m]\delta[m-n]$ for $1 \leq m, n \leq N-j$. Then we may write (4.67) via:

$$\widehat{\boldsymbol{\kappa}}_j^{\mathrm{f}} = \underset{\boldsymbol{\kappa}_j^{\mathrm{f}}}{\operatorname{argmin}}\ L(\boldsymbol{\kappa}_j^{\mathrm{f}}; \boldsymbol{\theta}_{j-1}^{\mathrm{f}}) = \left(\boldsymbol{e}_{j-1}^{\mathrm{f}} + \boldsymbol{E}_{j-1}^{\mathrm{b}}\boldsymbol{\kappa}_j^{\mathrm{f}}\right)^T \left(\boldsymbol{e}_{j-1}^{\mathrm{f}} + \boldsymbol{E}_{j-1}^{\mathrm{b}}\boldsymbol{\kappa}_j^{\mathrm{f}}\right),$$
$$\widehat{\boldsymbol{\kappa}}_j^{\mathrm{b}} = \underset{\boldsymbol{\kappa}_j^{\mathrm{b}}}{\operatorname{argmin}}\ L(\boldsymbol{\kappa}_j^{\mathrm{b}}; \boldsymbol{\theta}_{j-1}^{\mathrm{b}}) = \left(\boldsymbol{e}_{j-1}^{\mathrm{b}} + \boldsymbol{E}_{j-1}^{\mathrm{f}}\boldsymbol{\kappa}_j^{\mathrm{b}}\right)^T \left(\boldsymbol{e}_{j-1}^{\mathrm{b}} + \boldsymbol{E}_{j-1}^{\mathrm{f}}\boldsymbol{\kappa}_j^{\mathrm{b}}\right).$$

Consequently, the estimators are given by:

$$\widehat{\boldsymbol{\kappa}}_j^{\mathrm{f}} = -\left({\boldsymbol{E}_{j-1}^{\mathrm{b}}}^T \boldsymbol{E}_{j-1}^{\mathrm{b}}\right) \boldsymbol{e}_{j-1}^{\mathrm{f}}, \qquad \text{and} \qquad \widehat{\boldsymbol{\kappa}}_j^{\mathrm{b}} = -\left({\boldsymbol{E}_{j-1}^{\mathrm{f}}}^T \boldsymbol{E}_{j-1}^{\mathrm{f}}\right) \boldsymbol{e}_{j-1}^{\mathrm{b}}, \tag{4.68}$$

or pointwise according to:

$$\widehat{\kappa}_j^{\mathrm{f}}[n] = -\frac{e_{j-1}^{\mathrm{f}}[n]e_{j-1}^{\mathrm{b}}[n-1]}{e_{j-1}^{\mathrm{b}}[n]e_{j-1}^{\mathrm{b}}[n]}, \qquad \text{and} \qquad \widehat{\kappa}_j^{\mathrm{b}}[n] = -\frac{e_{j-1}^{\mathrm{f}}[n]e_{j-1}^{\mathrm{b}}[n-1]}{e_{j-1}^{\mathrm{f}}[n]e_{j-1}^{\mathrm{f}}[n]},$$

which may be viewed as a simple plug-in estimator based on (4.40). This is the best estimator available in the absence of further assumptions.

4.6.2 Constrained Estimation

Since estimating $(N-p)p$ parameters from N observations is an ill-posed problem, the estimators of (4.66) and (4.67) need to be constrained. Here we discuss two strategies based, respectively, on a set of *local* and *global* constraints on how fast the autoregressive or lattice coefficients are allowed to vary.

4.6.2.1 Time-Varying Autoregressive Coefficients

It is possible to constrain the speed of temporal variation of a TVAR process, locally, by bounding finite-difference approximations to the first d derivatives of each TVAR coefficient trajectory. In the case of $d = 1$, for instance, this yields the following convex optimization problem:

$$\begin{aligned} \widehat{\boldsymbol{a}} = \underset{\boldsymbol{a}}{\operatorname{argmin}} \; & \|\boldsymbol{x}_{N-p} - \boldsymbol{H}_{\boldsymbol{x}}\boldsymbol{a}\|_2^2 \\ \text{subject to} \quad & |a_i[n] - a_i[n-1]| \leq \epsilon \quad \forall\; 1 < n \leq N;\; 1 \leq i \leq p, \\ & \epsilon \geq 0. \end{aligned} \tag{4.69}$$

Note that a time-invariant AR process may be recovered by setting $\epsilon = 0$, in which case the estimator of (4.69) reduces to the covariance method of linear prediction.

The *local* constraints of (4.69) preclude large local fluctuations since the amount of coefficient variation per timestep is bounded. An alternative is to consider a *global* regularizer based on the total "length" of all AR trajectories captured via a q-norm-derived distance as follows:

$$L_q(\boldsymbol{a}) \triangleq \left(\sum_{i=1}^{p} \sum_{n=p+1}^{N-1} |a_i[n] - a_i[n-1]|^q \right)^{1/q}. \tag{4.70}$$

If one were to think of a TVAR process as a path on the manifold of AR processes, with each frozen-time AR process as a point along the path, then (4.70) serves as proxy for measuring the path length—we discuss this interpretation further in Section 4.6.4. In this manner (4.70) gives rise to a class of time-varying AR processes for some $C > 0$ described by the set of solutions to:

$$\begin{aligned} \widehat{\boldsymbol{a}} = \underset{\boldsymbol{a}}{\operatorname{argmin}} \; & \|\boldsymbol{x}_{N-p} - \boldsymbol{H}_{\boldsymbol{x}}\boldsymbol{a}\|_2^2 \\ \text{subject to} \quad & L_q(\boldsymbol{a}) \leq C. \end{aligned} \tag{4.71}$$

When $q = 1$, the solution to the quadratic program of (4.71) may be obtained by an algorithm called the Lasso, popular in the statistics literature; its properties were originally studied in [142] in the context of shrinkage estimators.

Two alternatives to the quadratic program of (4.71) include minimizing the path-length while constraining the energy of the residual:

$$\begin{aligned} \widehat{\boldsymbol{a}} = \underset{\boldsymbol{a}}{\operatorname{argmin}} \; & L_q(\boldsymbol{a}) \\ \text{subject to} \quad & \|\boldsymbol{x}_{N-p} - \boldsymbol{H}_{\boldsymbol{x}}\boldsymbol{a}\|_2^2 < C, \end{aligned} \tag{4.72}$$

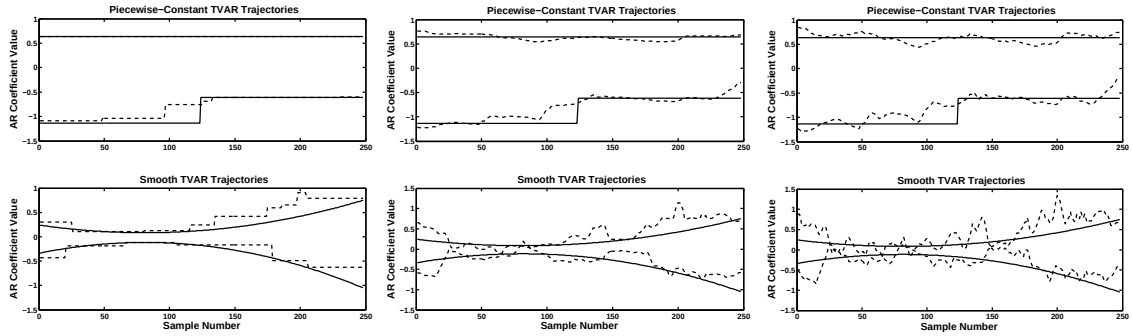


Figure 4.10: Fitting TVAR processes with a constraint on the overall "path-length" of coefficient trajectories. The quadratic program of (4.71) is applied to two TVAR(2) processes with piecewise-constant (top panels) and smooth (bottom panels) coefficient trajectories. The three examples correspond to different norms used: $q = 1$ (left), $q = 1.5$ (middle) and $q = 2$ (right).

or minimizing a dual-objective function as in:

$$\widehat{\boldsymbol{a}} = \underset{\boldsymbol{a}}{\operatorname{argmin}}\ \lambda L_q(\boldsymbol{a}) + (1-\lambda)\|\boldsymbol{x}_{N-p} - \boldsymbol{H}_{\boldsymbol{x}}\boldsymbol{a}\|_2^2, \tag{4.73}$$

for some $\lambda \in [0, 1]$. Note that setting $C = \sigma^2$ in (4.72) allows us to effectively measure the length of the process; however, σ^2 is not always known in practice. The sum-of-norms objective function of (4.73) has also been recently studied in the context of time-series segmentation algorithms [141].

We make the following observations about the problems (4.71), (4.72), and (4.73):

1. All three problems are *convex* when $1 < q \leq \infty$, and solutions may be efficiently found using freely-available software packages such as CVX [131]. Highly-efficient algorithms are available if $q \in \{1, 2, \infty\}$.

2. Using the ℓ_1-norm ($q = 1$) induces a sparse solution by penalizing the number of changes in value within the TVAR coefficient trajectories.

3. In the formulation of (4.71), a time-invariant AR process may be recovered by setting $C = 0$. But, the quadratic programs in (4.72) and (4.73) do not allow such a specialization for any value of C or λ.

As an illustration, Figure 4.10 shows the results of estimating the parameters two TVAR(2) processes, with piecewise-constant and smooth coefficient trajectories, using the globally-constrained estimator of (4.71) with different norms. In each case, the value of the length upper bound C was set based on the true path length of the generated process, i.e., by calculating $L_q(\boldsymbol{a})$ using the true coefficient trajectories.

A second example illustrates the application of (4.71) to a temporarily-unstable TVAR(2) process generated by filtering white Gaussian noise through a second-order digital resonator whose minimum-phase poles ($z, z^* \,|\, z = re^{-j\theta}$, $r = .8$, $\theta = \pi/4$) are briefly moved to their conjugate-reciprocal locations outside the unit circle resulting in an amplitude spike seen in the top panel of Figure 4.11. As shown in the middle panel of Figure 4.11, the power

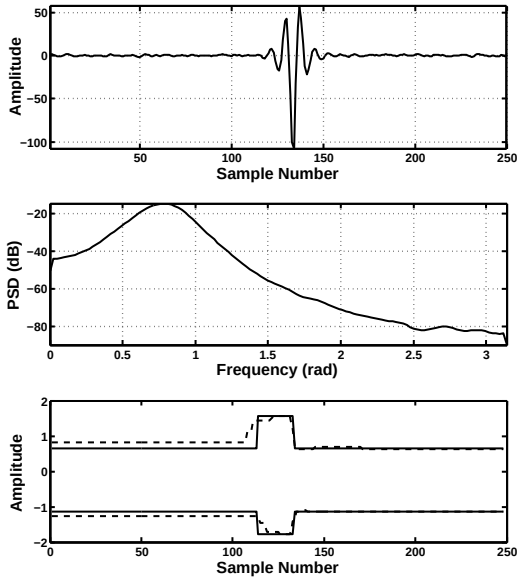


Figure 4.11: Fitting a TVAR(2) process with an ℓ_1 constraint on the overall "path-length" of coefficient trajectories. The time-domain signal showing the temporary instability (top) is shown together with its estimated power spectral density (middle), and with the true (solid) and estimated (dashed) estimates of TVAR coefficient trajectories (bottom).

spectral density remains unchanged, since reflecting poles about the boundary of the unit circle only changes the phase of the filter and leaves the magnitude spectrum unchanged. Nonetheless, the ℓ_1-constrained estimator of (4.71) yields accurate estimates of the TVAR trajectories at every time instant.

4.6.2.2 Time-varying Lattice Filters

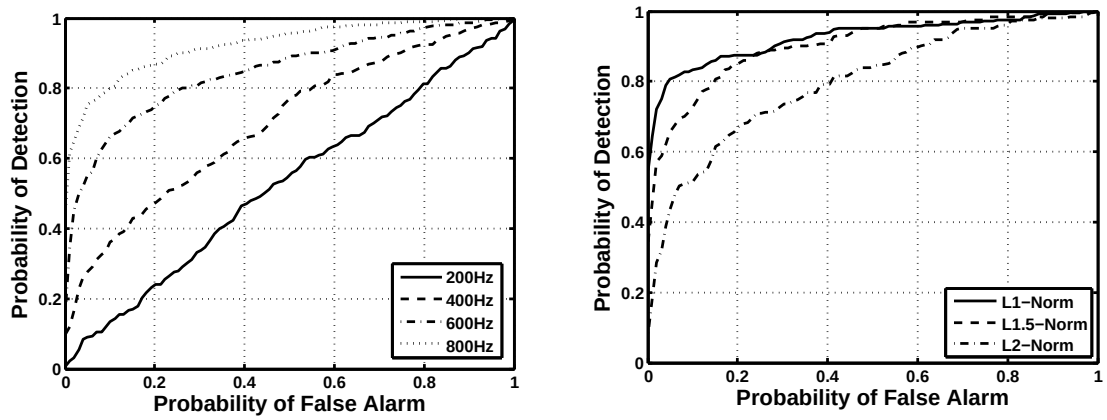
In order to constrain the lattice coefficient estimator of (4.68), we first make a standard assumption (see, e.g., [58]) that:

$$\|e_j^{\mathrm{f}}[n]\|^2 = \|e_j^{\mathrm{b}}[n-1]\|^2 \quad \text{for all} \quad 0 \leq n \leq N-1,$$

which together with (4.40) implies that $\kappa_j^{\mathrm{f}}[n] = \kappa_j^{\mathrm{b}}[n]$, and reduces the number of coefficients that need to be estimated by a factor of two. In this case, $\boldsymbol{\kappa}_j \triangleq \boldsymbol{\kappa}_j^{\mathrm{f}} = \boldsymbol{\kappa}_j^{\mathrm{b}}$ and the estimator of (4.68) becomes:

$$\widehat{\boldsymbol{\kappa}}_j = \underset{\boldsymbol{\kappa}_j}{\operatorname{argmin}}\ L(\boldsymbol{\kappa}_j; \boldsymbol{\theta}_{j-1}) \triangleq \sum_{n=j-1}^{N-1} \|e_j^{\mathrm{f}}[n]\|^2 + \|e_j^{\mathrm{b}}[n]\|^2. \tag{4.74}$$

The estimator of (4.74) may be further constrained in the same vein as (4.69), (4.71), (4.72), or (4.73). For instance, in the case of a path-length constraint we obtain the following quadratic program (convex if $1 < q \leq \infty$) for some $C > 0$ for the jth time-varying



(a) (b)

Figure 4.12: Example of GLRT detection performance for a 100-sample synthetic TVAR(2) signal: (a) GLRT operating characteristics for various frequency jumps $\delta \in \{\pi/40, \pi/20, 3\pi/40, \pi/10\}$ radians; (b) GLRT operating characteristics for different norms $q \in \{1, 1.5, 2\}$.

reflection coefficient:

$$\begin{aligned} \widehat{\boldsymbol{\kappa}}_j &= \underset{\boldsymbol{\kappa}_j}{\operatorname{argmin}} \sum_{n=j-1}^{N-1} \|e_j^{\mathrm{f}}[n]\|^2 + \|e_j^{\mathrm{b}}[n]\|^2 \\ &\text{subject to } \left(\sum_{i=1}^{p} \sum_{n=p+1}^{N-1} |\kappa_i[n] - \kappa_i[n-1]|^q \right)^{1/q} \leq C, \end{aligned} \tag{4.75}$$

where it is assumed that the first $j-1$ lattice coefficients have already been estimated. Since time-varying lattice coefficients are estimated iteratively, the quadratic program of (4.75) has a factor of p (at each stage) less constraints than the TVAR formulation of (4.71), which leads to overall computational savings.

4.6.3 Testing for Stationarity

Setting $\epsilon = 0$ in (4.69) or setting $C = 0$ in (4.71) and (4.75) constrains the set of feasible solutions to those with time-invariant coefficients. This allows us to formulate statistical hypothesis tests for the *stationarity* of $x[n]$ as follows:[2]

$$\begin{array}{llll} \text{Local:} & \mathcal{H}_0: \ \epsilon = 0 & \text{Global:} & \mathcal{H}_0: \ C = 0 \\ & \mathcal{H}_1: \ \epsilon = \epsilon_0 > 0 & & \mathcal{H}_1: \ C = C_0 > 0 \end{array} \tag{4.76}$$

Given a vector $\boldsymbol{x} \in \mathbb{R}^{N \times 1}$ of N observations, both of these hypothesis tests may be realized using a likelihood-ratio statistic (for fixed ϵ_0 or C_0) according to:

$$T(\boldsymbol{x}) \triangleq 2 \ln \frac{\sup_{\boldsymbol{a},\sigma^2} p_{\mathcal{H}_1}(\boldsymbol{x}; \boldsymbol{a}, \sigma^2)}{\sup_{\boldsymbol{a},\sigma^2} p_{\mathcal{H}_0}(\boldsymbol{x}; \boldsymbol{a}, \sigma^2)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma, \tag{4.77}$$

[2]We consider stationarity tests for the parametric model of (4.1) in Chapter 5.

where $p_{\mathcal{H}_i}(\boldsymbol{x};\boldsymbol{\theta})$ denotes the likelihood of $\boldsymbol{a}, \sigma^2$ given the data $\boldsymbol{x}$ under hypothesis $i \in \{0,1\}$.

In the case of the AR coefficient parameterization, the maximum likelihood estimates of $\boldsymbol{a}$ and σ^2 are obtained using (4.69) or (4.71) under $\mathcal{H}_1$ with $\epsilon = \epsilon_0$ or $C = C_0$, and under $\mathcal{H}_0$ using the same estimators with $\epsilon = 0$ or $C = 0$, respectively. In the case of an orthogonal parametrization, estimates of $\{\boldsymbol{\kappa}_1, \ldots, \boldsymbol{\kappa}_p\}$ are obtained under $\mathcal{H}_1$ with $C = C_0$ using (4.75) and under $\mathcal{H}_0$ using (4.75) with $C = 0$. Then an estimate of $\boldsymbol{a}$ may be obtained from the estimated time-varying lattice parameters via the generalized Levinson recursion of (4.38) and (4.39). Finally, note that in the case of a global regularizer on the coefficient variation, when (4.71) or (4.75) are employed, the test statistic depends on the choice of q-norm in (4.70).

To illustrate typical behavior of the likelihood-ratio test statistic of (4.77), we consider a synthetic 100-sample TVAR(2) signal obtained by filtering white Gaussian noise through a second-order digital resonator. The resonator's center frequency is increased by δ radians halfway through the duration of the signal, while its bandwidth is kept constant. The detection performance of the GLRT statistic of (4.77), computed for $q = 1$ to encourage sparsity in the resultant estimates, is illustrated in the left panel of Figure 4.12, which shows receiver operating characteristic (ROC) curves computed for different frequency jump sizes $\delta \in \{\pi/40, \pi/20, 3\pi/40, \pi/10\}$ radians (200 Hz increments). To generate data under $\mathcal{H}_0$, δ was set to zero and 500 trial simulations were performed for each combination. The value of C_0 was set based on the true path length of the generated process, i.e., by calculating $L_q(\boldsymbol{a})$ using the true coefficient trajectories.

In agreement with our intuition, detection performance improves when δ is increased—larger changes are easier to detect. We also considered the effect of the regularization norm q on the power of the likelihood ratio test—it is illustrated in the right panel of Figure 4.12. As expected, the detection performance improves with decreasing q since the underlying signal is sparse.

Note that the formulation of the hypothesis tests in (4.76) requires the specification of the constants ϵ_0 and C_0, which is natural since these constants define nested classes of nonstationary processes. Estimating these constants from data to arrive at a hypothesis test of the form

$$\begin{array}{llll} \text{Local:} & \mathcal{H}_0: & \epsilon = 0 & \qquad \text{Global:} \quad \mathcal{H}_0: \quad C = 0 \\ & \mathcal{H}_1: & \epsilon > 0 & \qquad\qquad\quad \mathcal{H}_1: \quad C > 0 \end{array} \tag{4.78}$$

is not a straightforward hypothesis testing question in our framework, since changing ϵ or C changes the underlying class of stochastic processes. It may be possible to address (4.78) through a multiple-testing approach whereby ϵ_0 and C_0 are systematically increased in the test of (4.76) until the null hypothesis is rejected.

4.6.4 Relationship to Metrics on Manifold of Power Spectral Densities

So far we have considered classes of nonstationary processes defined by limiting variation of the time-varying AR or lattice coefficients using a coefficient-domain distance measure such as (4.70). Below, we explore the relationship of these coefficient-domain distances to intrinsic (Riemannian) metrics on the manifold of AR processes. We touch upon these concepts from a point of view consistent with the spirit of our approach—in

order to highlight viable alternatives for regularizing paths (or geodesics) across "frozen-time" spectra of autoregressive processes.

The central question boils down to finding the right way to measure distance between AR processes. A wide variety of so-called "distortion measures" have been proposed and used for this purpose, primarily motivated by applications in speech analysis and coding [143, 144]. One of the earliest proposals [145] was the "error-matching measure:"

$$d_{\mathrm{IS}}(f_0, f_1) \triangleq \int_{-\pi}^{\pi} \left| \frac{f_0(\omega)}{f_1(\omega)} - \log\left(\frac{f_0(\omega)}{f_1(\omega)}\right) - 1 \right| \frac{d\omega}{2\pi},$$

between power spectral densities f_0 and f_1, commonly referred to as the Itakura-Saito distance. This indexing is consistent with the earlier development and suggests that $i \in \{0, 1\}$ represents two different points in time where signal statistics (e.g., power spectral densities) have been estimated from time-series data. Itakura also introduced the "gain-optimized" distortion $d_{\mathrm{I}}(f_0, f_1)$ that quantifies the differences in "shape" not total power between f_0 and f_1 and is defined by:

$$d_{\mathrm{I}}(f_0, f_1) \triangleq \min_{\lambda \geq 0} d_{\mathrm{IS}}(f_0, \lambda f_1) = \log \int_{-\pi}^{\pi} \frac{f_0(\omega)/\sigma_0^2}{f_1(\omega)/\sigma_1^2} \frac{d\omega}{2\pi}, \tag{4.79}$$

where σ_i^2 is given by the Szego-Kolmogorov formula:

$$\sigma_i^2 \triangleq \exp\left(\int_{-\pi}^{\pi} \log(f_i(\omega)) \frac{d\omega}{2\pi} \right), \tag{4.80}$$

and is the variance of the optimal one-step-ahead prediction error.

The gain-optimized distortion of (4.79) is intimately related to Riemannian metrics on the manifold of power spectral densities. This relationship can be established by considering the so-called "degradation of the prediction error variance"—the ratio of two prediction error variances $\sigma_{01}^2/\sigma_{00}^2$, where σ_{ij}^2 is the variance of the prediction error obtained when a random process with power spectrum f_i is predicted using the optimal predictor designed based on the power spectrum f_j [146]. This measure evaluates how well the optimal predictor designed for one process works when applied to predicting the other, and is equal to the ratio of the *arithmetic* and *geometric* means of the quantity f_0/f_1:

$$\rho(f_0, f_1) \triangleq \frac{\int_{-\pi}^{\pi} \frac{f_0(\omega)}{f_1(\omega)} \frac{d\omega}{2\pi}}{\exp\left(\int_{-\pi}^{\pi} \log\left(\frac{f_0(\omega)}{f_1(\omega)}\right) \frac{d\omega}{2\pi} \right)} = \exp(d_{\mathrm{I}}(f_0, f_1)), \tag{4.81}$$

where the last equality follows from (4.80).

It was shown in [146, Proposition 5] that the quantity $\rho(\cdot, \cdot) - 1$ can be used to induce a Riemannian metric on the topological space of power spectral densities. Indeed, for a small perturbation Δ—thought of as an element of the tangent space of the power spectral densities—$\rho(f, f + \Delta) - 1$ can be locally approximated by the Riemannian metric

$$g_f(\Delta) \triangleq \int_{-\pi}^{\pi} \left(\frac{\Delta(\omega)}{f(\omega)} \right)^2 \frac{d\omega}{2\pi} - \left(\int_{-\pi}^{\pi} \frac{\Delta(\omega)}{f(\omega)} \frac{d\omega}{2\pi} \right)^2,$$

which endows the manifold of power spectral densities with Riemannian structure. It can be shown [146, Proposition 7] that geodesics are easy to compute in this setting, and that geodesic distances between power spectra take the form of (normalized) L_2-distances between their respective logarithms:

$$d(f_0, f_1) \triangleq \sqrt{\int_{-\pi}^{\pi} \left(\log \frac{f_0(\omega)}{f_1(\omega)}\right)^2 \frac{d\omega}{2\pi} - \left(\int_{-\pi}^{\pi} \log \frac{f_0(\omega)}{f_1(\omega)} \frac{d\omega}{2\pi}\right)^2}.$$

Some insight can be obtained by considering the relationship between the degradation of the prediction error variance (4.81) and the coefficient-based distance measure of (4.70). Let $f_i(\omega)$ denote the power spectrum of an autoregressive process for $i \in \{0, 1\}$:

$$f_i(\omega) \triangleq \frac{\sigma_i^2}{|\sum_{k=0}^{p} a_k[i] e^{-jk\omega}|^2}, \tag{4.82}$$

where $a_0[\cdot] \triangleq 1$. Substituting (4.82) into (4.81) yields the following form for $\rho(f_0, f_1)$:

$$\begin{aligned}
\rho(f_0, f_1) &= \left(\frac{\sigma_0^2}{\sigma_1^2} \int_{-\pi}^{\pi} \frac{\left|\sum_{k=0}^{p} a_k[1] e^{-jk\omega}\right|^2}{\left|\sum_{k=0}^{p} a_k[0] e^{-jk\omega}\right|^2} \frac{d\omega}{2\pi}\right) \exp\left(-\int_{-\pi}^{\pi} \log\left(\frac{\sigma_0^2 |\sum_{k=0}^{p} a_k[1] e^{-jk\omega}|^2}{\sigma_1^2 |\sum_{k=0}^{p} a_k[0] e^{-jk\omega}|^2}\right) \frac{d\omega}{2\pi}\right) \\
&= \left(\frac{\sigma_0^2}{\sigma_1^2} \int_{-\pi}^{\pi} \frac{\left|\sum_{k=0}^{p} a_k[1] e^{-jk\omega}\right|^2}{\left|\sum_{k=0}^{p} a_k[0] e^{-jk\omega}\right|^2} \frac{d\omega}{2\pi}\right) \frac{\sigma_1^2}{\sigma_0^2} = \int_{-\pi}^{\pi} \frac{\left|\sum_{k=0}^{p} a_k[1] e^{-jk\omega}\right|^2}{\left|\sum_{k=0}^{p} a_k[0] e^{-jk\omega}\right|^2} \frac{d\omega}{2\pi},
\end{aligned}$$

where the first equality follows directly from the Szego-Kolmogorov formula of (4.80). Simplifying further and calculating the integral we obtain:

$$\begin{aligned}
\rho(f_0, f_1) &= \int_{-\pi}^{\pi} \frac{\left|\sum_{k=0}^{p} a_k[1] e^{-jk\omega}\right|^2}{\left|\sum_{k=0}^{p} a_k[0] e^{-jk\omega}\right|^2} \frac{d\omega}{2\pi} = \int_{-\pi}^{\pi} \left|\sum_{k=0}^{p} a_k[1] e^{-j\omega k}\right|^2 \frac{f_0(\omega)}{\sigma_0^2} \frac{d\omega}{2\pi} \\
&= \int_{-\pi}^{\pi} \sum_{k=0}^{p} \sum_{k'=0}^{p} a_k[1] a_{k'}[1] e^{jk\omega} e^{-jk'\omega} \frac{f_0(\omega)}{\sigma_0^2} \frac{d\omega}{2\pi} = \sum_{k=0}^{p} \sum_{k'=0}^{p} a_k[1] a_{k'}[1] \int_{-\pi}^{\pi} e^{j(k-k')\omega} \frac{f_0(\omega)}{\sigma_0^2} \frac{d\omega}{2\pi} \\
&= \sum_{k=0}^{p} \sum_{k'=0}^{p} a_k[1] a_{k'}[1] r_0[k-k'] = \frac{\boldsymbol{a}_1^T \boldsymbol{R}_{p+1} \boldsymbol{a}_1}{\sigma_0^2},
\end{aligned}$$

where $\boldsymbol{a}_i \triangleq (1, a_1[i], \ldots, a_p[i])^T$, $r_0[k]$ is the positive-definite autocorrelation sequence of the process with power spectral density f_0; $\boldsymbol{R}_{p+1}$ is the associated $(p+1) \times (p+1)$ *Toeplitz* covariance matrix whose entries are given by $\boldsymbol{R}_{p+1}(k, k') = r_0[k-k']$. Thus, if $\boldsymbol{\delta} \triangleq \begin{pmatrix} \delta_1 & \delta_2 & \cdots & \delta_p \end{pmatrix}^T$ denotes a tangent direction in the space of autoregressive coefficients and quantifies a "small" perturbation in the sense $\boldsymbol{a}_0 \to \boldsymbol{a}_1 = \boldsymbol{a}_0 + \begin{pmatrix} 0 & \boldsymbol{\delta}^T \end{pmatrix}^T$, then the degradation of predictive-error variance $\rho(f_0, f_1) - 1$ between the two nearby spectra is given by:

$$\rho(f_0, f_1) - 1 = \rho(f_0, f_1) - \rho(f_0, f_0) = \frac{\boldsymbol{a}_1^T \boldsymbol{R}_p \boldsymbol{a}_1}{\sigma^2} - \frac{\boldsymbol{a}_0^T \boldsymbol{R}_p \boldsymbol{a}_0}{\sigma^2} = \frac{\boldsymbol{\delta}^T \boldsymbol{R}_p \boldsymbol{\delta}}{\sigma^2}.$$

Therefore the differential quantity

$$g_{AR}(\boldsymbol{a}_0, \delta) \triangleq \sqrt{\rho(f_0, f_1) - 1} = \sqrt{\frac{\boldsymbol{\delta}^T \boldsymbol{R}_p \boldsymbol{\delta}}{\sigma^2}} \tag{4.83}$$

may be integrated along geodesics to calculate distances between points on the manifold of AR processes.

In the case of a stable AR(1) process, this calculation is relatively straightforward. Consider two AR(1) processes with spectral density functions f_0 and f_1 defined via:

$$f_0 \triangleq \frac{\sigma_0^2}{|1 - ae^{-j\omega}|^2} \qquad \text{and} \qquad f_1 \triangleq \frac{\sigma_1^2}{|1 - be^{-j\omega}|^2}.$$

Since $r_0[0] = \sigma_0^2/(1-a^2)$ and $r_0[1] = -a\sigma_0^2/(1-a^2)$, we have that:

$$\rho(f_0, f_1) = \begin{pmatrix} 1 & b \end{pmatrix} \begin{pmatrix} \frac{1}{1-a^2} & \frac{-a}{1-a^2} \\ \frac{-a}{1-a^2} & \frac{1}{1-a^2} \end{pmatrix} \begin{pmatrix} 1 \\ b \end{pmatrix} = \frac{b^2 - 2ab + 1}{1 - a^2} = \frac{(b-a)^2}{1-a^2} + 1.$$

Setting $b = a + \delta$ we obtain that:

$$g_{AR}(a, \delta) = \sqrt{\rho(f_0, f_1) - 1} = \frac{\delta}{\sqrt{1 - a^2}}. \tag{4.84}$$

Observe that the closer $|a|$ is to unity the larger the local fluctuation for fixed δ. This is a very desirable feature from a system-theoretic point of view since the metric depends on the placement of the poles. Since geodesics on the space of stable AR(1) processes are straight lines, the geodesic distance from an AR(1) process with parameter $|a| < 1$ to an AR(1) process with parameter $|b| < 1$ is reduces to the following integral:

$$d(f_0, f_1) = \int_a^b \frac{dx}{\sqrt{1 - x^2}} = \arcsin(b) - \arcsin(a). \tag{4.85}$$

It is crucial to observe the differences between the metric of (4.83) and the q-norm-based distance of (4.70). Both forms depend on the perturbation $\boldsymbol{\delta}$, but (4.83) also takes the location of the poles into account and is, therefore, more sensitive to small perturbations of poles near the unit circle than to perturbations of the poles near the origin. On the other hand, it is not clear how to use (4.83) in order to bound the rate of variation of time-varying AR or lattice coefficients in the manner of (4.71) because the resultant optimization problem is no longer convex. Thus, it is desirable to obtain a convenient expression for geodesic distances between AR models based on (4.83) and an efficient approach to computing or approximating it.

4.7 Summary

In this chapter we developed various aspects of time-varying autoregressive models, which generalize classical linear prediction that is fundamental for speech analysis. In the case of TVAR models whose coefficients are modeled via basis function expansions

we established a closed-form expression for the covariance structure of a TVAR process, showed how to apply the results to computing Cramér-Rao lower bounds, and obtained a new way of visualizing time-frequency content. Two new estimators were presented: an EM algorithm for estimating TVAR coefficients from noisy observations, and a convex optimization approach for estimating lattice coefficients that guarantees frozen-time stability of the associated inverse-prediction-error filter.

We proposed two new ways of modeling the temporal evolution of TVAR coefficients. The first approach combines the functional-expansion approach with a random walk, while the second has a geometric flavor—each TVAR process is viewed as a path on the manifold of AR processes. In this latter case, appropriately constraining the underlying coefficient trajectories results in estimators that may be implemented efficiently via convex programming. Interestingly, nonstationary processes defined in this manner admit a nesting structure—stationary counterparts are a special case—that leads to natural hypothesis tests for stationarity, which is the subject of the next chapter.

4.A Appendix: Time-Varying Lattice Filters

In this appendix, we derive the generalized Levinson recursions of (4.38) and (4.39), the time-varying lattice filter equations (4.42) and recursive relationship for the variances of the forward and backward errors.

Recall, that the generalized Levinson recursion is obtained by applying the Gram-Schmidt procedure to the variables $\{x[n-1], x[n-2], \ldots, x[n-j]\}$ in order starting from $x[n-1]$. Its jth stage proceeds as we now describe. The MMSE predictor of $x[n]$ of order j (denoted by $\widehat{x}_j[n]$) can be written as a sum of the MMSE predictor of $x[n]$ of order $j-1$ and the best prediction of $x[n]$ based on the part of $x[n-j]$ in a direction orthogonal to the subspace spanned by $\{x[n-1], x[n-2], \ldots, x[n-j+1]\}$. By the orthogonality principle, the direction orthogonal to this subspace is given *exactly* by the unit vector $e^{\mathrm{b}}_{j-1}[n-1]/\|e^{\mathrm{b}}_{j-1}[n-1]\|^2$ leading to the following decomposition:

$$\widehat{x}_j[n] = \widehat{x}_{j-1}[n] + \frac{\langle x[n], e^{\mathrm{b}}_{j-1}[n-1]\rangle}{\|e^{\mathrm{b}}_{j-1}[n-1]\|^2} e^{\mathrm{b}}_{j-1}[n-1]. \tag{4.86}$$

Now note that by orthogonality we have that:

$$\left\langle x[n] - e^{\mathrm{f}}_{j-1}[n], e^{\mathrm{b}}_{j-1}[n-1] \right\rangle = \left\langle \sum_{i=1}^{j-1} a_{i,j-1}[n] x[n-i], e^{\mathrm{b}}_{j-1}[n-1] \right\rangle = 0$$

since the terms $\{x[n-1], x[n-2], \ldots, x[n-j+1]\}$ all appear in $e^{\mathrm{b}}_{j-1}[n-1]$ according to (4.37), and the coefficients $b_{i,j}[n]$ form the optimal MSE predictor of $x[n-j]$. Thus

$$\left\langle x[n], e^{\mathrm{b}}_{j-1}[n-1] \right\rangle = \left\langle e^{\mathrm{f}}_{j-1}[n] + \sum_{i=1}^{j-1} a_{i,j}[n] x[n-i], e^{\mathrm{b}}_{j-1}[n-1] \right\rangle = \left\langle e^{\mathrm{f}}_{j-1}[n], e^{\mathrm{b}}_{j-1}[n-1] \right\rangle.$$

Since the forward time-varying lattice (reflection) coefficients $\kappa^{\mathrm{f}}_j[n]$ are defined in (4.40) as

$$\kappa^{\mathrm{f}}_j[n] \triangleq -\frac{\langle x[n], e^{\mathrm{b}}_{j-1}[n-1]\rangle}{\|e^{\mathrm{b}}_{j-1}[n-1]\|^2} = -\frac{\langle e^{\mathrm{f}}_{j-1}[n], e^{\mathrm{b}}_{j-1}[n-1]\rangle}{\|e^{\mathrm{b}}_{j-1}[n-1]\|^2},$$

and substituting (4.40) into (4.86) leads to:

$$\begin{aligned}\widehat{x}_j[n] &= \widehat{x}_{j-1}[n] - \kappa_j^{\mathrm{f}}[n] e_{j-1}^{\mathrm{b}}[n-1] \\ &= \sum_{i=1}^{j-1} a_{i,j-1}[n] x[n-i] - \kappa_j^{\mathrm{f}}[n] \left(x[n-j] - \sum_{i=0}^{j-1} b_{i,j-1}[n-1] x[n-j+i] \right) \\ &= \sum_{i=1}^{j-1} a_{i,j-1}[n] x[n-i] - \kappa_j^{\mathrm{f}}[n] \left(x[n-j] - \sum_{i=0}^{j-1} b_{j-i,j-1}[n-1] x[n-i] \right), \end{aligned} \tag{4.87}$$

where (4.87) was obtained by substituting i' for $j-i$ and then relabelling i' as i. Of course (4.87) must also be equal to:

$$\widehat{x}_j[n] = \sum_{i=1}^{j} a_{i,j}[n] x[n-i]. \tag{4.88}$$

Consequently, equating (4.87) and (4.88) we arrive at:

$$a_{i,j}[n] = \begin{cases} a_{i,j-1}[n] + \kappa_j^{\mathrm{f}}[n] b_{j-i,j-1}[n-1] \text{ for } 1 \le i < j \\ -\kappa_j^{\mathrm{f}}[n] \text{ if } i = j \end{cases}.$$

A similar development yields

$$b_{i,j}[n] = \begin{cases} b_{i,j-1}[n-1] + \kappa_j^{\mathrm{b}}[n] a_{j-i,j-1}[n] \text{ for } 1 \le i < j \\ -\kappa_j^{\mathrm{b}}[n] \text{ if } i = j. \end{cases}$$

Consequently, the optimal forward and backward linear prediction coefficients of order j can be computed from those of order $j-1$ via the generalized Levinson recursions of (4.38) and (4.39).

Next, observe that (4.42) follows from substituting (4.38) and (4.39) into (4.36) and (4.37), respectively. To see this explicitly note

$$\begin{aligned} e_j^{\mathrm{f}}[n] &\triangleq x[n] - \sum_{i=1}^{j} a_{i,j}[n] x[n-i] \\ &= x[n] - \sum_{i=1}^{j-1} \Big(a_{i,j-1}[n] + \kappa_j^{\mathrm{f}}[n] b_{j-i,j-1}[n-1] \Big) x[n-i] + \kappa_j^{\mathrm{f}}[n] x[n-j] \\ &= x[n] - \sum_{i=1}^{j-1} a_{i,j-1}[n] x[n-i] + \kappa_j^{\mathrm{f}}[n] \left(x[n-j] - \sum_{i=1}^{j-1} b_{j-i,j-1}[n-1] x[n-i] \right) \\ &= e_{j-1}^{\mathrm{f}}[n] + \kappa_j^{\mathrm{f}}[n] \left(x[n-j] - \sum_{i=0}^{j-1} b_{i,j-1}[n-1] x[n-j+i] \right) = e_{j-1}^{\mathrm{f}}[n] + \kappa_j^{\mathrm{f}}[n] e_{j-1}^{\mathrm{b}}[n-1]. \end{aligned}$$

Next, define the forward and backward error variances $\rho_j^{\mathrm{f}}[n]$ and $\rho_j^{\mathrm{b}}[n]$ according to

$$\rho_j^{\mathrm{f}}[n] = \langle e_j^{\mathrm{f}}[n], e_j^{\mathrm{f}}[n] \rangle, \qquad \text{and} \qquad \rho_j^{\mathrm{b}}[n] = \langle e_j^{\mathrm{b}}[n], e_j^{\mathrm{b}}[n] \rangle.$$

Using (4.42), we can develop a recursive relationship for the error variances as follows:

$$\begin{aligned}
\rho_j^{\mathrm{f}}[n] &= \langle e_j^{\mathrm{f}}[n], e_j^{\mathrm{f}}[n]\rangle = \langle e_{j-1}^{\mathrm{f}}[n] + \kappa_j^{\mathrm{f}}[n] e_{j-1}^{\mathrm{b}}[n-1], e_{j-1}^{\mathrm{f}}[n] + \kappa_j^{\mathrm{f}}[n] e_{j-1}^{\mathrm{b}}[n-1]\rangle \\
&= \langle e_{j-1}^{\mathrm{f}}[n], e_{j-1}^{\mathrm{f}}[n]\rangle + 2\kappa_j^{\mathrm{f}}[n]\langle e_{j-1}^{\mathrm{f}}[n], e_{j-1}^{\mathrm{b}}[n-1]\rangle + \left(\kappa_j^{\mathrm{f}}[n]\right)^2 \langle e_{j-1}^{\mathrm{b}}[n-1], e_{j-1}^{\mathrm{b}}[n-1]\rangle \\
&= \rho_{j-1}^{\mathrm{f}}[n] - 2\left(\kappa_j^{\mathrm{f}}[n]\right)^2 \rho_{j-1}^{\mathrm{b}}[n-1] + \left(\kappa_j^{\mathrm{f}}[n]\right)^2 \rho_{j-1}^{\mathrm{b}}[n-1] \\
&= \rho_{j-1}^{\mathrm{f}}[n] - \left(\kappa_j^{\mathrm{f}}[n]\right)^2 \rho_{j-1}^{\mathrm{b}}[n-1],
\end{aligned}$$

In the case of backward errors, a similar recursion is obtained:

$$\rho_j^{\mathrm{b}}[n] = \rho_{j-1}^{\mathrm{b}}[n-1] - \left(\kappa_j^{\mathrm{b}}[n]\right)^2 \rho_{j-1}^{\mathrm{f}}[n].$$

If $\rho_j^{\mathrm{f}}[n] = \rho_j^{\mathrm{b}}[n-1]$—which occurs if $\kappa_j^{\mathrm{f}}[n] = \kappa_j^{\mathrm{b}}[n]$—then the above recursions reduce to $\rho_j[n] = \rho_{j-1}[n](1 - \kappa_j^2[n])$. In the time-invariant case, the last recursion reduces to the familiar: $\rho_j = \rho_{j-1}(1 - \kappa_j^2)$.

Chapter 5

Parametric and Nonparametric Tests for Stationarity

A fundamental first step in applying statistical techniques designed for nonstationary stochastic processes is to confirm whether or not the data in question is, in fact, nonstationary. To this end, this chapter considers the general problem of whether a time-series of N observations contains sufficient evidence to reject the null hypothesis of stationarity. In the context of this question, we study the propriety of modeling speech time series using time-varying autoregressions, and consider cases when such parametric models are inappropriate and a more general, nonparametric approach is warranted.

In the first part of this chapter, we pursue a parametric approach to stationarity testing based on time-varying autoregressive (TVAR) models and evaluate the resultant methods in the context of speech analysis. A generalized likelihood ratio test (GLRT) is derived and studied both empirically for short data records, using formant-like synthetic examples, and asymptotically, leading to constant false alarm rate hypothesis tests. We show that the resultant computationally-efficient procedure may be applied to detect the presence of temporal vocal tract variation in speech waveform data. This is illustrated using two in-depth case studies conducted across different time scales of speech dynamics: first, the hypothesis testing framework is applied to detecting formant changes on the scale of tens of milliseconds of data, and second, to the problem of identification of glottal opening and closing instants, within individual pitch periods, on time scales below ten milliseconds.

Next, we examine the question of testing for stationarity when a suitable parametric model for the data under consideration is not available, thus rendering parametric tests of limited use. In this case, the time-invariance of coefficients arising in nonparametric representations can be used to test for stationarity—in analogy to the parametric setting whereby the constancy of the model parameters over time is necessary to guarantee the stationarity of the stochastic process they represent. Our main contribution in this direction is to propose an efficient Monte Carlo method for characterizing the null distribution of two test statistics based on empirical short-time Fourier coefficients. The approach is illustrated using synthetic and audio examples, and is then used as a method of obtaining a signal-adaptive variable-resolution Fourier analysis and a corresponding signal enhancement scheme.

5.1 Time-Varying Autoregressions in Speech: Detection Theory and Applications

We now develop a statistical detection framework for identifying vocal tract dynamics in speech data across different time scales. Since the source-filter view of speech production motivates modeling a stationary vocal tract using the standard linear-predictive or autoregressive (AR) model [3], it is natural to represent temporal variation in the vocal tract using a time-varying autoregressive process. Consequently, we propose here to detect vocal tract changes via a generalized likelihood ratio test to determine whether an AR or TVAR model is most appropriate for a given speech data segment. Our main methodological contribution is to derive this test and describe its asymptotic behavior. Our contribution to speech analysis is then to consider two specific, in-depth case studies of this testing framework: detecting change in speech spectra, and detecting glottal opening and closing instants from waveform data.

Earlier work in this direction began with the fitting of piecewise-constant AR models to test for nonstationarity [147–149]. However, in reality, the vocal tract often varies slowly, rather than as a sequence of abrupt jumps; to this end, [58, 92–98, 100–102] studied time-varying linear prediction using TVAR models. Associated estimation methods were further studied in [99, 111, 115, 116, 118, 121, 150]. In a more general setting, Kay [151] recently proposed a version of the Rao test for AR vs. TVAR determination; however, when available, likelihood ratio tests often outperform their Rao test counterparts for finite sample sizes [152]. Our approach is the first to develop and apply the GLRT for AR vs. TVAR determination in the speech analysis setting. In the context of array processing, Abramovich et. al. [108] considered a similar problem when *multiple* time-series of observations are available. Nonparametric approaches to detecting spectral change in acoustic signals were proposed by the current authors in [153, 154].

Detecting spectral variation across multiple scales is an important first step toward appropriately exploiting vocal tract dynamics. This can lead to improved speech analysis algorithms on time scales on the order of tens of milliseconds for speech enhancement [100, 153, 155, 156], classification of time-varying phonemes such as unvoiced stop consonants [97], and forensic voice comparison [157]. At the sub-segmental time scale (i.e., less than one pitch period), sliding-window AR analysis has been used to capture vocal tract variation and to study the excitation waveform as a key first step in applications including inverse filtering [158], speaker identification [42], synthesis [159], and clinical voice assessment [160].

In the first part of this section, we develop a general detection theory for speech analysis based on TVAR models. In Section 5.1.1, we formally introduce these models, derive their corresponding maximum-likelihood estimators, and develop a GLRT appropriate for speech waveforms. After providing examples using real and synthetic data, including an analysis of vowels and diphthongs from the TIMIT database [48], we then formulate in Section 5.1.2 a constant false alarm rate (CFAR) test and characterize its asymptotic behavior. In Section 5.1.3, we discuss the relationship of our framework to classical methods, including the piecewise-constant AR approach of [147].

Next, we consider two prototype speech analysis applications: in Section 5.1.4, we apply our GLRT framework to detect formant changes in both whispered and voiced

speech. We then show how to detect glottal opening *and* closing instants via the GLRT in Section 5.1.5. We evaluate our results on the more difficult problem of detecting glottal openings [161] using ground-truth data obtained by electroglottograph (EGG) analysis, and also show performance comparable to methods based on linear prediction and group delay for the task of identifying glottal closures.

5.1.1 Time-Varying Autoregressions and Testing

5.1.1.1 Model Specification

Recall the classical pth-order linear predictive model for speech, also known as an AR(p) autoregression [3]:

$$\text{AR}(p): \quad x[n] = \begin{cases} \sum_{i=1}^{p} a_i x[n-i] + \sigma w[n] & \text{if } n \geq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{5.1}$$

where the sequence $w[n]$ is a zero-mean white Gaussian process with unit variance, scaled by a gain parameter $\sigma > 0$. A more flexible pth-order *time-varying* autoregressive model is given by a one-sided, *shifted* TVAR process:

$$\text{TVAR}(p): \quad x[n] = \begin{cases} \sum_{i=1}^{p} a_i[n] x[n-i] + \sigma w[n] & \text{if } n \geq 0, \\ 0 & \text{otherwise,} \end{cases} \tag{5.2}$$

first discussed in Section 4.1. Note that time-dependence of the linear prediction coefficients $a_i[n]$ of (5.2) implies that (5.2) is a *nonstationary* random process.

As in Chapter 4, the temporal evolution of the linear prediction coefficients is specified in terms of an expansion in a set of $q+1$ functions $f_j[n]$ weighted by coefficients α_{ij} as follows:

$$a_i[n] = \sum_{j=0}^{q} \alpha_{ij} f_j[n], \quad \text{for all } 1 \leq i \leq p. \tag{5.3}$$

We assume throughout that the unit-valued function $f_0[n] = 1$ is included in the chosen basis set, so that the classical AR(p) model of (5.1) is recovered as $a_i \equiv \alpha_{i0} \cdot 1$ whenever $\alpha_{ij} = 0$ for all $j > 0$.

5.1.1.2 AR vs. TVAR Generalized Likelihood Ratio Test (GLRT)

We now describe how to test the hypothesis $\mathcal{H}_0$ that a given signal segment $\boldsymbol{x} \triangleq (x[0]\ x[1]\ \cdots\ x[N-1])^T$ has been generated by an AR(p) process according to (5.1), against the alternative hypothesis $\mathcal{H}_1$ of a TVAR(p) process as specified by (5.2) and (5.3) above. We introduce a GLRT to examine evidence of *change* in linear prediction coefficients over time, and consequently in the vocal tract resonances that they represent in the classical source-filter model of speech.

According to the functional expansion of (5.3), the TVAR(p) model of (5.2) is fully described by $p(q+1)$ expansion coefficients α_{ij} and the gain term σ. For convenience we group the coefficients α_{ij} into $q+1$ vectors $\boldsymbol{\alpha}_j$, $0 \leq j \leq q$, as

$$\boldsymbol{\alpha}_j \triangleq \begin{pmatrix} \alpha_{1j} & \alpha_{2j} & \cdots & \alpha_{pj} \end{pmatrix}^T.$$

We may then partition a vector $\boldsymbol{\alpha} \in \mathbb{R}^{p(q+1)\times 1}$ into blocks associated to the AR(p) portion of the model $\boldsymbol{\alpha}_{\mathrm{AR}}$, and the remainder $\boldsymbol{\alpha}_{\mathrm{TV}}$, which captures time variation:

$$\boldsymbol{\alpha} \triangleq \left(\boldsymbol{\alpha}_{\mathrm{AR}}^T \;\middle|\; \boldsymbol{\alpha}_{\mathrm{TV}}^T\right)^T = \left(\boldsymbol{\alpha}_0^T \;\middle|\; \boldsymbol{\alpha}_1^T \;\; \boldsymbol{\alpha}_2^T \;\; \cdots \;\; \boldsymbol{\alpha}_q^T\right)^T. \tag{5.4}$$

Recalling that the TVAR(p) model (hypothesis $\mathcal{H}_1$) reduces to an AR(p) model (hypothesis $\mathcal{H}_0$) precisely when $\boldsymbol{\alpha}_j = \mathbf{0}$ for all $j > 0$, we may formulate the following hypothesis test:

$$\begin{aligned} &\text{Model}: \quad \text{TVAR}(p) \text{ with parameters } \boldsymbol{\alpha}, \sigma^2; \\ &\text{Hypotheses}: \quad \begin{cases} \mathcal{H}_0 : \boldsymbol{\alpha}_j = \mathbf{0} & \text{for } \textit{all } j > 0, \\ \mathcal{H}_1 : \boldsymbol{\alpha}_j \neq \mathbf{0} & \text{for } \textit{at least one } j > 0. \end{cases} \end{aligned} \tag{5.5}$$

Each of these two hypotheses in turn induces a data likelihood in the observed signal $\boldsymbol{x} \in \mathbb{R}^{N\times 1}$, which we denote by $p_{\mathcal{H}_i}(\cdot)$ for $i = 0, 1$. The corresponding generalized likelihood ratio test comprises evaluation of a test statistic $T(\boldsymbol{x})$, and rejection of $\mathcal{H}_0$ in favor of $\mathcal{H}_1$ if $T(\boldsymbol{x})$ exceeds a given threshold γ:

$$T(\boldsymbol{x}) \triangleq 2 \ln \frac{\sup_{\boldsymbol{\alpha},\sigma^2} p_{\mathcal{H}_1}(\boldsymbol{x}; \boldsymbol{\alpha}, \sigma^2)}{\sup_{\boldsymbol{\alpha}_0,\sigma^2} p_{\mathcal{H}_0}(\boldsymbol{x}; \boldsymbol{\alpha}_0, \sigma^2)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma. \tag{5.6}$$

5.1.1.3 Evaluation of the GLRT Statistic

The numerator and denominator of (5.6) respectively imply maximum-likelihood (ML) parameter estimates of $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_{\mathrm{AR}}^T \mid \boldsymbol{\alpha}_{\mathrm{TV}}^T)^T$ and $\boldsymbol{\alpha}_0$ in (5.4) under the specified TVAR(p) and AR(p) models, along with their respective gain terms σ^2. Intuitively, when $\mathcal{H}_0$ is in force, estimates of $\boldsymbol{\alpha}_{\mathrm{TV}}$ will be small; we formalize this notion in Section 5.1.2.1 by showing how to set the test threshold γ to achieve a constant false alarm rate.

As previously discussed in Section 4.3.1, conditional ML estimates are easily obtained in closed form, and terms in (5.6) reduce to estimates of σ^2 under hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively. Given N observations, partitioned according to

$$\boldsymbol{x} = (\boldsymbol{x}_p \;\; \boldsymbol{x}_{N-p})^T \triangleq \left(x[0] \cdots x[p-1] \;\middle|\; x[p] \cdots x[N-1]\right)^T,$$

the joint probability density function of $\boldsymbol{\alpha}, \sigma^2$ is given by:

$$p(\boldsymbol{x}\,;\, \boldsymbol{\alpha}, \sigma^2) = p(\boldsymbol{x}_{N-p} \,|\, \boldsymbol{x}_p\,;\, \boldsymbol{\alpha}, \sigma^2) p(\boldsymbol{x}_p\,;\, \boldsymbol{\alpha}, \sigma^2). \tag{5.7}$$

As is standard practice, we approximate the *unconditional* data likelihood of (5.7) by the *conditional* likelihood $p(\boldsymbol{x}_{N-p} \,|\, \boldsymbol{x}_p\,;\, \boldsymbol{\alpha}, \sigma^2)$, whose maximization yields an estimator that converges to the exact (unconditional) ML estimator as $N \to \infty$ (see, e.g., [4] for this argument under $\mathcal{H}_0$). Gaussianity of $w[n]$ implies the conditional likelihood

$$p(\boldsymbol{x}_{N-p} \,|\, \boldsymbol{x}_p; \boldsymbol{\alpha}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{(N-p)/2}} \exp\left(-\sum_{n=p}^{N-1} \frac{e^2[n]}{2\sigma^2}\right), \tag{5.8}$$

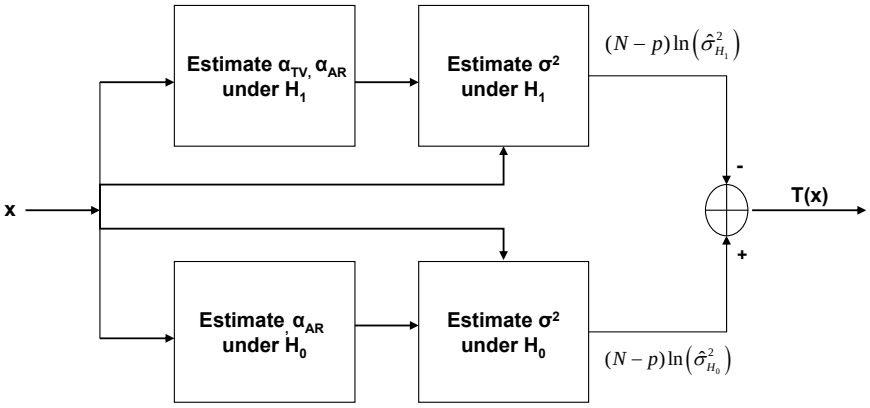


Figure 5.1: Computation of the GLRT statistic $T(\boldsymbol{x})$ according to Section 5.1.1.3.

where $e[n] \triangleq x[n] - \sum_{i=1}^{p} \sum_{j=0}^{q} \alpha_{ij} f_j[n] x[n-i]$ is the associated prediction error. As shown in Section 4.3.1, maximizing the logarithm of (5.8) with respect to $\boldsymbol{\alpha}$ yields the conditional ML estimate of $\boldsymbol{\alpha}$ as

$$\widehat{\boldsymbol{\alpha}} = \left(\boldsymbol{H}_{\boldsymbol{x}}^T \boldsymbol{H}_{\boldsymbol{x}}\right)^{-1} \boldsymbol{H}_{\boldsymbol{x}}^T \boldsymbol{x}_{N-p}, \tag{5.9}$$

where the $(n-p+1)$th row of the matrix $\boldsymbol{H}_{\boldsymbol{x}} \in \mathbb{R}^{(N-p)\times p(q+1)}$ is given by the Kronecker product $(x[n-1] \ \cdots \ x[n-p]) \otimes (f_0[n] \ f_1[n] \ \cdots \ f_q[n])$ for any $p \leq n \leq N-1$.

The estimator of (5.9) corresponds to a generalization of the *covariance method* of linear prediction—to which it exactly reduces when the number q of non-constant basis functions employed is set to zero [94]; we discuss the corresponding generalization of the autocorrelation method in Section 5.1.3.2.

The conditional ML estimate of σ^2 is obtained by substituting (5.9) into (5.8) and maximizing with respect to σ^2, yielding

$$\widehat{\sigma^2} = \frac{1}{N-p} \sum_{n=p}^{N-1} \left(x[n]x[n] - \sum_{i=1}^{p} \sum_{j=0}^{q} \widehat{\alpha_{ij}} f_j[n] x[n] x[n-i] \right). \tag{5.10}$$

Under $\mathcal{H}_0$ (the time-invariant case), the estimator of (5.10) reduces to the familiar estimator of the variance: $\widehat{\sigma^2} = \widehat{r}_{xx}[0] - \sum_{i=1}^{p} \alpha_{i0} \widehat{r}_{xx}[i]$, where $r_{xx}[\tau]$ is the autocorrelation function of $x[n]$ at lag τ.

In summary, the conditional ML estimates of $\boldsymbol{\alpha}_{\mathrm{AR}}, \boldsymbol{\alpha}_{\mathrm{TV}}$ and σ^2 under $\mathcal{H}_1$ are obtained using (5.9) and (5.10), respectively. Estimates of $\boldsymbol{\alpha}_{\mathrm{AR}}$ and σ^2 under $\mathcal{H}_0$ are obtained by setting $q = 0$ in (5.9) and (5.10). Substituting these estimates into the GLRT statistic of (5.6), we recover the following intuitive form for $T(\boldsymbol{x})$, whose computation is illustrated in Figure 5.1:

$$T(\boldsymbol{x}) = (N-p) \ln \left(\widehat{\sigma^2}_{\mathcal{H}_0} / \widehat{\sigma^2}_{\mathcal{H}_1} \right). \tag{5.11}$$

The computational complexity of evaluating the GLRT statistic of (5.11) depends on the complexity of inverting the $p(q+1) \times p(q+1)$ matrix $\boldsymbol{H}_{\boldsymbol{x}}^T \boldsymbol{H}_{\boldsymbol{x}}$ in (2.26), and, consequently, on the number of TVAR coefficients p and basis functions $q+1$. Under the null

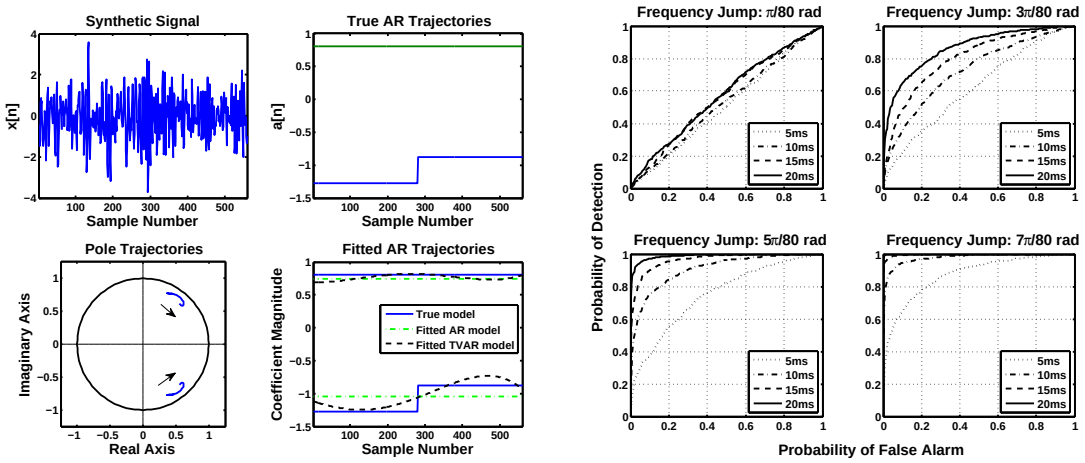


Figure 5.2: Example of GLRT detection performance for a "formant-like" synthetic TVAR(2) signal. Left: A test signal and its TVAR coefficients are shown at top, with pole trajectories and AR vs. TVAR estimates below. Right: Operating characteristics of the corresponding GLRT ($p = 2$ TVAR coefficients, $q = 4$ Legendre polynomials, $f_s = 16$ kHz) shown for various frequency jumps and data lengths.

hypothesis $\mathcal{H}_0$ ($q = 0$) $\boldsymbol{H}_{\boldsymbol{x}}^T \boldsymbol{H}_{\boldsymbol{x}}$ is Toeplitz and may be inverted in $\mathcal{O}(p^2)$ operations using the algorithm of Morf [162]. An extension of the same algorithm may be used to invert $\boldsymbol{H}_{\boldsymbol{x}}^T \boldsymbol{H}_{\boldsymbol{x}}$ in $\mathcal{O}(p^3(q+1)^2)$ operations under the alternative hypothesis $\mathcal{H}_1$; computational savings arise because $\boldsymbol{H}_{\boldsymbol{x}}^T \boldsymbol{H}_{\boldsymbol{x}}$ may be written as a product of two block-Toeplitz matrices in this case [58]. Consequently, $\mathcal{O}(Np^3(q+1)^2)$ total operations are required to compute (5.11) from N waveform samples.

5.1.1.4 Evaluation of GLRT Detection Performance

To demonstrate typical GLRT behavior, we first consider an example detection scenario involving a "formant-like" signal synthesized by filtering white Gaussian noise through a second-order digital resonator. The resonator's center frequency is increased by δ radians halfway through the duration of the signal, while its bandwidth is kept constant; an example 560-sample signal with $\delta = 7\pi/80$ radians is shown in Figure 5.2.

Detection performance in this setting is summarized in the right-hand panel of Figure 5.2, which shows receiver operating characteristic (ROC) curves for different signal lengths N and frequency jump sizes δ. These were varied in the ranges $N \in \{80, 240, 400, 560\}$ samples (10 ms increments) and $\delta \in \{\pi/80, 3\pi/80, 5\pi/80, 7\pi/80\}$ radians (200 Hz increments), and 1000 trial simulations were performed for each combination. To generate data under $\mathcal{H}_0$, δ was set to zero. In agreement with our intuition, detection performance improves when δ is increased while N is fixed, and vice versa—simply put, larger changes and those occurring over longer intervals are easier to detect. Moreover, even though the span of the chosen Legendre polynomials does not include the actual piecewise-constant coefficient trajectories, the norm of their projection onto this basis set is sufficiently large to trigger a detection with high probability.

Table 5.1: Vocal Tract Variation in TIMIT Vowels & Diphthongs.

Vowel	[ɛ]	[ɪ]	[æ]	[ʌ]	[ʊ]	[ə]
$T(\boldsymbol{x})$	67.5	60.5	94.6	63.8	58.9	32.1
Diphthong	[oʊ]	[ɔɪ]	[aɪ]	[eɪ]	[aʊ]	[ɝ]
$T(\boldsymbol{x})$	134.1	302.4	187.4	130.6	161.6	133.0

We next consider a large-scale experiment designed to test the sensitivity of the test statistic $T(\boldsymbol{x})$ to vocal tract variation in real speech data. To this end, we fitted AR(10) and TVAR(10) models (with $q = 4$ Legendre polynomials) to all instances of the vowels [ɛ], [ɪ], [æ], [ʌ], [ʊ], [ə] (as in "bet,' "bit," "bat," "but," "book," and "about") and the diphthongs [oʊ], [ɔɪ], [aɪ], [eɪ], [aʊ], [ɝ] (as in "boat," "boy," "bite," "bait," "bout," and "bird") in the training portion of the TIMIT database [48]. Data were downsampled to 8 kHz, and values of $T(\boldsymbol{x})$ were averaged across all dialects, speakers, and sentences ($50,000$ vowel and $25,000$ diphthong instances in total).

Per-phoneme averages are reported in Table 5.1, and indicate considerably stronger detections of vocal tract variation in diphthongs than in vowels—and indeed a two-sample t test easily rejects (p-value $\cong 0$) the hypothesis that the average values of $T(\boldsymbol{x})$ for the two groups are equal. This finding is consistent with the physiology of speech production, and demonstrates the sensitivity of the GLRT in practice.

5.1.2 Analysis of Detection Performance

To apply the hypothesis test of (5.5), it is necessary to select a threshold γ as per (5.6), such that the null hypothesis of a best-fit AR(p) model is rejected in favor of the fitted TVAR(p) model whenever $T(\boldsymbol{x}) > \gamma$. Below we describe how to choose γ to guarantee a constant false alarm rate (CFAR) for large sample sizes, and give the asymptotic (in N) distribution of the GLRT statistic under $\mathcal{H}_0$ and $\mathcal{H}_1$, showing how these results yield practical consequences for speech analysis.

5.1.2.1 Derivation of GLRT Asymptotics and CFAR Test

Under suitable technical conditions [28], likelihood ratio statistics take on a chi-squared distribution $\chi^2_d(0)$ as the sample size N grows large whenever $\mathcal{H}_0$ is in force, with the degrees of freedom d equal to the number of parameters restricted under the null hypothesis. In our setting, $d = pq$ since the pq coefficients $\boldsymbol{\alpha}_{\mathrm{TV}}$ are restricted to be zero under $\mathcal{H}_0$, and we may write that $T(\boldsymbol{x}) \sim \chi^2_{pq}(0)$ under $\mathcal{H}_0$ as $N \to \infty$.

Thus, we may specify an allowable asymptotic *constant false alarm rate* for the GLRT of (5.5), defined as follows:

$$\lim_{N\to\infty} \Pr\left\{T(\boldsymbol{x}) > \gamma; \mathcal{H}_0\right\} = \Pr\left\{\chi^2_{pq}(0) > \gamma\right\}. \tag{5.12}$$

Since the asymptotic distribution of $T(\boldsymbol{x})$ under $\mathcal{H}_0$ depends *only* on p and q, which are set in advance, we can determine a CFAR threshold γ by fixing a desired value (say, 5%) for the right-hand side of (5.12), and evaluating the inverse cumulative distribution function

of $\chi^2_{pq}(0)$ to obtain the value of γ that guarantees the specified (asymptotic) constant false alarm rate.

When $\boldsymbol{x}$ is a TVAR process so that the alternate hypothesis $\mathcal{H}_1$ is in force, $T(\boldsymbol{x})$ instead takes on (as $N \to \infty$) a *noncentral* chi-squared distribution $\chi^2_d(\lambda)$. Its noncentrality parameter $\lambda > 0$ depends on the *true but unknown* parameters of the model under $\mathcal{H}_1$; thus in general

$$T(\boldsymbol{x}) \overset{N\to\infty}{\sim} \chi^2_{pq}(\lambda), \quad \begin{cases} \lambda = 0 & \text{under } \mathcal{H}_0, \\ \lambda > 0 & \text{under } \mathcal{H}_1. \end{cases} \tag{5.13}$$

As shown in Appendix 5.A, it is easy to establish that the expression for λ in the case at hand is given by

$$\lambda = \boldsymbol{\alpha}_{\mathrm{TV}}^T (\overline{\boldsymbol{F}^T \boldsymbol{F} \otimes \sigma^{-2} \boldsymbol{R}}) \boldsymbol{\alpha}_{\mathrm{TV}}, \tag{5.14}$$

where $\overline{\cdot}$ denotes the Schur complement with respect to the first $p \times p$ matrix block of its argument, the $(j+1)$th column of the matrix $\boldsymbol{F} \in \mathbb{R}^{(N-p)\times(q+1)}$ is given by $\begin{pmatrix} f_j[p] & f_j[p+1] & \cdots & f_j[N-1] \end{pmatrix}^T$ and $\boldsymbol{R}$ is given by:

$$\boldsymbol{R} \triangleq \begin{pmatrix} r_{xx}[0] & r_{xx}[1] & \cdots & r_{xx}[p-1] \\ r_{xx}[1] & r_{xx}[0] & \cdots & r_{xx}[p-2] \\ \vdots & \vdots & \ddots & \vdots \\ r_{xx}[p-1] & r_{xx}[p-2] & \cdots & r_{xx}[0] \end{pmatrix}. \tag{5.15}$$

Here $\{r_{xx}[0], r_{xx}[1], \ldots, r_{xx}[p-1]\}$ is the autocorrelation sequence corresponding to $\boldsymbol{\alpha}_{\mathrm{AR}}$ (given, e.g., by the "step-down algorithm" [4]). The expression of (5.14) follows from the fact that $\boldsymbol{F}^T \boldsymbol{F} \otimes \sigma^{-2} \boldsymbol{R}$ is the Fisher information matrix for our TVAR(p) model; its Schur complement arises from the composite form of our hypothesis test, since the parameters $\boldsymbol{\alpha}_{\mathrm{AR}}, \sigma^2$ are unrestricted under $\mathcal{H}_0$.

More generally, we may relate this result to the underlying TVAR coefficient trajectories $a_i[n]$, arranged as columns of a matrix $\boldsymbol{A}$, with each column-wise mean trajectory value a corresponding entry in a matrix $\bar{\boldsymbol{A}}$. Letting $\tilde{\boldsymbol{A}} \triangleq \boldsymbol{A} - \bar{\boldsymbol{A}}$ denote the centered columns of $\boldsymbol{A}$, and noting both that $\overline{\boldsymbol{F}^T \boldsymbol{F} \otimes \boldsymbol{R}} = \overline{\boldsymbol{F}^T \boldsymbol{F}} \otimes \boldsymbol{R}$ and that $\boldsymbol{F}(\boldsymbol{F}^T \boldsymbol{F})^{-1} \boldsymbol{F}^T \boldsymbol{A} = \boldsymbol{A}$ when $\mathcal{H}_1$ is in force, properties of Kronecker products [163] can be used to show that (5.14) may be written as

$$\lambda = \sigma^{-2} \operatorname{tr}(\tilde{\boldsymbol{A}} \boldsymbol{R} \tilde{\boldsymbol{A}}^T); \tag{5.16}$$

details are provided in Appendix 5.A. Thus λ depends on the centered columns of $\boldsymbol{A}$, which contain the true but unknown coefficient trajectories $a_i[n]$ minus their respective mean values.

5.1.2.2 Model Order Selection

The above results yield not only a *practical* CFAR threshold-setting procedure, but also a full asymptotic description of the GLRT statistic of (5.6) under both $\mathcal{H}_0$ and $\mathcal{H}_1$. In light of this analysis, it is natural to ask how the TVAR model order p should be chosen in practice, along with the number q of non-constant basis functions. In deference to the large literature on the former subject [3], we adopt here the standard "2 coefficients per 1 kHz of speech bandwidth" rule of thumb.

Intuitively, the choice of basis functions should be well matched to the expected characteristics of the coefficient trajectories $a_i[n]$. To make this notion quantitatively precise, we appeal to the results of (5.13)–(5.16) as follows. First, the statistical *power* of our test to successfully detect small departures from stationarity is measured by the quantity $\Pr\left\{\chi_d^2(\lambda) > \gamma\right\}$. A result of [164] then shows that for fixed γ, the power function $\Pr\left\{\chi_d^2(\lambda) > \gamma\right\}$ is:

1. Strictly monotonically *increasing* in λ, for fixed d;
2. Strictly monotonically *decreasing* in d for fixed λ.

Each of these properties in turn yields a direct and important consequence for speech analysis:

- *Test power is maximized when λ attains its largest value:* For fixed p and q, the noncentrality parameter λ of (5.16) determines the power of the test as a function of σ^2 and the true but unknown coefficient trajectories $\boldsymbol{A}$.
- *Overfitting the data reduces test power:* Choosing p or q to be larger than the true data-generating model will result in a quantifiable loss in power, as λ will remain fixed while the degrees of freedom increase.

The first of these consequences follows from Property 1 above, and reveals how test power depends on the energy of the centered TVAR trajectories $\tilde{\boldsymbol{A}} = \boldsymbol{A} - \bar{\boldsymbol{A}}$ for fixed $\bar{\boldsymbol{A}}$ and p, q, σ^2. To verify the second consequence, observe that the product $\tilde{\boldsymbol{A}}\boldsymbol{R}\tilde{\boldsymbol{A}}^T$ remains unaffected by an increase in either p or q beyond that of the true TVAR(p) model. Then by Property 2, the corresponding increase in the degrees of freedom pq will lead to a loss of test power.

This analysis implies that care should be taken to adequately capture the energy of TVAR coefficient trajectories while guarding against overfitting; this formalizes our earlier intuition and reinforces the importance of choosing a relatively low-dimensional subspace formed by the span of low-frequency basis functions whose degree of smoothness is matched to the expected TVAR(p) signal characteristics under $\mathcal{H}_1$. This conclusion is further illustrated in Figure 5.3, which considers the effects of overfitting on the "formant-like" synthetic example of Section 5.1.1.4, with $p = 2$, $N = 100$ samples, $\delta = 7\pi/80$ radians, and piecewise-constant coefficient trajectories. Not only is the effect of overfitting p apparent in the left-hand panel, but the detection performance also suffers as the degree q of the Legendre polynomial basis is increased, as shown in the right-hand panel.

5.1.3 Relationship to Classical Approaches

We now relate our hypothesis testing framework to two classical approaches in the literature. First, we compare its performance to that of Brandt's test [147], which has seen wide use both in earlier [148, 149] and more recent studies [156, 165, 166], for purposes of transient detection, automatic segmentation for speech recognition, and concatenative speech synthesis. Second, we demonstrate its advantages relative to the autocorrelation method of time-varying linear prediction [94], showing that data windowing can adversely affect detection performance in this nonstationary setting.

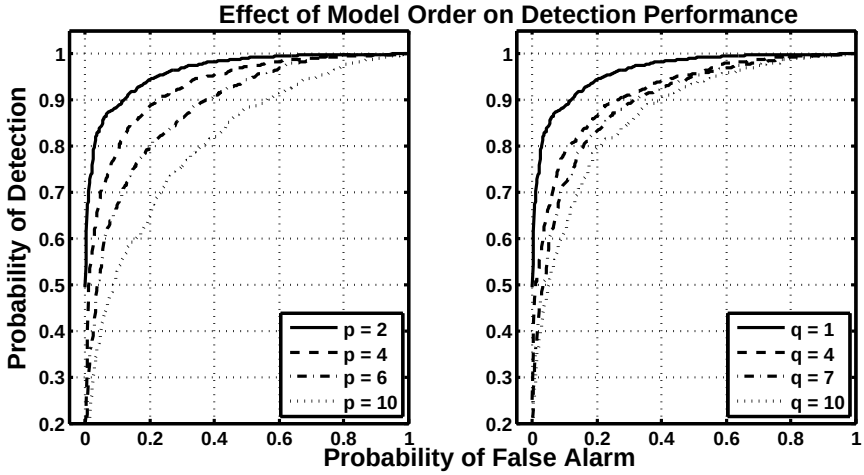


Figure 5.3: The effect of overfitting on the detection performance of the GLRT statistic for the synthetic signal of Figure 5.2. An increase in the model order—p (left) and q (right)—decreases the probability of detection at any CFAR level.

5.1.3.1 Classical Piecewise-Constant AR Approach

A related previous approach is to model $\boldsymbol{x}$ as an AR process with piecewise-constant parameters that can undergo at most a single change [167]. The essence of this approach, first employed in the speech setting by [147], is to split $\boldsymbol{x}$ into two parts according to $\boldsymbol{x} = (\boldsymbol{x}_r \,|\, \boldsymbol{x}_{N-r}) = (x[0] \cdots x[r-1] \,|\, x[r] \cdots x[N-1])^T$ for some *fixed* r, and to assume that under $\mathcal{H}_0$, $\boldsymbol{x}$ is modeled by an AR(p) process with parameters $\boldsymbol{\alpha}_0$, whereas under $\mathcal{H}_1$, $\boldsymbol{x}_r$ and $\boldsymbol{x}_{N-r}$ are described by *distinct* AR(p) processes with parameters $\boldsymbol{\alpha}_r$ and $\boldsymbol{\alpha}_{N-r}$, respectively.

In this context, testing for change in AR parameters at some *known* r can be realized as a likelihood ratio test; the associated test statistic $T'_r(\boldsymbol{x})$ is obtained by applying the covariance method to $\boldsymbol{x}$, $\boldsymbol{x}_r$, and $\boldsymbol{x}_{N-r}$ in order to estimate $\boldsymbol{\alpha}_0$, $\boldsymbol{\alpha}_r$, and $\boldsymbol{\alpha}_{N-r}$, respectively. However, since the value of r is *unknown* in practice, $T'_r(\boldsymbol{x})$ must also be maximized over r, yielding a test statistic $T'(\boldsymbol{x})$ as follows:

$$T'(\boldsymbol{x}) \triangleq \max_r T'_r(\boldsymbol{x}) \quad 2p \le r < N - 2p, \text{ with} \tag{5.17}$$

$$T'_r(\boldsymbol{x}) \triangleq \frac{\sup\limits_{\boldsymbol{\alpha}_r, \boldsymbol{\alpha}_{N-r}} p_{\mathcal{H}_1}(\boldsymbol{x}_r; \boldsymbol{\alpha}_r) p_{\mathcal{H}_1}(\boldsymbol{x}_{N-r}; \boldsymbol{\alpha}_{N-r})}{\sup\limits_{\boldsymbol{\alpha}_0} p_{\mathcal{H}_0}(\boldsymbol{x}; \boldsymbol{\alpha}_0)}. \tag{5.18}$$

We compared the detection performance of the GLRT statistic of (5.6) with that of (5.17) on both the piecewise-*constant* signal of Figure 5.2 and a piecewise-*linear* TVAR(2) signal to illustrate their respective behaviors—the resulting ROC curves are shown in Figure 5.4. In both cases, it is evident that the TVAR-based statistic of (5.6) has more power than that of (5.17), in part due to the extra variability introduced by maximizing over all values of r in (5.17)—especially those near the boundaries of its range. Even in the case of the piecewise-constant signal, correctly matched to the assumptions underlying (5.17), the

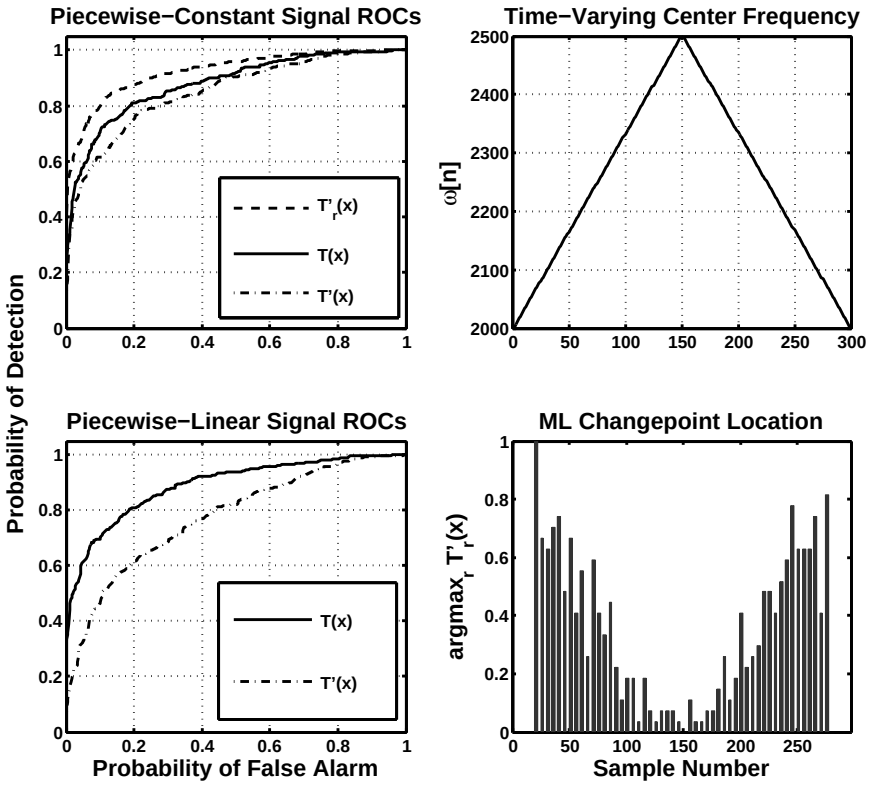


Figure 5.4: Comparing the detection performance of the statistic of (5.6) and that of (5.17): (top-left) comparison using the piecewise-constant signal ($N = 100$, $\delta = 5\pi/80$) of Section 5.1.1.4 with $p = 2$ and $q = 2$ Legendre polynomials used for computing (5.6); (top-right) piecewise-linear center frequency of the digital resonator used to generate the 2nd synthetic example; (bottom-left) comparison using the piecewise-linear signal ($N = 300$) with $p = 2$ and $q = 3$ Legendre polynomials used for computing (5.6); (bottom-right) histogram of the changepoint r that maximizes the test statistic $T'_r(\boldsymbol{x})$ for each instantiation of the signal with piecewise-linear TVAR coefficient trajectories.

TVAR-based test is outperformed only when r is known a priori, and (5.18) is used. This effect is particularly acute in the small sample size setting—an important consideration for the single-pitch-period case study of Section 5.1.5.

This example demonstrates that *any* estimates of r can be misleading under model mismatch. As shown in the bottom-right panel of Figure 5.4, the detected changepoint is often estimated to be near the start or end of the data segment, but no "true" changepoint exists since the time-varying center frequency is *continuously* changing. Thus piecewise-constant models are only simple approximations to potentially complex TVAR coefficient dynamics; in contrast, flexibility in the choice of basis functions implies applicability to a broader class of time-varying signals.

Note also that computing (5.17) requires brute-force evaluation of (5.18) for all values of r, whereas (5.6) need be calculated once. Moreover, $T'(\boldsymbol{x})$ fails to yield chi-squared (or any closed-form) asymptotics [167], thus precluding the design of a CFAR test and any quantitative evaluation of test power.

5.1.3.2 Classical Linear Prediction and Windowing

Recall that our GLRT formulation of Section 5.1.1, stemming from the TVAR model of (5.2), generalized the covariance method of linear prediction to the time-varying setting. As discussed in Section 4.3.2, the classical autocorrelation can also be generalized to the time-varying setting using the *synchronous* TVAR model

$$x[n] = \sum_{i=1}^{p} a_i[n-i]x[n-i] + \sigma w[n], \tag{5.19}$$

in lieu of (5.2). Grouping the coefficients α_{ij} into p vectors $\widetilde{\boldsymbol{\alpha}}_i \triangleq (\alpha_{i0}\ \alpha_{i1}\ \cdots\ \alpha_{iq})^T, 1 \leq i \leq p$, induces a partition of the expansion coefficients given by $\widetilde{\boldsymbol{\alpha}} \triangleq (\widetilde{\boldsymbol{\alpha}}_1^T\ \widetilde{\boldsymbol{\alpha}}_2^T\ \cdots\ \widetilde{\boldsymbol{\alpha}}_p^T)^T$—a permutation of elements of $\boldsymbol{\alpha}$ in (5.4). The autocorrelation estimator of $\widetilde{\boldsymbol{\alpha}}$ is then obtained by minimizing the prediction error over all $n \in \mathbb{Z}$, while assuming that $x[n] = 0$ for all $n \notin [0, \ldots, N-1]$, and is equivalent to the least-squares solution of the following linear regression problem:

$$\boldsymbol{x} = \widetilde{\boldsymbol{H}}_{\boldsymbol{x}}\widetilde{\boldsymbol{\alpha}} + \sigma\boldsymbol{w}, \tag{5.20}$$

where $\boldsymbol{w} = \big(w[0]\ \cdots\ w[N-1]\big)^T$ and the nth row of $\widetilde{\boldsymbol{H}}_{\boldsymbol{x}} \in \mathbb{R}^{N \times p(q+1)}$ is given by $(f_0[n-1]x[n-1]\ \cdots\ f_0[n-p]x[n-p]\ \cdots\ f_q[n-1]x[n-1]\ \cdots\ f_q[n-p]x[n-p])$. The *autocorrelation* estimate of $\widetilde{\boldsymbol{\alpha}}$ then follows from (5.20) as: [1]

$$\widehat{\widetilde{\boldsymbol{\alpha}}} = (\widetilde{\boldsymbol{H}}_{\boldsymbol{x}}^T\widetilde{\boldsymbol{H}}_{\boldsymbol{x}})^{-1}\widetilde{\boldsymbol{H}}_{\boldsymbol{x}}^T\boldsymbol{x}. \tag{5.21}$$

Moreover, when the autocorrelation method is used for spectral estimation in the stationary setting, $\boldsymbol{x}$ is often pre-multiplied by a smooth window. To empirically examine the role of data windowing in the time-varying setting, we generated a short 196-sample synthetic TVAR(2) signal $\boldsymbol{x}$ using $q = 0$ ($\mathcal{H}_0$) and $q = 2$ ($\mathcal{H}_1$) non-constant Legendre polynomials, and fitted $\boldsymbol{x}$ using AR(3) and TVAR(3) models—with the extra autoregressive order expected to capture the effects of data windowing. We then generated an ROC curve associated with the GLRT statistic of (5.11), shown in the top panel of Figure 5.5, along with ROC curves corresponding to an evaluation of (5.11) following the autocorrelation—rather than the covariance—method, both with and without windowing.

The bottom panel of Figure 5.5 shows the empirical distributions of both autocorrelation-based test statistics under $\mathcal{H}_0$, and indicates how windowing has the inadvertent effect of hindering detection performance in this setting. We have observed the effects of Figure 5.5 to be magnified for even shorter data records, implying greater precision of the covariance-based GLRT approach, which also has the advantage of known test statistic asymptotics under correct model specification.

5.1.4 Case Study I: Detecting Formant Motion

We now introduce a GLRT-based sequential detection algorithm to identify vocal tract variation on the scale of tens of milliseconds of speech data, and undertake a more

[1] As noted in Section 4.3.2, $\widetilde{\boldsymbol{H}}_{\boldsymbol{x}}^T\widetilde{\boldsymbol{H}}_{\boldsymbol{x}}$ is a *block-Toeplitz* matrix and may efficiently inverted using $\mathcal{O}(p^3(q+1)^2)$ operations [125].

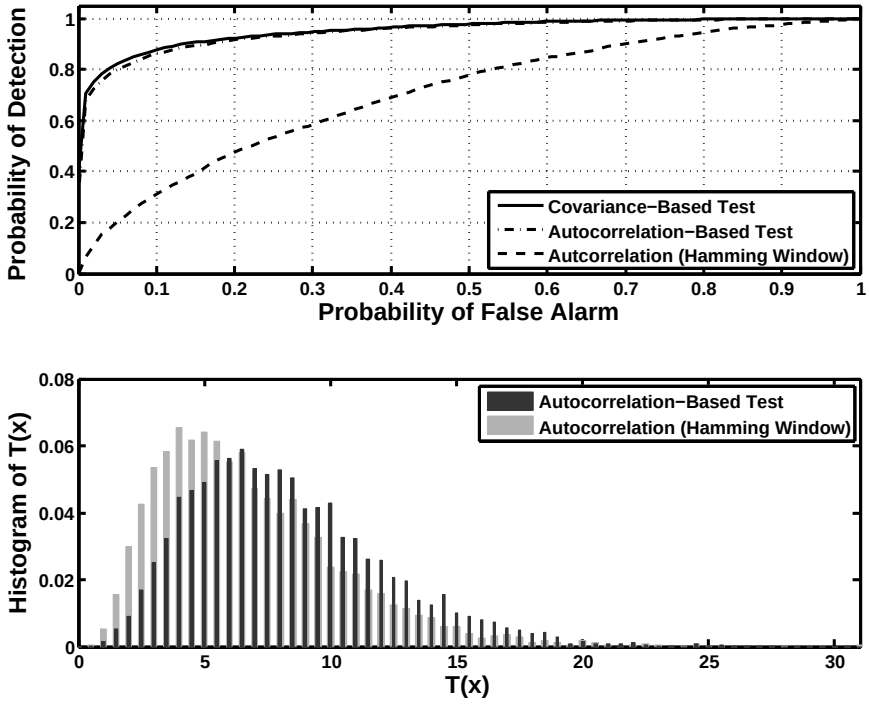


Figure 5.5: Comparison of covariance-and autocorrelation-based test statistics, based on 5000 trials with a short (196-sample) data record. Top: ROC curves showing the effects of data windowing on detection performance. Bottom: Detail of how windowing changes the distribution of $T(\boldsymbol{x})$ under $\mathcal{H}_0$.

refined analysis than that of Section 5.1.1.4 to demonstrate its efficacy on both whispered and voiced speech. Our results yield strong empirical evidence that appropriately specified TVAR models can capture vocal tract dynamics, just as AR models are known to provide a time-invariant vocal tract representation that is robust to glottal excitation type.

5.1.4.1 Sequential Change Detection Scheme

Our basic approach is to divide the waveform into a sequence of K short-time segments $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_K\}$ using shifts of a single N_0-sample rectangular window, and then to merge these segments, from left to right, until *spectral change* is detected via the GLRT statistic of (5.6). The procedure, detailed in Algorithm 5.1, begins by merging the first pair of adjacent short-time segments $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ into a longer segment $\boldsymbol{x}_m$ and computing $T(\boldsymbol{x}_m)$; failure to reject $\mathcal{H}_0$ implies that $\boldsymbol{x}_m$ is stationary. Thus, the short-time segments remain merged and the next pair considered is $(\boldsymbol{x}_m, \boldsymbol{x}_3)$. This procedure continues until $\mathcal{H}_0$ is rejected, indicating the presence of change within the merged segment under consideration. In this case, the scheme is re-initialized, and adjacent short-time segments are once again merged until a subsequent change in the spectrum is detected.

In Algorithm 5.1, the CFAR threshold γ of (5.12) is set *prior* to observing any data, by appealing to the asymptotic distribution of $T(\boldsymbol{x})$ under $\mathcal{H}_0$ developed in Section 5.1.2.1.

Algorithm 5.1 Sequential Formant Change Detector

1. Initialization: set γ via (5.12), input waveform data $\boldsymbol{x}$
   - Compute K short-time segments $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K\}$ of $\boldsymbol{x}$ using shifts of a rectangular window
   - Set $k = 1$, $\boldsymbol{x}_l = \boldsymbol{x}_1$, $\boldsymbol{x}_r = \boldsymbol{x}_2$
   - Set a marker array $\mathrm{C}[k] = 0$ for all $1 \leq k < K$
2. While $k < K$
   - Set $\boldsymbol{x}_m = \boldsymbol{x}_l + \boldsymbol{x}_r$ and compute $T(\boldsymbol{x_m})$ via (5.6)
   - If $T(\boldsymbol{x_m}) < \gamma$ (no formant motion within $\boldsymbol{x}_m$)
     - Set $\boldsymbol{x}_l = \boldsymbol{x}_m$, $\mathrm{C}[k] = 0$

     Else (formant motion detected within $\boldsymbol{x}_m$)
     - Set $\boldsymbol{x}_l = \boldsymbol{x}_k$, $\mathrm{C}[k] = 1$
   - Set $\boldsymbol{x}_r = \boldsymbol{x}_{k+1}$, $k = k + 1$
3. Return the set of markers $\{k : \mathrm{C}[k] = 1\}$

In principle, the time resolution to within which change can be detected is limited only by N_0. Using arbitrarily short windows, however, increases the variance of the test statistic and results in an increase in false alarms—a manifestation of the Fourier uncertainty principle. Decreasing γ also serves to increase the (constant) false alarm rate, and leads to spurious labeling of local fluctuations in the estimated coefficient trajectories (e.g., due to the position of the sliding window relative to glottal closures) as vocal tract variation.

5.1.4.2 Evaluation with Whispered Speech

In order to evaluate the GLRT in a gradually more realistic setting, we first consider the case of whispered speech to avoid the effects of voicing, and apply the formant change detection scheme of Algorithm 5.1 to whispered utterances containing *slowly-varying* and *rapidly-varying* spectra, respectively.

The waveform used in the first experiment comprises a whispered vowel [ɑ] (as in "father") followed by a diphthong [aɪ] (as in "liar"). It was downsampled to 4 kHz in order to focus on changes in the *first two* formants, and Algorithm 5.1 was applied to this waveform as well as to its 0–1 kHz and 1–2 kHz subbands (containing the first and second formants, respectively).

Results are summarized in Figure 5.6, and clearly demonstrate that the GLRT is sensitive to formant motion. All three spectrograms indicate that spectral change is first detected near the boundary of the vowel and diphthong—precisely when the vocal tract configuration starts to change. Subsequent consecutive changes are found when sufficient formant change has been observed relative to data duration—a finding consistent with our earlier observation in Section 5.1.1.4 that more data are required to detect small changes in the AR coefficient trajectories, and by proxy the vocal tract, at the same level of statistical significance (i.e., same false alarm rate).

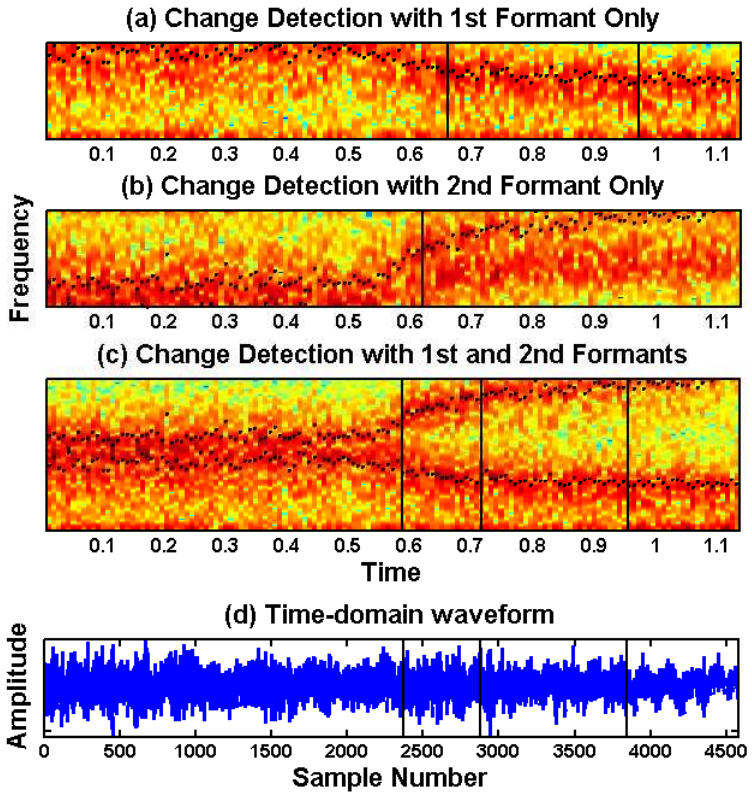


Figure 5.6: Result of applying Algorithm 5.1 (16 ms rectangular windows, $p = 4$, $q = 2$ Legendre polynomials, 1% CFAR) to detect formant movement in the whispered waveform [ɑaɪ]. Spectrograms corresponding to subbands containing the first formant only (a) second formant only (b) and both formants (c) were computed using 16 ms Hamming windows with 50% overlap, and are overlaid with formant tracks computed by WaveSurfer [49]. Black lines demarcate times at which formant motion was detected; the time-domain waveform overlaid with these boundaries is also shown (d).

Next observe that whereas three "changepoints" are found when the waveform contains two moving resonances, a *total* of three "changepoints" are marked in the single-resonance waveforms shown in Figs. 5.6(a) and 5.6(b). Intuitively, each of these signals can be thought of as having "less" spectral change than the waveform shown in Figure 5.6(c), which contains both formants. Thus, since the corresponding amounts of spectral change are smaller, longer short-time segments are required to detect formant movement—as indicated by the delays in detecting the vowel-diphthong transition seen in Figs. 5.6(a) and (b) relative to (c).

We next conducted a second experiment to demonstrate that the GLRT can also detect a more rapid onset of spectral change as compared to, e.g., the relatively slow change in the spectrum of the diphthong. To this end we applied Algorithm 5.1 to a sustained whispered vowel ([i ː] as in "beet"), followed by the plosive [t] at 10 kHz. The results, shown in Figure 5.7, indicate that no change is detected during the sustained vowel, whereas the plosive is clearly identified.

Finally, we have observed change detection results such as these to be robust to

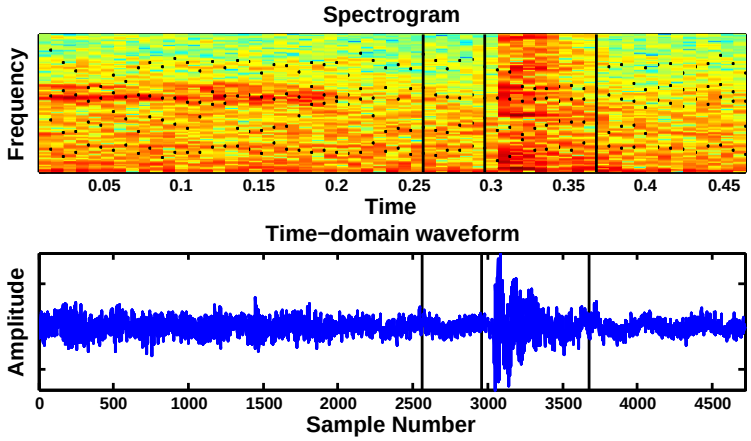


Figure 5.7: Algorithm 5.1 (16 ms windows, $p=10$, $q=4$ Legendre polynomials, 1% CFAR), applied to detect formant movement in the whispered waveform [i ː] t]. Its spectrogram (top) is overlaid with formant tracks computed by WaveSurfer [49] and black lines demarcating the time instants at which formant motion was detected; the time-domain signal is also shown (bottom).

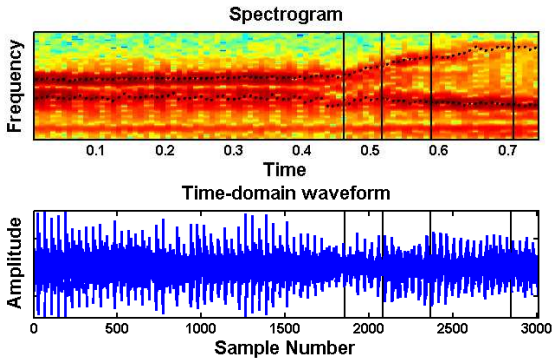
not only reasonable choices of p (roughly 2 coefficients per 1 kHz of speech bandwidth) and q (1–10), but also to the size of the initial window length (10–40 ms), and the constant false alarm rate (1–20%).

5.1.4.3 Extension to Voiced Speech

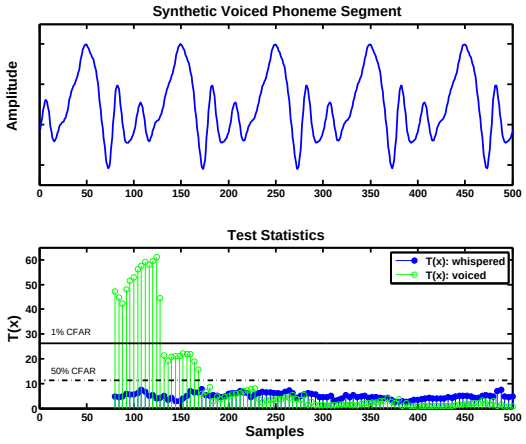
We next conducted an experiment to show that the TVAR-based GLRT is *robust* to the presence of voicing. We repeated the first experiment of Section 5.1.4.2 above using a *voiced* vowel-diphthong pair [ɑaɪ] over the range 0–4 kHz. The same parameter settings were employed, except for the addition of two poles to take into account the shape of the glottal pulse during voicing [3]. Algorithm 5.1 yields the results shown in Figure 5.8(a), which parallel those shown in Figure 5.6 for the whispered case. Indeed, the first change occurs at approximately the vowel-diphthong boundary, with subsequent "changepoints" marked when sufficient formant movement has been observed.

The similarities in these results are due in part to the fact that the analysis windows employed in both cases span at least one pitch period. To wit, consider the synthesized voiced phoneme [ɑ] and the associated GLRT statistic of (5.6) shown in the top and bottom panels of Figure 5.8(b), respectively. Even though the formants of the synthesized phoneme are constant, the value of $T(\boldsymbol{x})$ undergoes a stepwise decrease from over the 1% CFAR threshold when < 1 pitch period is observed, to just above the 50% CFAR threshold when < 1.5 periods are observed—and finally stabilizes to a level below the 50% CFAR threshold after more than two periods are seen. In contrast, the GLRT statistic computed for the associated *whispered* phoneme, generated by filtering white noise by a vocal tract parameterized by the *same* formant values and shown in the bottom panel of Figure 5.8(b), remains time-invariant.

These results indicate that the periodic excitation during voicing has negligible



(a) Spectrogram of the voiced waveform [ɑaɪ] is overlaid with formant tracks computed by WaveSurfer [49] and black lines demarcating the time instants at which formant motion was detected; the time domain signal is shown for reference (bottom).



(b) Formant-synthesized voiced phoneme [ɑ] (top) and associated GLRT statistic (bottom, green) are shown along with 1% (solid black) and 50% CFAR (dashed-black) thresholds. Window lengths of $5 - 35$ ms at 1 ms (16-sample) increments with $p = 6$, $q = 3$ Legendre polynomials were used to calculate $T(\boldsymbol{x})$. Values of $T(\boldsymbol{x})$ for a whispered [ɑ] (bottom, blue) generated using the same formant values are shown for comparison.

Figure 5.8: Detecting vocal tract dynamics in voiced speech (a) and the impact of the quasi-periodic glottal flow on the GLRT statistic $T(\boldsymbol{x})$ (b).

impact on the GLRT statistic when longer (i.e., $>$ 2 pitch periods) speech segments are used, and explain the robustness of the GLRT statistic $T(\boldsymbol{x})$ to the presence of voicing in the experiments of this section. On the other hand, the GLRT is sensitive to the glottal flow when shorter speech segments are employed, suggesting that it can be also used effectively on sub-segmental time scales, as we show in Section 5.1.5.

5.1.5 Case Study 2: Sub-Segmental Speech Analysis

We now demonstrate that our GLRT framework can be used not only to detect formant motion across multiple pitch periods, as discussed above in Section 5.1.4, but also to detect vocal tract variations *within* individual pitch periods. Since the vocal tract configuration is relatively constant during the glottal airflow closed phase, and undergoes change at its boundaries [42], a hypothesis test for vocal tract variation provides a natural way to identify both glottal opening and closing instants within the same framework.

We show below that this framework is especially well suited to detecting the gradual change associated with glottal openings, and can also be used to successfully detect glottal closures. Glottal closure identification is a classical problem (see, e.g., [159] for a recent review), with mature engineering solutions typically based on features of the linear prediction residual or the group delay function (see, e.g., [158, 159, 161] and references therein).

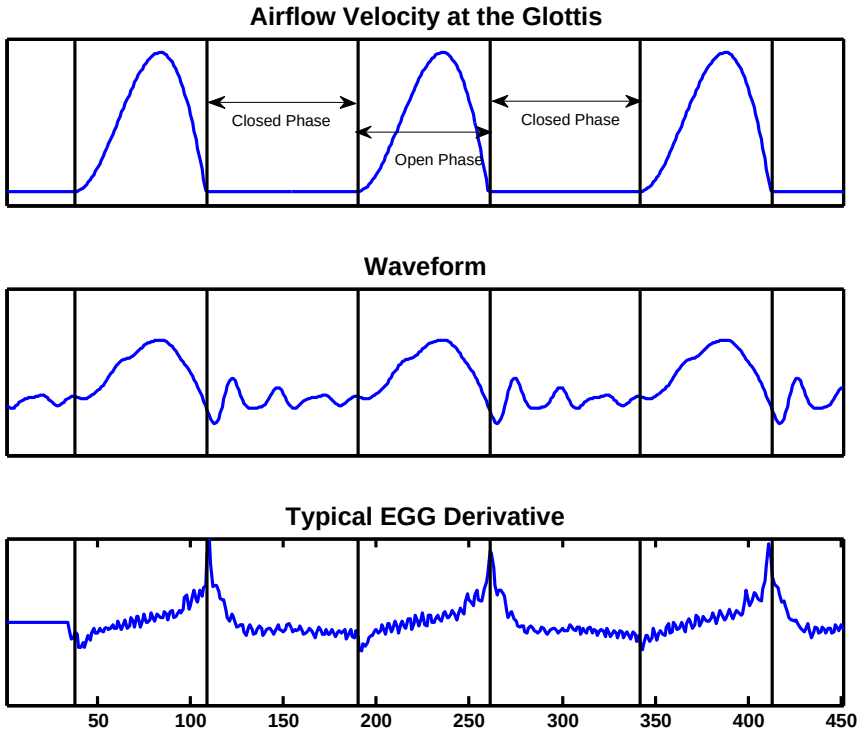


Figure 5.9: Glottal openings and closures demarcated over two pitch periods of a typical vowel, shown with idealized glottal flow (top), speech (middle), and EGG derivative (bottom) waveforms as a function of time.

In contrast, the slow onset of the open phase results in a difficult detection problem, and glottal opening detection has received relatively little attention in the literature [161], with preliminary results reported only in recent conference proceedings [168, 169].

5.1.5.1 Physiology of Sub-Segmental Variations

Figure 5.9 illustrates the idealized open and closed glottal phases associated with a typical vowel, along with the corresponding waveform and derivative electroglottograph (DEGG) data indicating approximate opening and closing instants [160, 170]. As discussed earlier in Section 2.2.1, the glottal closure instant (GCI) in each pitch period is defined as the moment at which the vocal folds close, and marks the start of the closed phase—an interval during which no airflow volume velocity is measured at the glottis (top panel), and the acoustic output at the lips takes the form of exponentially-damped oscillations (middle panel). Nominally, the glottal opening instant (GOI) indicates the start of the open phase: the vocal folds gradually begin to open until airflow velocity reaches its maximum amplitude, after which they begin to close, leading to the next GCI.

Time-invariance of the vocal tract suggests the use of linear prediction to estimate formant values during the closed phase [3], and then to use changes in these values across sliding windows to determine GOI and GCI locations [42]. Indeed, as the vocal folds

begin to open at the GOI, the vocal tract gradually lengthens, resulting in a change in the frequency and bandwidth of the first formant [50]—an effect that can be explained by a source-filter model with a *time-varying* vocal tract. Furthermore, the assumption that short-term statistics of the speech signal undergo maximal change in the vicinity of a GCI implies that such regions will exhibit large linear-prediction errors.

5.1.5.2 Detection of Glottal Opening Instants

We first give a sequential algorithm to detect GOIs via the GLRT statistic $T(\boldsymbol{x})$. To study the efficacy of the proposed method, we assume that the timings of the glottal closures are available, and use these to process each pitch period *independently*. In addition to evaluating the absolute error rates of our proposed scheme using recordings of sustained vowels, we also compare it with the method of [158]—a standard prediction-error-based approach that remains in wide use, and effectively underlies more recent approaches such as [168].

Sequential GOI Detection Procedure In contrast to the "merging" procedure of Algorithm 5.1, our basic approach here is to scan a sequence of short-time segments $\boldsymbol{x}_w$, induced by shifts of an N_0-sample rectangular window initially left-aligned with a glottal closure instant, until spectral change is detected via the GLRT statistic of (5.6).

At each iteration, the window slides one sample to the right, and $T(\boldsymbol{x}_w)$ is evaluated; this procedure continues until $T(\boldsymbol{x}_w)$ exceeds a specified CFAR threshold γ, indicating that *spectral change* was detected, and signifying the beginning of the open phase. In this case, the GOI location is declared to be at the *right* edge of $\boldsymbol{x}_w$. On the other hand, a missed detection results if a GOI has not been identified by the time the right edge of the sliding window coincides with the next glottal closure instant. The exact procedure is summarized in Algorithm 5.2.

Algorithm 5.2 Sequential Glottal Opening Instant Detector

1. Initialization: input one pitch period of data $\boldsymbol{x}$ between two consecutive glottal closure locations g_1 and g_2
   - Set $w_l = g_1$, $w_r = w_l + N_0$, and set γ via (5.12)
   - Set $\boldsymbol{x}_w = (x[w_l] \ \cdots \ x[w_r])$
2. While $T(\boldsymbol{x}_w) < \gamma$ and $w_r < g_2$
   - Increment w_l and w_r (slide window to right)
   - Recompute $\boldsymbol{x}_w$ and evaluate $T(\boldsymbol{x}_w)$
3. If $w_r < g_2$, then return w_r as the estimated glottal opening location, otherwise report a missed detection.

Since each instantiation of Algorithm 5.2 is confined to a *single* pitch period, the parameters N_0, p, and q must be chosen carefully. To ensure robust estimates of the TVAR coefficients, the window length N_0 cannot be too small; on the other hand, if it exceeds

the length of the entire closed-phase region, then the GOI cannot be resolved. Likewise, choosing a small number of TVAR coefficients results in smeared spectral estimates, whereas using large values of p leads to high test statistic variance and a subsequent increase in false alarms; this same line of reasoning also leads us to keep q small. Thus, in all the experiments reported in Section 5.1.5.2, we employ $N_0 = 50$-sample windows, $p = 4$ TVAR coefficients and the first 2 Legendre polynomials as basis functions ($q = 1$). We also evaluated the robustness of our results with respect to these settings, and observed that using window lengths of 40–60 samples, 3–6 TVAR coefficients, and 2–4 basis functions also leads to reasonable results in practice.

Evaluation We next evaluated the ability of Algorithm 5.2 to identify the glottal opening instants in five sustained vowels uttered by a male speaker (109 Hz average F0), synchronously recorded with an EGG signal (Center for Laryngeal Surgery and Voice Rehabilitation, Massachusetts General Hospital), and subsequently downsampled to 16 kHz. The Speech Filing System [171] was used to extract DEGG peaks and dips, which in turn provided a means of experimentally measured ground truth for our evaluations.

A typical example of GOI detection is illustrated in Figure 5.10, which shows the results of applying Algorithm 5.2 to an excised segment of the vowel [ɑ]. The detected GOI in this example was declared to be at the right edge of the first short-time segment $\boldsymbol{x}_w$ for which $T(\boldsymbol{x}_w)$ exceed the 15% CFAR threshold γ, and is marked by a dashed black line in all four panels of Figure 5.10. As can be seen in the bottom-right panel, the estimated GOI coincides precisely with a dip in the DEGG waveform. Moreover, as the top-right panel shows, this location corresponds to a significant change in the estimated coefficient trajectories, likely due both to a change in the frequency and bandwidth of the first formant (resulting from nonlinear source-filter interaction [42, 50]), as well as an increase in airflow volume velocity (from zero) at the start of the open phase.

Detection rates were then computed over 75 periods of each vowel, and detected GOIs were compared to DEGG dips in every pitch period that yielded a GOI detection. The resultant detection rates and root mean-square errors (RMSE, conditioned on successful detection) are reported in Table 5.2, along with a comparison to the prediction-error-based approach of [158], which we now describe.

Table 5.2: GOI Detection Accuracy (ms, No. missed detections).

	[ɑ]	[e]	[i]	[o]	[u]
GLRT RMSE (ms)	0.69	1.03	1.00	1.15	0.69
WMG [158] RMSE (ms)	1.04	1.78	1.13	1.97	1.10
GLRT Missed Det.	0	5	0	0	0
WMG [158] Missed Det.	8	18	4	6	4

Comparison with approach of Wong, Markel, and Gray (WMG) [158] The approach of [158] involves first computing a normalized error measure $\eta(\boldsymbol{x}_w)$ for each short-time segment $\boldsymbol{x}_w$ (induced by a sliding window as in Algorithm 5.2), and then identifying

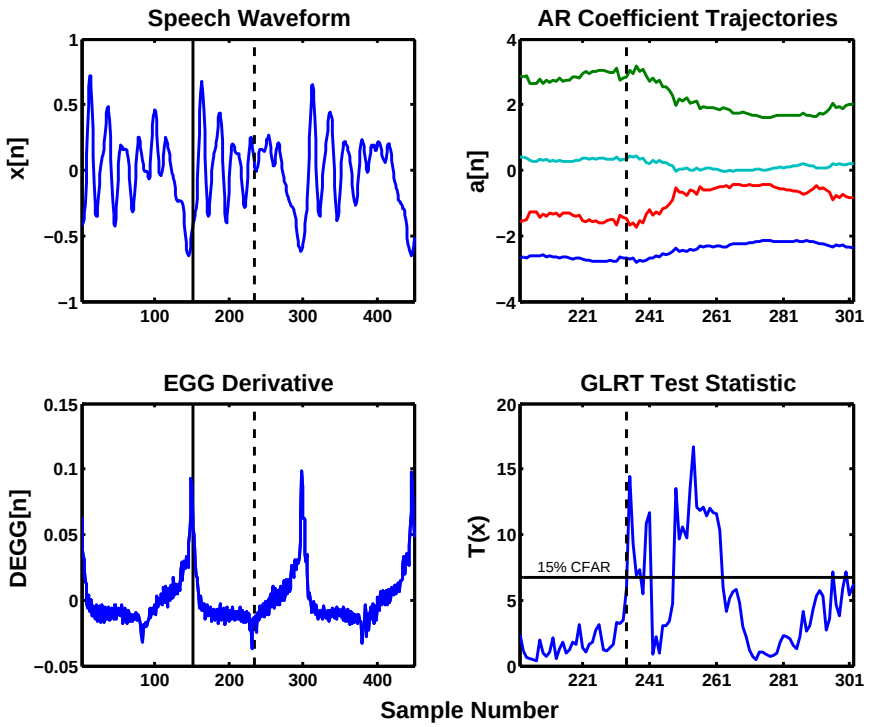


Figure 5.10: Algorithm 5.2 applied to detect the GOI in a pitch period of the vowel [ɑ] (top left), shown together with its EGG derivative (bottom left). The sliding window is left-aligned with the GCI (solid black line); estimated AR coefficients (top right) and the GLRT statistic $T(\boldsymbol{x})$ of (5.6) (bottom right) are then computed for each subsequent window position. The detected GOI (dashed black line) corresponds to the location of the first determined change (at the 15% CFAR level) in vocal tract parameters.

the GOI instant with the right edge of $\boldsymbol{x}_w$ when a large increase in $\eta(\boldsymbol{x}_w)$ is observed. The measure $\eta(\boldsymbol{x}_w)$ is obtained by fitting a *time-invariant* AR(p) model to $\boldsymbol{x}_w$ (using (5.9) with $q = 0$), calculating the norm of the resultant prediction error, and normalizing by the energy of short-time segment $\boldsymbol{x}_w$.

Figure 5.11 provides a comparison of this approach to that of Algorithm 5.2, over 8 periods of the vowel [ɑ]. Here Algorithm 5.2 is implemented with a 15% CFAR level, but the threshold for $\eta(\boldsymbol{x})$ must be set manually, since no theoretical guidelines are available [158]. Indeed, as illustrated in the bottom panel of Figure 5.11, variability in the dynamic range of $\eta(\boldsymbol{x})$ across pitch periods implies that *any* fixed threshold will necessarily introduce a tradeoff between detection rates and RMSE. In this example, lowering the threshold to intersect with $\eta(\boldsymbol{x})$ in the second pitch period—and thereby removing the missed detection—results in a 25% increase in RMSE.

The denominator of the GLRT statistic $T(\boldsymbol{x}_w)$ depends on the *same* prediction error residual used to calculate $\eta(\boldsymbol{x}_w)$; however, as indicated by Figure 5.11, it remains much more stable across pitch periods. Thus, while the approach of [158] relies on large *absolute* changes in AR residual energy to detect glottal openings, that of Algorithm 5.2 explicitly

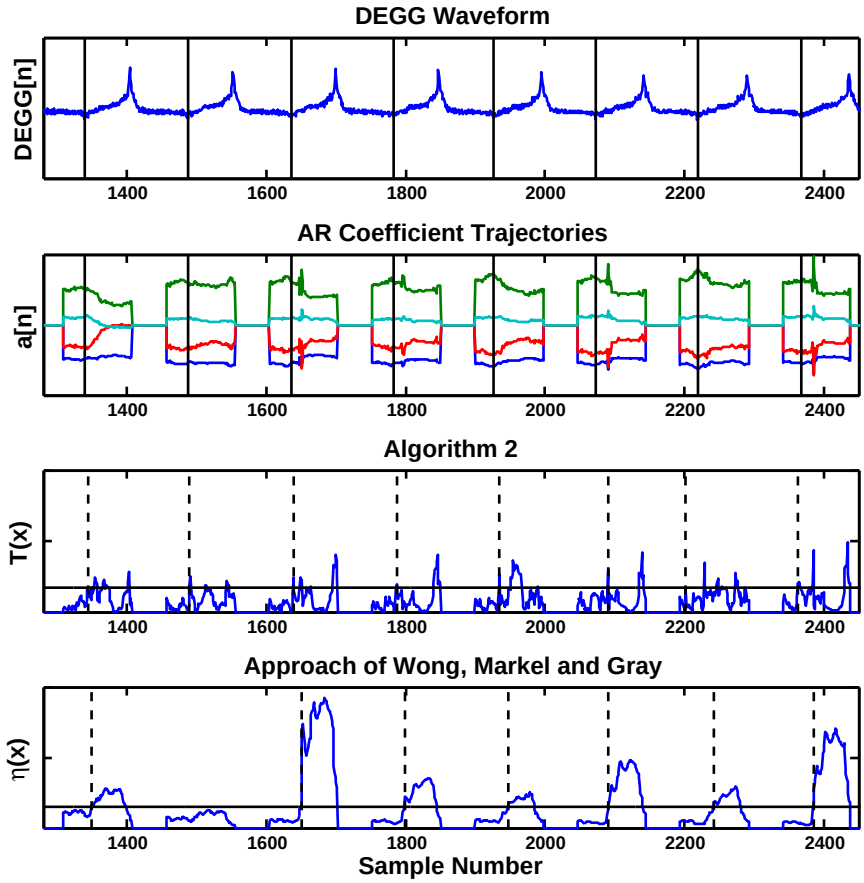


Figure 5.11: Comparison of Algorithm 5.2 and the approach of [158] for GOI detection. The EGG derivative for 8 periods of the vowel [ɑ], and estimated AR coefficients, are shown for all sliding window positions (top two panels) along with the associated values of $T(\boldsymbol{x})$ and $\eta(\boldsymbol{x})$ (bottom two panels). True and estimated GOI locations are indicated by solid and dashed black lines, respectively. Note the variability in the dynamic range of $\eta(\boldsymbol{x})$ from one pitch period to the next, and the missed detection (2nd pitch period, bottom panel).

takes into account the *ratio* of AR to TVAR residual energies—resulting in improved overall performance. Indeed, though thresholds were set individually for each vowel of Table 5.2, and manually adjusted to obtain the best RMSE performance while keeping the number of missed detections reasonably small, Algorithm 5.2 with a 15% CFAR threshold exhibits both superior detection rates *and* RMSE.

5.1.5.3 Detection of Glottal Closure Instants

Although our main focus here is on GOI detection, the GLRT statistic of (5.6) may also be employed to detect glottal closures. Indeed, under the assumption stated earlier that the speech signal undergoes locally maximal change in the vicinity of a GCI, a simple GCI detection algorithm immediately suggests itself: Compute (5.6) for every location of an appropriate sliding analysis window, and declare the glottal closure to occur at the *midpoint*

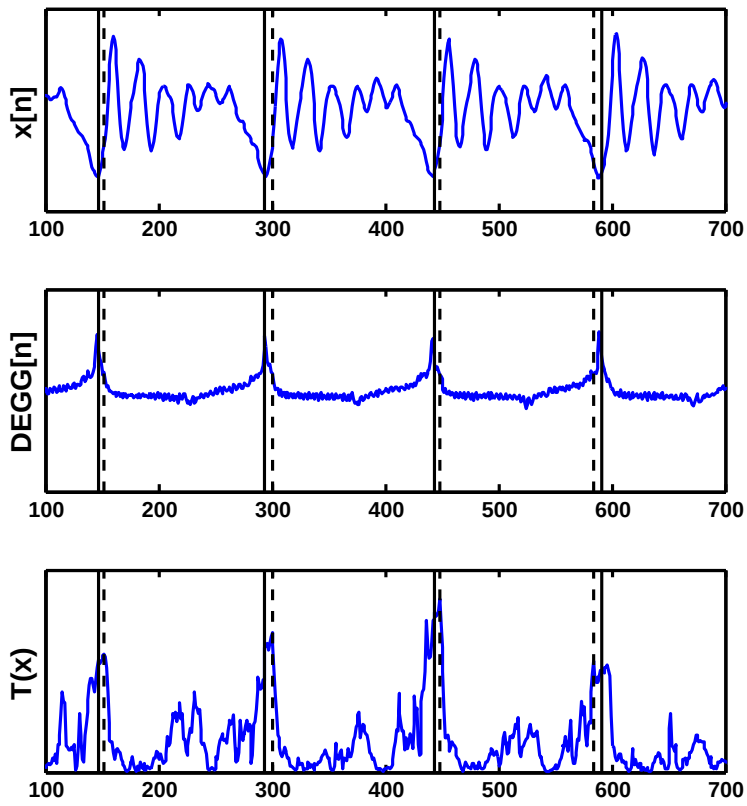


Figure 5.12: Using the GLRT statistic of (5.6) to find GCI locations in a segment of the vowel /ɑ/. The speech waveform (top), the EGG derivative (middle) and the GLRT statistic (bottom) are overlaid with the true (solid, black line) and estimated (dashed, black line) glottal closure locations in each pitch period.

of the window with the largest associated value of $T(\boldsymbol{x})$. In this formulation, $T(\boldsymbol{x})$ is being treated simply as a signal with features that may be helpful in finding the GCI locations; no test threshold need be set. A typical result is shown in the third panel of Figure 5.12, obtained using the same parameter settings (50-sample window, $p = 4, q = 2$) as in the GOI detection scheme of Section 5.1.5.2.

We compared this method to two others based on linear prediction and group delay, as described above. First, we implemented the alternative likelihood-ratio epoch detection (LRED) approach of [149], which tests for a single change in AR parameters. Second we used the "front end" of the popular DYPSA algorithm for GOI detection [161], comprising the generation of GCI candidates and their weighting by the ideal phase-slope function deviation cost as implemented in the Voicebox online toolbox [172]. Table 5.3 summarizes the GCI estimation results under the same conditions as reported in Table 5.2. All three methods are comparable in terms of accuracy, though the GLRT approach proposed here can be used—with the same parameter settings—for both GCI and GOI detection.

Results from both our approach and the DYPSA front end can in turn be propagated across pitch periods (using, e.g., dynamic programming [161]) to inform a broader

Table 5.3: GCI Detection Accuracy (ms)

Vowel/ GCI RMSE	[ɑ]	[e]	[i]	[o]	[u]
GLRT ($N_0 = 50, p = 4, q = 2$)	0.47	1.05	0.73	0.97	1.03
LRED [149] ($N_0 = 72, p = 6$)	1.02	0.69	1.00	0.65	1.12
DYPSA Front End [161]($N_0 = 50$)	0.61	0.68	0.70	0.79	1.10

class of group-delay methods [159], though we leave such a system-level comparison as the subject of future work.

5.2 Nonparametric Testing: Empirical Short-Time Coefficients and the Bootstrap

We now examine the question of testing for stationarity in a broader context. When processes are characterized by parametric models, the constancy of the model parameters over time is necessary to establish stationarity. Time-varying AR models treated in Chapter 4 and in Section 5.1 are canonical examples—AR coefficients must be time-invariant for the associated process to be stationary. While TVAR models are well-suited for modeling speech and TVAR-based hypothesis tests are appropriate for establishing the stationarity of speech waveforms, no appropriate parametric model may be available in many other applications, rendering parametric tests of limited use. In this case, however, the time-invariance of coefficients arising in nonparametric (transform-based) representations could be used to test for stationarity. Indeed, a number of test statistics based on this idea have been proposed in the extant literature [173–177].

In addition to specifying a test statistic, its distribution under the null hypothesis of stationarity is necessary to set thresholds when using the associated hypothesis test in practice. In some cases, it is possible to obtain these distributions [173,174,176,178] under certain regularity assumptions and in an appropriate asymptotic regime. However, these results are only approximate when the number of observations is small.

Our main contribution is an efficient Monte Carlo or bootstrap method for characterizing the null distribution of two test statistics based on empirical short-time Fourier coefficients. Accordingly, in Section 5.2.1, we define two test statistics that measure the temporal variation of short-time spectral coefficients, and characterize their sampling distributions under the null hypothesis of stationarity for the case of a white Gaussian noise process, in Section 5.2.2.1. Based on these calculations, we introduce, in Section 5.2.2.2, a simple and efficient Monte Carlo procedure based on the Wold representation to simulate from the null distribution. The approach is similar to the one developed independently in [177], but their method is based on alternative definitions of the null hypothesis and test statistics. After validating the procedure on synthetic examples, we demonstrate one potential application as a method of obtaining a signal-adaptive means of local Fourier analysis and corresponding signal enhancement scheme, in Section 5.2.3.

5.2.1 Hypothesis Test For Wide-Sense Stationarity

Given N observations $\boldsymbol{x} \triangleq \begin{pmatrix} x[0] & x[1] & \cdots & x[N-1] \end{pmatrix}$ of a zero-mean discrete-time random process $(x[n], n \in \mathbb{Z})$, we are interested in testing the hypothesis:

$$\begin{aligned} &\mathcal{H}_0 : x[n] \text{ is wide-sense stationary (WSS)} \\ &\mathcal{H}_1 : x[n] \text{ is nonstationary.} \end{aligned} \tag{5.22}$$

Wide-sense stationarity of $x[n]$ implies that its autocorrelation function $r_{xx}[n,m] \triangleq E(x[n]x[m])$ depends only on the lag $\tau = n - m$, therefore we write $r_{xx}[n,m] = r_{xx}[\tau]$ in this case. If we partition the N observations into M non-overlapping rectangular windows each of length L, then we may use L measurements within the mth segment (i.e., window) to obtain an unbiased estimate of a *local* autocorrelation function $r_{xx}^m[\tau]$ defined over $0 \leq |\tau| \leq L/2 - 1$ via

$$\widehat{r_{xx}^m[\tau]} \triangleq \frac{1}{L - |\tau|} \sum_{l=0}^{L/2-1} x[l]x[l+\tau],$$

and the associated power spectral density (PSD) via

$$\widehat{S_{xx}^m}[k] = \frac{1}{\sqrt{L}} \sum_{\tau=-L/2}^{L/2-1} \widehat{r_{xx}^m[\tau]} e^{-i2\pi k\tau/L}. \tag{5.23}$$

Observe that if the process $x[n]$ were wide-sense stationary, then the variation among $\{\widehat{S_{xx}^1}[k], \ldots, \widehat{S_{xx}^M}[k]\}$—the set of estimated PSD values at frequency bin k—should be due *only* to the variance of the underlying power spectral density estimator (which in this case is the periodogram weighted by a sinc^2 kernel). On the other hand, if $x[n]$ were nonstationary in the second moment such that the frequency content of the underlying process were time-varying, we would expect to observe greater variation in M PSD values than could just be attributed to the estimation procedure.

Indeed, in the absence of estimator variance, $x[n]$ would be wide-sense stationary only if the relation $\widehat{S_{xx}^m}[k] = \widehat{S_{xx}^{m'}}[k]$ were to hold for every frequency bin $0 \leq k \leq (L-1)/2$ and all $m, m' \in [1, \ldots, M]$. This condition is necessary, but not sufficient since the power spectrum must be constant for all $\omega \in [0, 2\pi)$ not just for the set of discrete Fourier frequencies. Note, also that when the process $x[n]$ is nonstationary, we can still loosely interpret the quantity $\widehat{S_{xx}^m}[k]$ in (5.23) as the Fourier transform of the estimated *frozen-time* autocorrelation of the process.

Consequently, consider two test statistics that measure the amount of spectral variation in windowed data segments over time:

$$\widehat{V}(\boldsymbol{x}) \triangleq \frac{1}{ML} \sum_{k=0}^{L/2-1} \sum_{m=0}^{M-1} \left(\widehat{S_{xx}^m}[k] - \frac{1}{M} \sum_{p=0}^{M-1} \widehat{S_{xx}^p}[k] \right)^2 \tag{5.24}$$

$$\widetilde{V}(\boldsymbol{x}) \triangleq \frac{1}{ML} \sum_{k=0}^{L/2-1} \sum_{m=0}^{M-1} \left(\widetilde{S_{xx}^m}[k] - \frac{1}{M} \sum_{p=0}^{M-1} \widetilde{S_{xx}^p}[k] \right)^2 \tag{5.25}$$

where $\widehat{S_{xx}^m}[k]$ and $\widetilde{S_{xx}^m}[k]$ are the periodogram and multitaper periodogram [179] estimators of the PSD given by:

$$\widehat{S_{xx}^m}[k] \triangleq |X_m^w[k]|^2 \quad \text{and} \quad \widetilde{S_{xx}^m}[k] \triangleq \frac{1}{R}\sum_{r=1}^{R} |X_m^{w_r}[k]|^2. \tag{5.26}$$

Here, $X_m^w[k]$ is the discrete short-time Fourier Transform (STFT) defined by

$$X_m^w[k] = \frac{1}{\sqrt{L}} \sum_{n=mL+1}^{mL+L} w[n-Lm]x[n]e^{-2\pi ikn/L}, \tag{5.27}$$

with w is a rectangular window of length L, and $X_m^w[k]$ is the kth frequency component of the mth STFT window. The multitaper spectrum $X_m^{w_r}[k]$ is found by choosing the window $w = w_r$ in (5.27) to be the rth of R discrete prolate spheroidal sequences [179]. Note that the assumed that the windows do not overlap will be relaxed later on.

The main motivation behind considering two distinct test statistics is that even though both $\widehat{V}(\boldsymbol{x})$ and $\widetilde{V}(\boldsymbol{x})$ are asymptotically unbiased [179], the latter may have lower variance since the variance of the multitaper PSD estimator is smaller than that of the periodogram estimator. Specifically, we have that as L grows, $\mathrm{Var}(\widehat{S_{xx}^m}[k])$ tends to $(S_{xx}^m[k])^2$, while $\mathrm{Var}(\widetilde{S_{xx}^m}[k])$ tends to $(S_{xx}^m[k])^2/R$. Thus, for fixed R and large L, $\mathrm{Var}(\widetilde{S_{xx}^m}[k]) < \mathrm{Var}(\widehat{S_{xx}^m}[k])^2$. This latter fact is crucial—lower variance estimators of the PSD imply higher sensitivity of (5.25) to changes in the signal statistics, and tests based on $\widehat{V}(\boldsymbol{x})$ can be expected to be more powerful than those based on $\widetilde{V}(\boldsymbol{x})$.

To illustrate this point, we compare the *relative* detection performance of the STFT- and multitaper-based test statistics using synthetic time-varying MA(2) and time-varying AR(2) signals—corresponding to smooth and peaky spectra, respectively. The observations, each of length 5120, were obtained by changing the MA or AR coefficients at the midpoint of the signal. In particular, to generate data under $\mathcal{H}_1$, the initial MA coefficients of $(1, 0.4)$ were changed to $(1, 0.4 + \delta)$ with $\delta \in (0.1, 0.2)$ and the initial AR coefficients $(-1.273, 0.81)$ were changed to $(-1.196, 0.81)$ corresponding to a $\pi/50$ Hz shift in the center frequency of the associated bandpass filter. To generate data under $\mathcal{H}_0$ the initial parameters were simply left unchanged.

The test statistics of (5.24) and (5.25) were computed using 10 adjacent 512-sample rectangular windows; $R = 6$ tapers were used to obtain $\widetilde{V}(\boldsymbol{x})$. Six hundred Monte Carlo simulations were done for these scenarios and the resultant ROC curves are shown in left and right panels of Figure 5.13 for the AR and MA examples, respectively. The performance gains associated with using the multitaper estimator are evident for this example.

5.2.2 Constructing a CFAR Test

Next we discuss how to implement the hypothesis test of (5.22) using a constant false alarm rate threshold, which requires knowledge of $p(\widehat{V}(\boldsymbol{x}); \mathcal{H}_0)$ and $p(\widetilde{V}(x); \mathcal{H}_0)$—the

[2]Choosing R large, however, may lead to greater bias and a weaker test.

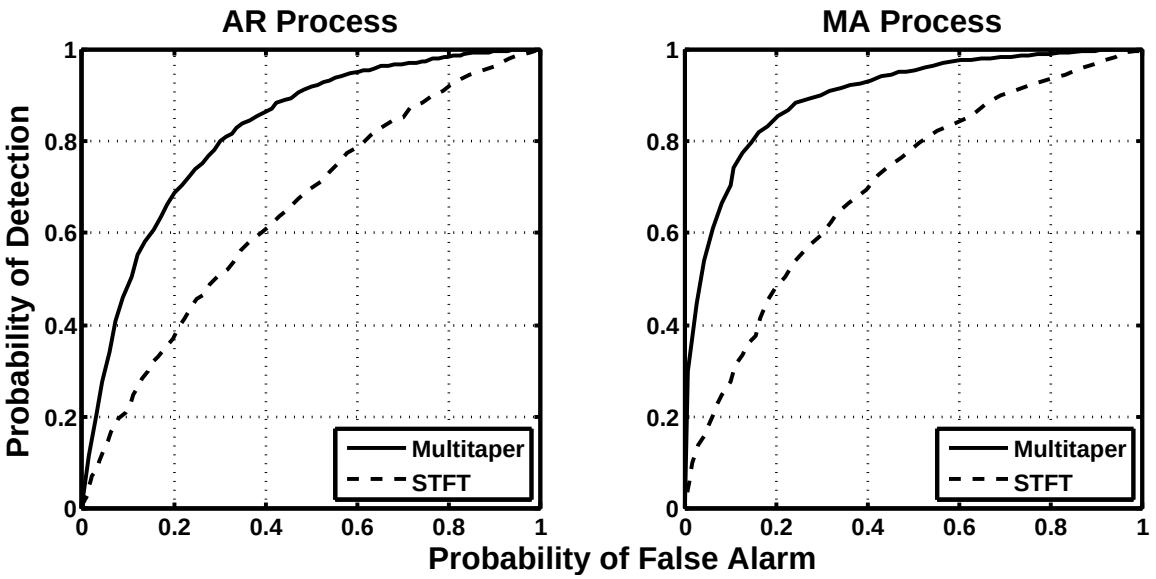


Figure 5.13: ROC curves summarizing test performance for time-varying AR (left) and MA (right) signals using STFT- and multitaper-based estimators. Signals were chosen to illustrate relative rather than absolute performance.

sampling distributions of (5.24) and (5.25), respectively, under $\mathcal{H}_0$. In this case, $S_{xx}^m[k]$ is independent of m and so we may estimate $S_{xx}[k]$ by:

$$\widehat{S_{xx}}[k] = \frac{1}{M}\sum_m \widehat{S_{xx}^m}[k] \quad \text{or} \quad \widetilde{S_{xx}}[k] = \frac{1}{M}\sum_m \widetilde{S_{xx}^m}[k]. \tag{5.28}$$

There are other choices for how to estimate $S_{xx}[k]$ under $\mathcal{H}_0$ (e.g., median of the short-time spectra), but the estimator in (5.28) is natural and leads to good performance.

If the sampling distributions of (5.24) and (5.25) for the class of WSS signals with power spectra given by (5.28) are known—indicating how to set a CFAR threshold γ—the null hypothesis is rejected when the test statistic exceeds γ. Therefore, we characterize $p_{\mathcal{H}_0}(\widehat{V}(\boldsymbol{x}))$ and $p_{\mathcal{H}_0}(\widetilde{V}(\boldsymbol{x}))$ next.

5.2.2.1 Asymptotic Analysis: White Noise Case

We begin by considering $p_{\mathcal{H}_0}(\widehat{V}(\boldsymbol{x}))$ for the special case when $x[n]$ is a white Gaussian noise process—not only do these calculations provide some intuition about the hypothesis test, but they shall also reappear in our analysis of the general case in Section 5.2.2.2. We begin by defining $l \triangleq L/2 - 1$ and rewriting $\widehat{V}(\boldsymbol{x})$ of (5.24) as:

$$\widehat{V}(\boldsymbol{x}) = \frac{1}{M^2 l}\sum_{k=0}^{l}\left[\sum_{m=0}^{M-1}(M-1)(\widehat{S_{xx}^m}[k])^2 - \sum_{r=0}^{M-1}\sum_{s=0,s\neq r}^{M-1}\widehat{S_{xx}^s}[k]\widehat{S_{xx}^r}[k]\right].$$

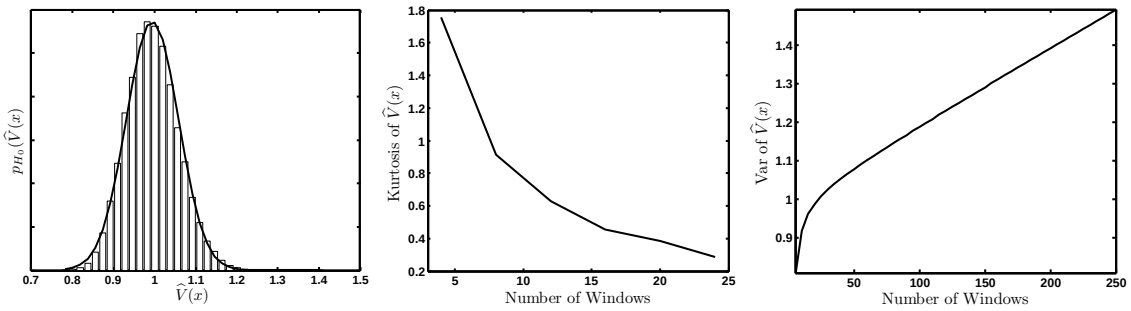


Figure 5.14: Understanding $p_{\mathcal{H}_0}(\widehat{V}(\boldsymbol{x}))$ when $x[n]$ is white Gaussian noise. Left: empirical (grey) and Gaussian (black) approximations of $p_{\mathcal{H}_0}(\widehat{V}(\boldsymbol{x}))$ with $M = 20$. The kurtosis of $\widehat{V}(\boldsymbol{x})$ under $\mathcal{H}_0$ decreases (middle) and its variance grows linearly with M (right)

.

It is well known that $\widehat{S^m_{xx}}[k] \sim \frac{1}{2}\chi^2_2$ for $0 < k \leq l$ and $\widehat{S^m_{xx}}[0] \sim \chi^2_1$, where χ^2_d is a chi-squared density with d degrees of freedom. Letting μ_{ki} denote the ith moment of the kth PSD bin, from (5.2.2.1) it follows that $\mathbb{E}\left(\widehat{V}(\boldsymbol{x})\right) = (M-1)/ML\sum_{k=0}^{l}(\mu_{k2} - \mu_{k1}^2)$ and:

$$\begin{aligned}\mathbb{E}\left(\widehat{V}(\boldsymbol{x})^2\right) = \frac{1}{(Ml)^2}\sum_{k=0}^{l}\Bigg(&\left(\frac{M-1}{M}\right)^2\left(M\mu_{k4} + M(M-1)\mu_{k2}^2\right)\\ &- 2\frac{(M-1)^2}{M}\left(2\mu_{k3}\mu_{k1} + (M-2)\mu_{k2}\mu_{k1}^2\right)\\ &+ \frac{1}{M^2}\left(2P^M_{M-2}\mu_{k2}^2 + 4P^M_{M-3}\mu_{k2}\mu_{k1}^2\ P^M_{M-4}\mu_{k1}^4\right)\Bigg)\\ &+ \frac{(M-1)^2}{(Ml)^2}\sum_k\sum_j\left(\mu_{k2}\mu_{j2} - 2\mu_{k2}\mu_{j1}^2 + \mu_{k1}^2\mu_{j1}^2\right),\end{aligned} \tag{5.29}$$

where $P^m_k \triangleq m!/(m-k)!$. A histogram of $p_{\mathcal{H}_0}(\widehat{V}(\boldsymbol{x}))$ obtained by simulation, overlaid with a Gaussian distribution fitted according to (5.29), is shown in the left panel of Figure 5.14 to illustrate this analysis. Even though the distribution of $\widehat{V}(\boldsymbol{x})$ is not Gaussian, the accuracy of the Normal approximation increases with the number of windows M. We have observed reasonable results when at least $15-20$ windows are used in calculating $\widehat{V}(\boldsymbol{x})$, as evidenced by the plot of empirical kurtosis as a function of the number of windows in the middle panel of Figure 5.14.

A key point is that we would ideally like to choose as large a window as possible, while still preserving the sensitivity of the periodogram to the presence of nonstationarity. Note that (5.29) implies that $\mathrm{Var}(\widehat{V}(\boldsymbol{x}))$ increases linearly in the number of analysis windows used, as confirmed by the plot in the right panel of Figure 5.14. Thus, L decreases as M increases and we conclude that using many short analysis windows may increase the overall mean-square error of our spectral estimates and thereby decrease the power of the hypothesis test, in agreement with our intuition.

5.2.2.2 Wold Representation: General Case

In the general case, when the spectrum (5.28) is not white or is estimated using overlapping windows, deriving $p_{\mathcal{H}_0}(\widehat{V}(\boldsymbol{x}))$ directly is either impossible or extremely tedious. Further, these calculations would be applicable only when M is large (see e.g., middle panel of Figure (5.14)). Instead, we proceed via simulation, by leveraging the Wold representation. Recall from Section 2.6.1, that the Wold representation of a stationary process $x[n]$ is given by:

$$x[n] = \sum_{m=0}^{\infty} h[m]\epsilon[n-m], \tag{5.30}$$

where $\epsilon[n]$ is an uncorrelated innovations sequence and $h[n]$ is the impulse response of a stable and minimum-phase filter. This allows us to express the PSD of $x[n]$ as follows:

$$S_{xx}[k] = |H[k]|^2 \, S_{\epsilon\epsilon}[k]. \tag{5.31}$$

Since $S_{\epsilon\epsilon}[k] = 1$, it is equivalent to estimate $S_{xx}[k]$ and $|H[k]|^2$. Thus, assuming that an observed signal $x[n]$ has been partitioned according to the same tiling of the time-frequency place that was used to define the test statistic in (5.24), we have that $|\widehat{H}[k]|^2 = \widehat{S_{xx}}[k]$ as defined by (5.28).

The key idea behind our simulation approach is to use (5.31) to generate realizations of $S_{xx}^m[k]$, under the null, by multiplying $|\hat{H}[k]|^2$ with different realizations of white noise spectra. In turn, this enables the computation of (5.24) for every realization so achieved. Specifically, let the ith realization of a white noise PSD be denoted by $\hat{S}_{\epsilon\epsilon}^{m,i}[k]$, then the ith realization of a WSS signal in the mth window is given by:

$$\widehat{S_{xx}^{m,i}}[k] = |\widehat{H}[k]|^2 \widehat{S_{\epsilon\epsilon}^{m,i}}[k]. \tag{5.32}$$

The ith instantiation of (5.24) under $\mathcal{H}_0$ is, therefore, given by:

$$\begin{aligned}\widehat{V}^i(\boldsymbol{x}) &= \frac{1}{Ml}\sum_{k=0}^{l}\sum_{m=0}^{M-1}\left(\widehat{S_{xx}^{m,i}}[k] - \frac{1}{M}\sum_{p=0}^{M-1}\widehat{S_{xx}^{p,i}}[k]\right)^2 \\ &= \frac{1}{Ml}\sum_{k=0}^{l}|\widehat{H}[k]|^4\sum_{m=0}^{M-1}\left(\widehat{S_{\epsilon\epsilon}^{m,i}}[k] - \frac{1}{M}\sum_{p=0}^{M-1}\widehat{S_{\epsilon\epsilon}^{p,i}}[k]\right)^2.\end{aligned} \tag{5.33}$$

Similarly, we may obtain replicates of $\widetilde{V}^i(\boldsymbol{x})$ and characterize $p_{\mathcal{H}_0}(\widetilde{V}(\boldsymbol{x}))$ by substituting multitaper estimates of the relevant PSDs into (5.32) and (5.33). In both cases, we may build up an empirical CDF of (5.24) or (5.25) and reject $\mathcal{H}_0$ if the observed statistic lies in the tail beyond a specified false alarm threshold γ.

It is crucial to observe that since $\hat{H}[k]$ is fixed, the only source of randomness in (5.33) is due to variance of the estimator for the PSD of white noise. Therefore, in practice, once a windowing scheme has been chosen, only the distribution of $\hat{S}_{\epsilon\epsilon}^{m,i}[k]$ has to be determined empirically (or via the Gaussian approximation of Section 5.2.2.1) for each k. Then $p_{\mathcal{H}_0}(\widehat{V}(\boldsymbol{x}))$ is readily obtained by using $\hat{H}[k]$ together with (5.33)—a powerful construct since all Monte-Carlo simulations may be done *offline*.

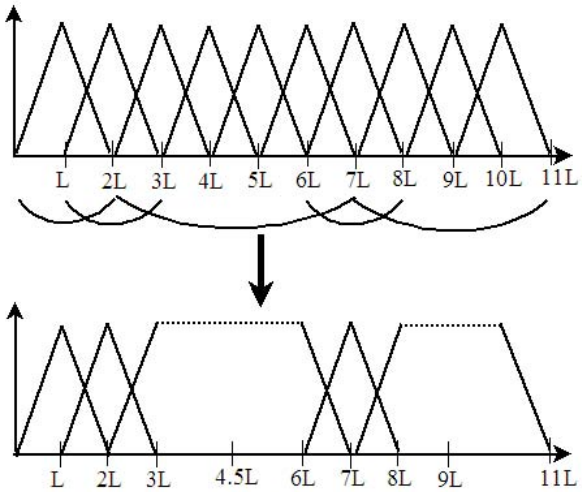


Figure 5.15: Example of how a fixed-resolution scheme using windows of length L (top) is modified to achieve an adaptive-resolution scheme by merging neighboring windows (bottom).

5.2.3 Enhancement Example

As an example of the applicability of the proposed testing framework to actual time series data, we illustrate its performance in the context of signal enhancement, with an eye toward the material presented next in Chapter 6. Audio noise reduction is generally achieved through the attenuation of spectral coefficients using *uniformly-sized* windows and local Fourier analysis. Instead, we propose a signal-adaptive enhancement scheme in which local Fourier analysis is employed using *variable-length* windows derived via the nonparametric hypothesis test of Section 5.2.1.

The intuition behind testing for stationarity as a means of signal adaptation is readily apparent: If the signal of interest $x[n]$ is a stationary random process, then it stands to reason that $x[n]$ should be analyzed and enhanced in its entirety. On the other hand, if $x[n]$ is nonstationary, then hypothesis testing serves to isolate segments over which the second-order statistics of $x[n]$ are inferred to be time-invariant.

5.2.3.1 Enhancement System

We use the test statistics of Section 5.24 to modulate a simple, sequential signal-adaptive enhancement scheme that can be implemented by "growing" a given window forward in time through successive attempts to merge it with its subsequent neighboring translates in a manner analogous to Algorithm 5.1 in Section 5.1.4.1. We defer a detailed description of this *online* scheme until Section 6.7.1 in Chapter 6. However, at a high level the scheme begins by tiling the signal using M uniformly spaced, possibly overlapping, windows of length L. Two adjacent windows are then merged if the null hypothesis of stationarity fails to be rejected for the short-time signal induced by the larger merged

window.[3] Whenever a proposed merge fails, the procedure resets and repeats, halting when the end of the data stream is reached (or, equivalently when the initial window is once again encountered, in a cyclic setting). In the end, we are left with a set of variable-length windows, an example is shown in Figure 5.15, which in turn induces a set of $M' \leq M$ variable-length short-time segments.

Assuming observations $y[n]$ degraded by additive white Gaussian noise, the resulting spectral slice $Y_{m'}[k]$ corresponding to the m'th window (with $1 \leq m' \leq M'$) is attenuated according to the standard Wiener suppression scheme:

$$\widehat{X}_{m'}[k] = \frac{\widehat{S_{xx}^{m'}}[k]}{\widehat{S_{xx}^{m'}}[k] + \sigma^2} Y_m[k], \tag{5.34}$$

where $\widehat{S_{xx}^{m'}}[k]$ is the periodogram estimator of the power spectrum based on the *clean* data in the m'th short-time segment. Here we use the periodogram of the clean (instead of the usual noisy) signal as the "oracle" estimate of the power spectral density because we are interested in illustrating the performance of our nonparametric test for stationarity, rather than evaluating competing spectral estimators. Waveform reconstruction from the modified variable-length short-time segments is achieved by overlap-add synthesis described in detail in Chapter 6.

5.2.3.2 Enhancement Example Results

We now apply the enhancement scheme to a stylized synthetic test signal akin to the example employed in [180] and as well as to a short segment of clarinet music. Both signals were corrupted by additive white Gaussian noise to yield a signal-to-noise (SNR) ratio of 5 dB. In both examples, the fixed-resolution scheme employed 200-sample triangular windows with 50% overlap, as in Figure 5.15. To decide whether two adjacent windows should be merged, each individual window is further subdivided into 4 parts and the test statistic of (5.24) is computed across all six windows tiling the two neighboring segments (the windows used to conduct the hypothesis are distinct from those used in the adaptive scheme). Using the techniques of Section 5.2.2.2 a 10% CFAR threshold was found and used in each decision of the adaptation scheme. Note that (5.33) allows us to approximate $p_{\mathcal{H}_0}(\widehat{V}(\boldsymbol{x}))$ even if windows are overlapping.

The adaptation scheme is applied to one instance of the synthetic signal and the resultant segmentation is shown in left panels of Figure 5.16. The hypothesis test aids in identifying stationary regions which, in turn, leads to improved enhancement performance (leading to an additional 0.5–2 dB SNR gain) relative to the fixed-resolution scheme as shown in right panels of Figure 5.16.

Adaptive segmentation of the clarinet recording, shown in Figure 5.17, accurately captures dominant signal features even in the presence of severe noise. Here the adaptive enhancement scheme provides an additional 1.2 dB SNR gain over the fixed-resolution approach. Moreover, informal listening tests have also indicated a significant reduction in musical noise—in agreement with results obtained for speech in [153].

[3] Note that even though we are here employing the nonparametric hypothesis test of Section 5.2.1, the parametric approach of 5.1.1.2 could be employed as well.

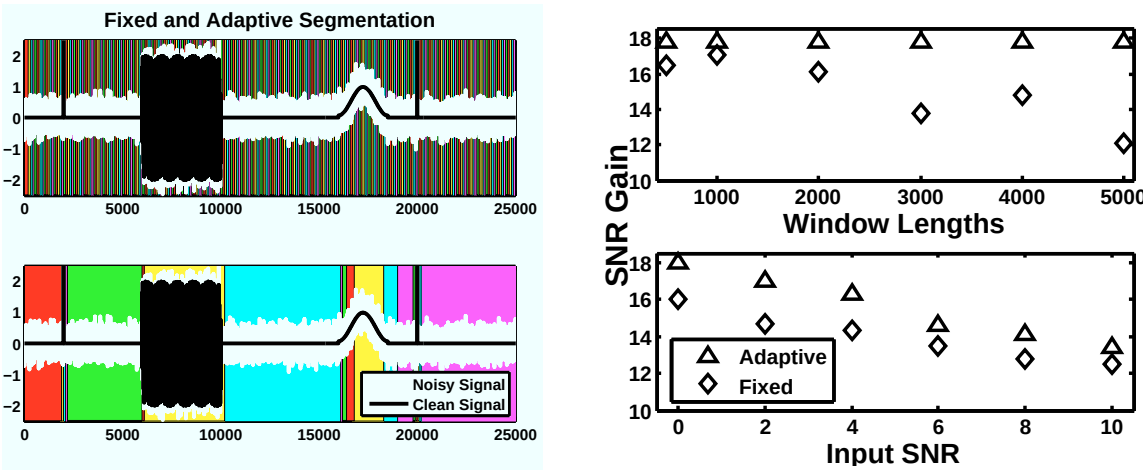


Figure 5.16: Left: Fixed- (top) and adaptive-resolution (bottom) segmentations of a synthetic test signal from [153], degraded with white Gaussian noise to yield an SNR of 5 dB. Boundaries of colored rectangles demarcate regions of stationarity. Right: SNR gain is shown as the lengths of the fixed-resolution windows are varied (top); and as a function of input SNR for fixed window lengths (bottom).

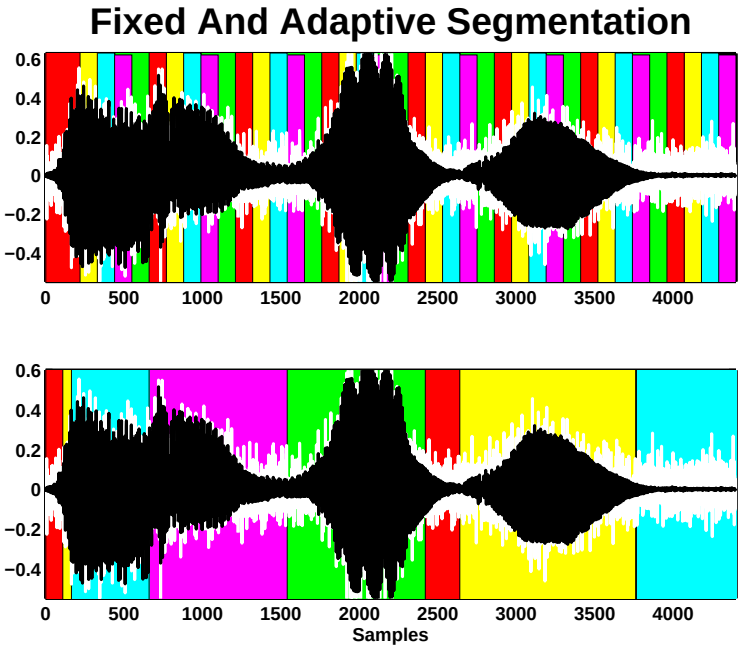


Figure 5.17: Fixed (top) and adaptive-resolution (bottom) segmentations of clarinet recording, degraded with white Gaussian noise to yield an SNR of 5 dB.

5.3 Summary

In this chapter we introduced parametric and nonparametric hypothesis tests for stationarity. In the parametric case, we developed a statistical framework based on TVAR models and applied it to the detection of nonstationarity in speech waveforms. This generalization of linear prediction was shown to yield efficient fitting procedures, as well as a corresponding generalized likelihood ratio test. Our study of GLRT detection performance yielded several practical consequences for speech analysis. Incorporating these conclusions, we presented two algorithms to identify changes in the vocal tract configuration in speech data at different time scales. At the segmental level we demonstrated the sensitivity of the GLRT to vocal tract variations corresponding to formant changes, and at the sub-segmental scale, we used it to identify both glottal openings and closures.

In the nonparametric setting, we proposed two test statistics based on the periodogram and multitaper estimators of the power spectral density and studied their sampling distribution under the null hypothesis using an efficient simulation scheme based on the Wold representation. We have illustrated our scheme in the context of enhancement for synthetic signals and explored its applicability for musical signals with promising results.

5.A Appendix: Asymptotic Behavior of the GLRT

In Section 5.1.2.1, we stated that the asymptotic distribution of the generalized likelihood ratio test (GLRT) statistic $T(\boldsymbol{x})$ was given by:

$$T(\boldsymbol{x}) \overset{N\to\infty}{\sim} \chi^2_{pq}(\lambda), \quad \begin{cases} \lambda = 0 & \text{under } \mathcal{H}_0, \\ \lambda > 0 & \text{under } \mathcal{H}_1. \end{cases}$$

Here, we derive (5.14) and (5.16)—the formulas for the non-centrality parameter λ.

Let $\boldsymbol{I}$ be the Fisher information matrix (FIM) with blocks corresponding to $\boldsymbol{\alpha}_{\mathrm{AR}}$, $\boldsymbol{\alpha}_{\mathrm{AR}}$, and σ^2 as follows:

$$\boldsymbol{I} \triangleq \begin{pmatrix} \boldsymbol{I}_{\boldsymbol{\alpha}_{\mathrm{AR}}\boldsymbol{\alpha}_{\mathrm{AR}}} & \boldsymbol{I}_{\boldsymbol{\alpha}_{\mathrm{AR}}\boldsymbol{\alpha}_{\mathrm{TV}}} & \boldsymbol{I}_{\boldsymbol{\alpha}_{\mathrm{AR}}\sigma^2} \\ \boldsymbol{I}_{\boldsymbol{\alpha}_{\mathrm{TV}}\boldsymbol{\alpha}_{\mathrm{AR}}} & \boldsymbol{I}_{\boldsymbol{\alpha}_{\mathrm{TV}}\boldsymbol{\alpha}_{\mathrm{TV}}} & \boldsymbol{I}_{\boldsymbol{\alpha}_{\mathrm{TV}}\sigma^2} \\ \boldsymbol{I}_{\sigma^2\boldsymbol{\alpha}_{\mathrm{AR}}} & \boldsymbol{I}_{\sigma^2\boldsymbol{\alpha}_{\mathrm{TV}}} & \boldsymbol{I}_{\sigma^2\sigma^2} \end{pmatrix}.$$

Below we will show that $\boldsymbol{I}_{\sigma^2\boldsymbol{\alpha}_{\mathrm{AR}}} = \boldsymbol{0}_{1\times p}$ and $\boldsymbol{I}_{\boldsymbol{\alpha}_{\sigma^2\mathrm{TV}}} = \boldsymbol{0}_{1\times pq}$. Consequently, in our composite hypothesis testing setting, the non-centrality parameter is given by:

$$\lambda \triangleq \boldsymbol{\alpha}_{\mathrm{TV}}^T \left(\boldsymbol{I}_{\boldsymbol{\alpha}_{\mathrm{TV}}\boldsymbol{\alpha}_{\mathrm{TV}}} - \boldsymbol{I}_{\boldsymbol{\alpha}_{\mathrm{TV}}\boldsymbol{\alpha}_{\mathrm{AR}}} \boldsymbol{I}^{-1}_{\boldsymbol{\alpha}_{\mathrm{AR}}\boldsymbol{\alpha}_{\mathrm{AR}}} \boldsymbol{I}_{\boldsymbol{\alpha}_{\mathrm{AR}}\boldsymbol{\alpha}_{\mathrm{TV}}} \right) \boldsymbol{\alpha}_{\mathrm{TV}} \tag{5.35}$$

where all the matrices are evaluated using the true parameter values under the null hypothesis $\mathcal{H}_0$. First, we will show that (5.35) is given by (5.14), and second that this expression is equivalent to (5.16).

The derivatives necssary for computing the Fisher information matrix $\boldsymbol{I}$ were

stated in Section 4.2.2. Thus, the required expectations under $\mathcal{H}_0$ are given by:

$$-\mathbb{E}\left(\frac{\ln p(\boldsymbol{x}_{N-p}\,|\,\boldsymbol{x}_p;\boldsymbol{\alpha},\sigma^2)}{\partial\alpha_{ij}\partial\alpha_{kl}}\right)=\frac{1}{\sigma^2}\sum_{n=p}^{N-1}f_j[n]f_l[n]\mathbb{E}(x[n-i]x[n-k])=\frac{r_{xx}[i-k]}{\sigma^2}\sum_{n=p}^{N-1}f_j[n]f_l[n]$$

$$-\mathbb{E}\left(\frac{\ln p(\boldsymbol{x}_{N-p}\,|\,\boldsymbol{x}_p;\boldsymbol{\alpha},\sigma^2)}{\partial\alpha_{ij}\partial\sigma^2}\right)=-\frac{1}{\sigma^4}\sum_{n=p}^{N-1}f_j[n]\mathbb{E}(e[n]x[n-i])=0$$

$$-\mathbb{E}\left(\frac{\ln p(\boldsymbol{x}_{N-p}\,|\,\boldsymbol{x}_p;\boldsymbol{\alpha},\sigma^2)}{\partial\sigma^2\partial\sigma^2}\right)=-\frac{N-p}{2\sigma^4}+\frac{1}{\sigma^6}\sum_{n=p}^{N-1}\mathbb{E}\big(e^2[n]\big)=\frac{N-p}{2\sigma^4},$$

where $\{r_{xx}[0],r_{xx}[1],\ldots,r_{xx}[p-1]\}$ is the autocorrelation sequence corresponding to $\boldsymbol{\alpha}_{\mathrm{AR}}$ (given, e.g., by the "step-down algorithm" [4]) and the second expectation evaluates to 0 by the orthogonality principle. For convenience define the vector $\boldsymbol{f}_j\triangleq\begin{pmatrix}f_j[p] & f_j[p+1] & \cdots & f_j[N-1]\end{pmatrix}^T\in\mathbb{R}^{(N-p)\times 1}$ and the matrix $\boldsymbol{F}\in\mathbb{R}^{(N-p)\times(q+1)}$ according to: $\boldsymbol{F}=\big(\boldsymbol{f}_0\,|\,\boldsymbol{F}_{\overline{0}}\big)=\begin{pmatrix}\boldsymbol{f}_0\,|\,\boldsymbol{f}_1 & \cdots & \boldsymbol{f}_q\end{pmatrix}$.

The Fisher information matrix, under $\mathcal{H}_0$, is therefore given by:

$$\boldsymbol{I}=\frac{1}{\sigma^2}\begin{pmatrix}\boldsymbol{f}_0^T\boldsymbol{f}_0\boldsymbol{R} & \boldsymbol{F}_{\overline{0}}^T\boldsymbol{f}_0\otimes\boldsymbol{R} & \boldsymbol{0}_{p\times 1}\\ \boldsymbol{f}_0^T\boldsymbol{F}_{\overline{0}}\otimes\boldsymbol{R} & \boldsymbol{F}_{\overline{0}}^T\boldsymbol{F}_{\overline{0}}\otimes\boldsymbol{R} & \boldsymbol{0}_{pq\times 1}\\ \boldsymbol{0}_{1\times p} & \boldsymbol{0}_{1\times pq} & \frac{N-p}{2\sigma^2}\end{pmatrix}=\frac{1}{\sigma^2}\begin{pmatrix}\boldsymbol{F}^T\boldsymbol{F}\otimes\boldsymbol{R} & \boldsymbol{0}_{p(q+1)\times 1}\\ \boldsymbol{0}_{1\times p(q+1)} & \frac{N-p}{2\sigma^2}\end{pmatrix},\tag{5.36}$$

where $\boldsymbol{R}$ is defined according to (5.15). The equations (5.35) and (5.36) together imply:

$$\lambda=\boldsymbol{\alpha}_{\mathrm{TV}}^T(\overline{\boldsymbol{F}^T\boldsymbol{F}\otimes\sigma^{-2}\boldsymbol{R}})\boldsymbol{\alpha}_{\mathrm{TV}},\tag{5.37}$$

where $\overline{\cdot}$ denotes the Schur complement with respect to the first $p\times p$ matrix block of its argument.

To show that (5.37) is equivalent to (5.16), first note the following relationship between Kronecker products and the Schur complement:

$$\begin{aligned}\overline{\boldsymbol{F}^T\boldsymbol{F}\otimes\sigma^{-2}\boldsymbol{R}}&=\left(\boldsymbol{F}_{\overline{0}}^T\boldsymbol{F}_{\overline{0}}\otimes\sigma^{-2}\boldsymbol{R}-\left(\boldsymbol{f}_0^T\boldsymbol{F}_{\overline{0}}\otimes\sigma^{-2}\boldsymbol{R}\right)^T\left(\boldsymbol{f}_0^T\boldsymbol{f}_0\sigma^{-2}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{f}_0^T\boldsymbol{F}_{\overline{0}}\otimes\sigma^{-2}\boldsymbol{R}\right)\right)\\&=\left(\boldsymbol{F}_{\overline{0}}^T\boldsymbol{F}_{\overline{0}}\otimes\sigma^{-2}\boldsymbol{R}-\left((\boldsymbol{f}_0^T\boldsymbol{F}_{\overline{0}})^T\otimes\sigma^{-2}\boldsymbol{R}^T\right)\left(\left(\boldsymbol{f}_0^T\boldsymbol{f}_0\right)^{-1}\boldsymbol{f}_0^T\boldsymbol{F}_{\overline{0}}\otimes\boldsymbol{I}\right)\right)\\&=\left(\boldsymbol{F}_{\overline{0}}^T\boldsymbol{F}_{\overline{0}}\otimes\sigma^{-2}\boldsymbol{R}-\left(\boldsymbol{F}_{\overline{0}}^T\boldsymbol{f}_0\left(\boldsymbol{f}_0^T\boldsymbol{f}_0\right)^{-1}\boldsymbol{f}_0^T\boldsymbol{F}_{\overline{0}}\otimes\sigma^{-2}\boldsymbol{R}\right)\right)\\&=\boldsymbol{F}_{\overline{0}}^T\left(\boldsymbol{I}-\boldsymbol{f}_0\left(\boldsymbol{f}_0^T\boldsymbol{f}_0\right)^{-1}\boldsymbol{f}_0^T\right)\boldsymbol{F}_{\overline{0}}\otimes\sigma^{-2}\boldsymbol{R}=\overline{\boldsymbol{F}^T\boldsymbol{F}}\otimes\sigma^{-2}\boldsymbol{R},\end{aligned}$$

where the second equality follows by the definition of a Kronecker product and the third equality follows from the mixed-product property of Kronecker products: $(\boldsymbol{B}_1\otimes\boldsymbol{C}_1)\cdot(\boldsymbol{B}_2\otimes\boldsymbol{C}_2)=\boldsymbol{B}_1\boldsymbol{B}_2\otimes\boldsymbol{C}_1\boldsymbol{C}_2$. Now, if we define the matrix $\boldsymbol{\Theta}\in\mathbb{R}^{p\times q}$ according to:

$$\boldsymbol{\Theta}=(\boldsymbol{\Theta}_0\,|\,\boldsymbol{\Theta}_{\overline{0}})\triangleq\begin{pmatrix}\boldsymbol{\alpha}_0\,|\,\boldsymbol{\alpha}_1 & \cdots & \boldsymbol{\alpha}_q\end{pmatrix},$$

and note that $\boldsymbol{\alpha}_{\mathrm{TV}}=\mathrm{vec}(\boldsymbol{\Theta}_{\overline{0}})$, we can write the noncentrality parameter according to:

$$\begin{aligned}\lambda&=\boldsymbol{\alpha}_{\mathrm{TV}}^T(\overline{\boldsymbol{F}^T\boldsymbol{F}\otimes\sigma^{-2}\boldsymbol{R}})\boldsymbol{\alpha}_{\mathrm{TV}}=\boldsymbol{\alpha}_{\mathrm{TV}}^T(\overline{\boldsymbol{F}^T\boldsymbol{F}}\otimes\sigma^{-2}\boldsymbol{R})\boldsymbol{\alpha}_{\mathrm{TV}}\\&=\mathrm{vec}(\boldsymbol{\Theta}_{\overline{0}})^T(\overline{\boldsymbol{F}^T\boldsymbol{F}}\otimes\sigma^{-2}\boldsymbol{R})\,\mathrm{vec}(\boldsymbol{\Theta}_{\overline{0}})=\mathrm{vec}(\boldsymbol{\Theta}_{\overline{0}})^T\,\mathrm{vec}\left(\sigma^{-2}\boldsymbol{R}\cdot\boldsymbol{\Theta}_{\overline{0}}\cdot\overline{\boldsymbol{F}^T\boldsymbol{F}}\right),\\&=\mathrm{tr}\left(\boldsymbol{\Theta}_{\overline{0}}^T\cdot\sigma^{-2}\boldsymbol{R}\cdot\boldsymbol{\Theta}_{\overline{0}}\cdot\overline{\boldsymbol{F}^T\boldsymbol{F}}\right)=\mathrm{tr}\left(\sigma^{-2}\boldsymbol{R}\cdot\boldsymbol{\Theta}_{\overline{0}}\overline{\boldsymbol{F}^T\boldsymbol{F}}\boldsymbol{\Theta}_{\overline{0}}^T\right),\end{aligned}\tag{5.38}$$

where we used the identities $\operatorname{vec}(\boldsymbol{BXC}) = (\boldsymbol{C}^T \otimes \boldsymbol{B}) \operatorname{vec}(\boldsymbol{X})$ and $\operatorname{tr}(\boldsymbol{B}^T\boldsymbol{C}) = \operatorname{vec}(\boldsymbol{B})^T \operatorname{vec}(\boldsymbol{C})$.

To write (5.38) as (5.16), we need to relate $\boldsymbol{\Theta}_{\overline{0}}\overline{\boldsymbol{F}^T\boldsymbol{F}}\boldsymbol{\Theta}_{\overline{0}}^T$ to the matrix $\boldsymbol{A}$ of TVAR coefficient trajectories previously defined in Section 5.1.2.1. Observe that we may write $\overline{\boldsymbol{F}^T\boldsymbol{F}}$ as $\boldsymbol{F}_{\overline{0}}^T \boldsymbol{P}_{\boldsymbol{f}_0}^{\perp} \boldsymbol{F}_{\overline{0}}$, and that when $\mathcal{H}_1$ is in force then $\boldsymbol{F}\boldsymbol{\Theta}^T = \boldsymbol{A}$, by construction. Thus,

$$\begin{aligned} \boldsymbol{\Theta}_{\overline{0}}\overline{\boldsymbol{F}^T\boldsymbol{F}}\boldsymbol{\Theta}_{\overline{0}}^T &= \boldsymbol{\Theta}_{\overline{0}}\boldsymbol{F}_{\overline{0}}^T\boldsymbol{P}_{\boldsymbol{f}_0}^{\perp}\boldsymbol{F}_{\overline{0}}\boldsymbol{\Theta}_{\overline{0}}^T = \boldsymbol{\Theta}_{\overline{0}}\boldsymbol{F}_{\overline{0}}^T\boldsymbol{P}_{\boldsymbol{f}_0}^{\perp}\boldsymbol{P}_{\boldsymbol{f}_0}^{\perp}\boldsymbol{F}_{\overline{0}}\boldsymbol{\Theta}_{\overline{0}}^T \\ &= \boldsymbol{\Theta}\boldsymbol{F}^T\boldsymbol{P}_{\boldsymbol{f}_0}^{\perp}\boldsymbol{P}_{\boldsymbol{f}_0}^{\perp}\boldsymbol{F}\boldsymbol{\Theta}^T = \boldsymbol{A}^T\boldsymbol{P}_{\boldsymbol{f}_0}^{\perp}\boldsymbol{P}_{\boldsymbol{f}_0}^{\perp}\boldsymbol{A} = (\boldsymbol{A}-\bar{\boldsymbol{A}})^T(\boldsymbol{A}-\bar{\boldsymbol{A}}) = \tilde{\boldsymbol{A}}^T\tilde{\boldsymbol{A}}. \end{aligned} \tag{5.39}$$

The development is now complete, since (5.39) together with (5.38) and the cyclic invariance of the trace implies

$$\lambda = \sigma^{-2} \operatorname{tr}(\tilde{\boldsymbol{A}}\boldsymbol{R}\tilde{\boldsymbol{A}}^T).$$

Chapter 6

Superposition Frames for Adaptive Time-Frequency Analysis and Fast Reconstruction

Short-time Fourier analysis is an essential tool in speech and signal processing, but its application implicitly assumes piecewise-stationarity of the underlying signal. On the other hand, signal-adaptive time-frequency (TF) representations can lead to improved algorithms as illustrated, in Section 5.2.3, by our use of nonparametric hypothesis testing for the construction of variable-resolution TF representations and their subsequent application to signal enhancement.

In this chapter, we address the question of how standard fixed-resolution short-time Fourier representations, such as the spectrogram, may be generalized in order to adapt to the TF structure of signals. To this end, we introduce a broad family of adaptive, linear time-frequency representations termed superposition frames, and show that they admit desirable fast overlap-add reconstruction properties akin to standard short-time Fourier techniques. This approach stands in contrast to many adaptive TF representations in the existing literature, which, while more flexible than standard fixed-resolution approaches, typically fail to provide for efficient reconstruction and often lack the regular structure necessary for precise frame-theoretic analysis. The main technical contributions come through the development of properties which ensure that our superposition construction provides for a numerically stable, invertible signal representation. The primary algorithmic contributions is the introduction and analysis of two signal adaptation schemes based on greedy selection and dynamic programming, respectively. We conclude by illustrating the framework in the context of speech enhancement in order to highlight potential applications of the approach.

6.1 Introduction

Overcomplete short-time Fourier methods are frequently used to analyze the time-varying spectral content of discrete-time waveforms $x[t]$ arising in a variety of signal processing applications (see e.g., Section 2.5 for a discussion of short-time analysis of speech waveforms.) Since the choice of localizing window function effectively controls the balance

between time and frequency resolution *a priori*, standard representations cannot modulate this trade-off to adapt to the local spectral content of $x[t]$. Over the past two decades, this shortcoming has motivated the development of various linear and nonlinear adaptive TF analysis methods [180–192], in applications ranging from biomedical engineering [193] to radar signal analysis [194] and speech processing [155].

Despite the recognized importance of overcomplete signal-adaptive time-frequency analysis, the above methods generally fail to admit *fast reconstruction* of the signal $x[t]$ from its TF representation, by which we mean any non-iterative method that avoids direct (pseudo-)inversion of the corresponding analysis operator. While approaches such as modulated lapped transforms for audio coding [195], wavelet packet decompositions via best basis [196, 197] and adaptive segmentation via dynamic programming [198–202] can lead to flexible tilings of the TF plane, the general goal of efficient reconstruction from signal-adaptive, *overcomplete time-frequency* representations remains an open problem. This issue is particularly relevant today, given the recent interest in oversampled, modulated filter banks [203–205].

Here, we introduce a broad family of adaptive, linear TF representations that admit a fast overlap-add reconstruction property akin to standard short-time Fourier techniques. We do so by adapting a discrete Gabor frame to an observed signal $x[t]$ via superpositions of neighboring translates of a single window function, to yield a representation termed a superposition frame. Related procedures include the multi-window constructions of [206], in which multiple systems are defined on the same TF lattice; and the multi-Gabor expansions of [207], in which multiple time lattices and windows are employed. However, neither of these schemes considers the use of subset selection to achieve a signal-adaptive system in the manner of the present chapter. More recent approaches [187–190] address subset selection from a Gabor frame or union of Gabor frames, but do not consider the structure of the corresponding canonical dual.

A very recent approach in this direction is the study of general nonstationary Gabor frames [208], and indeed our contribution can be viewed as one possible instantiation of this framework. However, as shown below, the additional structure induced by our superposition construction yields several important properties, including, among other results, a *preservation* of the lower frame bound of the original Gabor frame, a generalized constant overlap-add property that avoids the explicit computation of dual windows, and a means of generating new families of adaptive lapped frames.

The rest of the chapter is organized as follows. After reviewing the short-time Fourier transform and Gabor systems on $\mathbb{C}^L$ in Section 6.2, we introduce superposition windows in Section 6.3 and subsequently use them to construct superposition systems in Section 6.4. In Section 6.5 we prove that the resultant systems are in fact frames for $\mathbb{C}^L$ and study their frame-theoretic properties, and in Section 6.6 establish fast reconstruction via an analysis of the corresponding frame operator. In Section 6.7, we provide examples of signal-dependent adaptation algorithms and illustrate their application to superposition systems in the context of speech enhancement; results of informal listening tests are also reported. Detailed proofs of many results have been collected in Appendix 6.A.

6.2 Preliminaries

We first review some well-known properties of Gabor frames [205, 209, 210] and discuss their relationship to short-time Fourier analysis. We take as our setting the space $\mathbb{C}^L$, and interpret its members as discrete-time L-periodic signals $x \in \ell^2(\mathbb{Z}_L)$, with $\mathbb{Z}_L$ denoting the integers $\mathbb{Z}$ modulo L. The *short-time Fourier transform* (STFT) on $\mathbb{C}^L$ uses a well-concentrated window function in order to localize x in time prior to the analysis of its frequency content.

Definition 2 (Short-Time Fourier Transform)**.** *Fix a window $w \in \mathbb{C}^L$ and time-frequency lattice constants $a, b > 0$ that divide L, with a an integer, and define $M, N : Na = Mb = L$. Then for the mth frequency bin index and nth window shift, with $m \in \mathbb{Z}_M$ and $n \in \mathbb{Z}_N$, the Gabor or subsampled short-time Fourier transform $X[m, n]$ of $x \in \mathbb{C}^L$ is given by*

$$X[m,n] \triangleq \sum_{t=0}^{L-1} x[t]\overline{w[t-na]e^{2\pi imbt/L}}, \tag{6.1}$$

where $i = \sqrt{-1}$ and $\bar{\cdot}$ denotes complex conjugation. The expression of (6.1) can be viewed as a set of inner products of x with NM TF shifts of the chosen window w. To realize this correspondence, and to set notation, we introduce explicit translation and modulation operators as follows.

Definition 3 (Translation and Modulation Operators)**.** *Let the translation and modulation operators $\mathcal{T}$ and $\mathcal{M}$ be defined as maps from $\mathbb{C}^L$ to itself acting according to:*

$$\mathcal{T}_{na} w[t] \triangleq w[t-na], \quad \mathcal{M}_{mb} w[t] \triangleq w[t]\, e^{2\pi imbt/L}.$$

Through the action of these operators, TF shifts of the chosen window $w \in \mathbb{C}^L$ may be indexed as

$$\phi_{m,n}[t] \triangleq \mathcal{M}_{mb}\mathcal{T}_{na} w[t], \quad m \in \mathbb{Z}_M, n \in \mathbb{Z}_N, \tag{6.2}$$

and one speaks of a Gabor system $\mathcal{G}(w, a, b) = \{\phi_{m,n}\}$. In order to ensure a reconstruction property for any x from its subsampled short-time Fourier transform $X[m, n]$, the Gabor system $\mathcal{G}(w, a, b)$ must form a *frame* for $\mathbb{C}^L$ as follows.

Definition 4 (Gabor Systems and Frames)**.** *A denumerable set $\{\phi_{m,n}\}$ of vectors comprising TF shifts of a single window function $w \in \mathbb{C}^L$ is called a* Gabor system*, and is said to be a* Gabor frame *for $\mathbb{C}^L$ if there exist constants $0 < A \leq B < \infty$ termed* frame bounds *such that:*

$$\forall x \in \mathbb{C}^L,\ A\|x\|^2 \leq \textstyle\sum_{m,n} |\langle x, \phi_{m,n}\rangle|^2 \leq B\|x\|^2, \tag{6.3}$$

with inner product $\langle x, \phi_{m,n}\rangle \triangleq \sum_{t=0}^{L-1} x[t]\overline{\phi_{m,n}[t]} = X[m, n]$.

An upper frame bound B for (6.3) is guaranteed whenever the set $\{\phi_{m,n}\}$ is finite, and so the existence of a lower frame bound $A > 0$, for a finite Gabor system $\mathcal{G}(w, a, b)$, is equivalent to the requirement that its elements span $\mathbb{C}^L$. This occurs *if and only if* the *frame operator* is of full rank.

Definition 5 (Gabor Frame Operator)**.** *Let $\mathscr{G}(w,a,b) = \{\phi_{m,n}\}$ be a Gabor system on $\mathbb{C}^L$, and define the* frame operator $S : \mathbb{C}^L \to \mathbb{C}^L$ *through its action on x as $Sx = \sum_{m,n}\langle x, \phi_{m,n}\rangle \phi_{m,n}$. Then S is represented by the $L \times L$ symmetric and positive semi-definite matrix with entries*

$$S[t,t'] \triangleq \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} \mathcal{M}_{mb}\mathcal{T}_{na}w[t]\overline{\mathcal{M}_{mb}\mathcal{T}_{na}w[t']}. \tag{6.4}$$

Remark 1 (Strict Positive-Definiteness of Frame Operator)**.** *By Definition 5, the frame condition of* (6.3) *is equivalent to* strict *positive definiteness of S and hence a necessary condition is that $MN \geq L$ (i.e., $ab \leq L$). Moreover, the minimal and maximal eigenvalues of S yield optimal frame bounds, since* (6.3) *may be expressed as the requirement that $A\langle x,x\rangle \leq \langle Sx,x\rangle \leq B\langle x,x\rangle$, $\forall x \in \mathbb{C}^L$.*

The frame condition of (6.3) in turn implies the following reconstruction property:

$$\forall x \in \mathbb{C}^L, t \in \mathbb{Z}_L,\ x[t] = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} \langle x, \phi_{m,n}\rangle \widetilde{\phi}_{m,n}[t],$$

where the elements $\{\widetilde{\phi}_{m,n}\}$ comprise a (not necessarily unique) *dual frame*. However, to each frame may be associated a unique *canonical* dual, whose elements are given by the action of the frame operator inverse S^{-1} on each $\phi_{m,n}$. Moreover, in the Gabor setting, this canonical dual takes the form of another Gabor system $\mathscr{G}(\widetilde{w},a,b)$, with $\widetilde{w} \triangleq S^{-1}w$.

Any S can be written as a sum of outer products of each frame vector with itself; from (6.4) via the orthogonality relation

$$\sum_{m=0}^{M-1} e^{2\pi imb(t-t')/L} = \begin{cases} M & \text{when } M \text{ divides } t-t', \\ 0 & \text{otherwise,} \end{cases}$$

we obtain the so-called Walnut representation [211] of a Gabor frame operator S, which will be used repeatedly throughout.

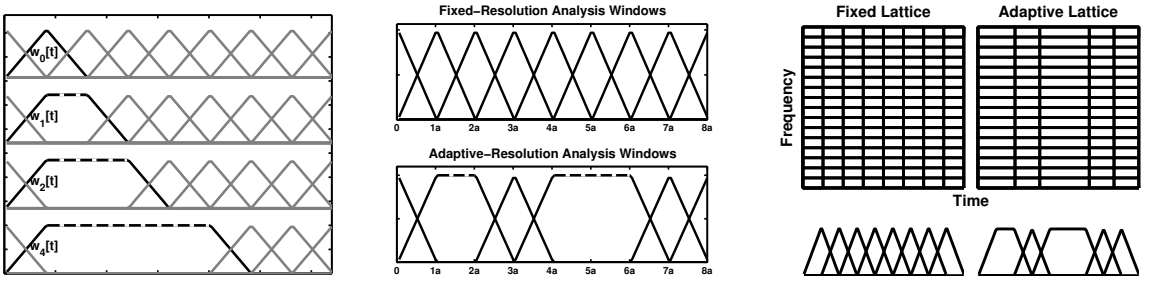
Definition 6 (Discrete Walnut Representation)**.** *Denote by $M\backslash(t-t')$ the condition that M divides $t-t'$, and by $\mathbb{I}_{M\backslash(t-t')}[t-t']$ the corresponding indicator function on $\mathbb{Z}_L$. Then the frame operator S of a finite Gabor system $\mathscr{G}(w,a,b)$ has banded structure, and satisfies the entrywise relation*

$$S[t,t'] = \mathbb{I}_{M\backslash(t-t')}[t-t'] \cdot M \sum_{n=0}^{N-1} \mathcal{T}_{na}w[t]\overline{\mathcal{T}_{na}w[t']}. \tag{6.5}$$

Remark 2 (Covering Condition)**.** *Note that if $\mathscr{G}(w,a,b)$ is a frame for $\mathbb{C}^L$, then* (6.5) *implies that the covering condition*

$$\sum_{n=0}^{N-1} |w[t-na]|^2 > 0, \forall\, t \in \mathbb{Z}_L \tag{6.6}$$

must be satisfied, since a necessary condition for positive definiteness of S is that its diagonal entries are positive.



(a) Superposition windows w_r, for $r = 0, 1, 2, 4$ merges of neighboring translates.

(b) Translates of windows from a Gabor system (top) and superposition system (bottom).

(c) Corresponding time-frequency lattices associated with the systems of Figure 6.1(b).

Figure 6.1: An example realization of a superposition system obtained via two and then three window merges.

Remark 3 (Window Length as Distinct from Support)**.** *The* support *of* $w \in \mathbb{C}^L$ *refers to the set of indices* t *for which* $w[t] \neq 0$*, with* $|\operatorname{supp}(w)|$ *its cardinality. Bearing in mind the summands of* (6.5) *and* (6.6)*, define the* length *of* w *by*

$$\operatorname{len}(w) \triangleq |\operatorname{supp}(w)| \tag{6.7}$$

if $\operatorname{supp}(w)$ *is contiguous, as is often the case in practice, and* $\min_{n\in\mathbb{Z}_L}(\max_{t,t'\in\mathbb{Z}_L}|t-t'+1|: \mathcal{T}_n(w[t]\overline{w[t']}) \neq 0)$ *otherwise.*

Remark 4 (Diagonal Frame Operator)**.** *It follows from* (6.5) *and* (6.7) *that* S *is diagonal if* $M \geq \operatorname{len}(w)$*, since* $\mathcal{T}_{na}w[t]\overline{\mathcal{T}_{na}w[t']} = 0$ *for all* $|t-t'| \geq \operatorname{len}(w)$*, including those for which* M *divides* $t-t'$*. In turn, this implies efficient computation of the dual frame* $\{\mathcal{M}_{mb}\mathcal{T}_{na}\widetilde{w}\}$*, with* $\widetilde{w} = S^{-1}w$ *obtained via element-wise division of* $w[t]$ *by* $S[t,t] = M\sum_{n=0}^{N-1}|w[t-na]|^2$*. In this case the condition of* (6.6) *is sufficient to guarantee the frame condition of* (6.3)*.*

We conclude by using the arguments of Remarks 3 and 4 to establish a result required for our subsequent development.

Lemma 1. *Fix any* $w \in \mathbb{C}^L$ *and* a, b *such that* $\mathscr{G}(w, a, b)$ *is a frame for* $\mathbb{C}^L$*, with* $M = L/b$*. Then for any integral* $M' \geq \operatorname{len}(w)$*, the Gabor system* $\mathscr{G}(w, a, L/M')$ *is also a frame for* $\mathbb{C}^L$*, with diagonal frame operator and maximal lower frame bound given by* $(M'/M)\min_{t\in\mathbb{Z}_L} S[t,t]$*.*

Proof. We must show that if the frame operator S of a Gabor system $\mathscr{G}(w, a, b)$ on $\mathbb{C}^L$ is full rank, then so is the frame operator S' of any system $\mathscr{G}(w, a, L/M')$. For any $M' \geq \operatorname{len}(w)$, the argument of Remark 4 implies that S' is diagonal, with eigenvalues $S'[t,t] = M'\sum_{n=0}^{N-1}|w[t-na]|^2$; the Walnut representation of (6.5) further implies that $S'[t,t] = (M'/M)S[t,t]$, for $M = L/b$. As $\mathscr{G}(w, a, b)$ is a frame for $\mathbb{C}^L$, it follows that $S[t,t] > 0$. Hence $S'[t,t] > 0$ for all $t \in \mathbb{Z}_L$, and thus S' is of full rank. □

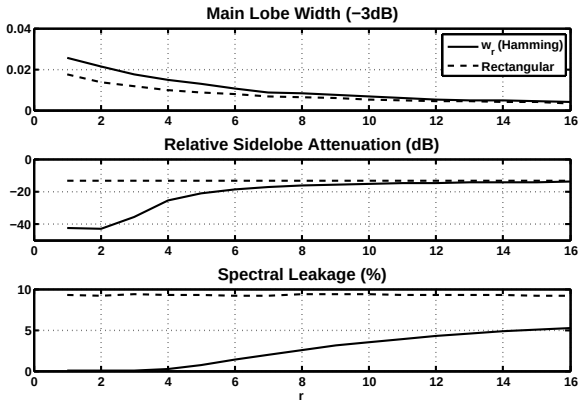


Figure 6.2: Frequency characteristics of superposition windows w_r derived from Hamming windows, with $a = \text{len}(w_0)/4$ (75% overlap), shown relative to those of rectangular windows of length $\text{len}(w_r)$.

6.3 Superposition Windows

Having outlined the connections between Gabor systems and the short-time Fourier transform, we now introduce the central ingredient of our signal-adaptive TF analysis framework—the *superposition window* construction, illustrated in Figure 6.1(a).

Definition 7 (Superposition Window)**.** *Fix a real, nonnegative window w on $\mathbb{C}^L$ and an integer $a = L/N$, along with some $r \in \mathbb{Z}_N$. We then define the* superposition window w_r *to be a linear sum of $r+1$ adjacent translates of $w[t]$ as follows:*

$$w_r[t] \triangleq \sum_{n=0}^{r} \mathcal{T}_{na} w[t], \quad r \in \mathbb{Z}_N. \tag{6.8}$$

Remark 5 (Fourier Transform Support)**.** *Let $\widehat{w}$ denote the (discrete) Fourier transform of $w \in \mathbb{C}^L$. Linearity of* (6.8) *implies that the support of $\widehat{w_r}$ is contained within that of $\widehat{w}$, as* $\text{supp}(\widehat{w_r}) = \text{supp}(\sum_{n=0}^{r} e^{-2\pi i n a(\cdot)/L} \widehat{w}) \subseteq \text{supp}(\widehat{w})$.

Remark 6 (Fourier Transform Decay)**.** *As r increases, it is clear that w_r can become more like a rectangular window (see, e.g., Figure 6.1(a)); this effect is illustrated in Figure 6.2 for the case of Hamming superposition windows. Consequently, the main lobe width of $\widehat{w_r}$ shrinks, leading to improved frequency resolution relative to $\widehat{w_0}$; this main lobe resolution, however, comes at the expense of decreasing sidelobe attenuation. Spectral leakage—the ratio of sidelobe power to total power—remains superior, as does overall spectral decay for small r.*

Our subsequent construction of superposition frames employs sets of modulated superposition windows, and to this end we establish the following two energy "conservation" properties, proved in Appendix 6.A.

Lemma 2 (Localized Parseval Property)**.** *Fix any* $w \in \mathbb{C}^L$ *and an integer* $M = L/b \geq \text{len}(w)$*. Then*

$$\forall x \in \mathbb{C}^L, \ \sum_{m=0}^{M-1} |\langle x, \mathcal{M}_{mb} w \rangle|^2 = M \sum_{t=0}^{L-1} |x[t]|^2 |w[t]|^2. \tag{6.9}$$

Lemma 3 (Superadditivity of Superposition Energy)**.** *Let real, nonnegative superposition windows* w_p *and* w_q *be derived from a Gabor system* $\mathscr{G}(w, a, b)$ *on* $\mathbb{C}^L$*, and merge them to obtain a new superposition window* $w_p + w_q{}' = w_p + \mathcal{T}_{(p+1)a} w_q$*. Then,* if *and* only if $M = L/b \geq \text{len}(w_p + w_q{}')$*, the following holds for every* $M_0 = L/b_0 \in \{\max(\text{len}(w_p), \text{len}(w_q)), \ldots, M\}$*:*

$$\sum_{m=0}^{M_0-1} |\langle x, \mathcal{M}_{mb_0} w_p \rangle|^2 + \left|\langle x, \mathcal{M}_{mb_0} w_q{}' \rangle\right|^2 \leq \sum_{m=0}^{M-1} \left|\langle x, \mathcal{M}_{mb}(w_p + w_q{}') \rangle\right|^2, \quad \forall x \in \mathbb{C}^L. \tag{6.10}$$

This superadditivity property, which is invariant to translation of $w_p + w_q{}'$, will be used in Section 6.5 to show that adapting a Gabor frame through the superposition construction *preserves* the original Gabor lower frame bound—an important consideration for numerical stability.

6.4 Construction of Superposition Systems

We now describe how to employ the superposition windows of Section 6.3 above to create a signal-adaptive analysis framework. Let $\mathscr{G}(w, a, b)$ represent a Gabor system, which induces a short-time Fourier transform on $\mathbb{C}^L$ according to Definition 2. Beginning with $\mathscr{G}(w, a, b)$, we then form a signal-dependent, variable-resolution STFT by using the superposition sum of (6.8) to adaptively merge neighboring translates from the set $\{\mathcal{T}_{na} w, n \in \mathbb{Z}_N\}$. Later we will demonstrate how this signal-adaptive analysis can be coupled with a variety of different algorithms; we begin, however, by studying the general set of *superposition systems* independently of any algorithmic construction. To this end, we introduce the notion of *ordered partition functions* as a means of indexing arbitrary sets of superposition windows, and then extend these to yield a full time-frequency analysis.

6.4.1 Ordered Partition Functions

Observe that exactly 2^{N-1} distinct sets of variable-length superposition windows may be derived by merging window translates from a given Gabor system $\mathscr{G}(w, a, \cdot)$. As a means of indexing these sets, the following "stick-breaking" analogy is helpful. Consider a "stick" composed of N ordered, unit-length pieces, representing elements of the set $\{\mathcal{T}_{na} w, n \in \mathbb{Z}_N\}$. Merging adjacent windows in this set can be thought of as fusing neighboring pieces of the stick. Each stick partition thus induces an ordered partition of the set $\{1, 2, \ldots, N\}$, with each piece uniquely identified by an initial index and length, and we may formalize this analogy as follows.

Definition 8 (Ordered Partition Functions)**.** *We call any* $\widetilde{I} : \mathbb{Z}_N \times \mathbb{Z}_N \to \{0, 1\}$ *an* ordered partition function *if it is not identically zero, and satisfies the following three properties:*

1. *Each piece of the stick is distinct:*

$$\widetilde{I}[n,r]=1 \;\Rightarrow\; \widetilde{I}[n,r']=0 \;\; \forall\, r'\neq r, r'\in\mathbb{Z}_N.$$

2. *The length of each piece is denoted by* $r+1$*:*

$$\widetilde{I}[n,r]=1 \;\Rightarrow\; \widetilde{I}[n',r']=0 \text{ on } \{n+1,\ldots,n+r\}\times\mathbb{Z}_N.$$

3. *All pieces of the stick are accounted for:*

$$\widetilde{I}[n,r]=1 \;\Rightarrow\; \widetilde{I}[n+r,r']=1 \text{ for exactly one } r'\in\mathbb{Z}_N.$$

Definition 8 clearly implies that the "length" of the stick remains unchanged:

$$\sum_{n=0}^{N-1}\sum_{r=0}^{N-1}\widetilde{I}[n,r]\,(r+1)=N; \tag{6.11}$$

moreover, each ordered partition function $\widetilde{I}$ can be associated with a set of translated superposition windows, which includes $T_{na}w_r$ whenever $\widetilde{I}[n,r]=1$. The following examples of ordered partition functions are illustrated in Figure 6.1(b).

Example 1 (N-Part Partition)**.** *The ordered partition function associated to the top panel of Figure 6.1(b) is*

$$\widetilde{I}[\cdot,r]\triangleq\begin{cases}1 & \text{if } r=0,\\ 0 & \text{otherwise.}\end{cases}$$

This clearly recovers the window translates of any Gabor system $\mathscr{G}(w,a,\cdot)$*. Note that in accordance with* (6.11)*, we have that* $\sum_{n=0}^{N-1}\sum_{r=0}^{N-1}\widetilde{I}[n,r]\,(r+1)=\sum_{n=0}^{N-1}\widetilde{I}[n,0]=N=8$*.*

Example 2 ($(N-3)$-Part Partition)**.** *The ordered partition function associated to the bottom panel of Figure 6.1(b) is*

$$\widetilde{I}[n,r]=\begin{cases}\widetilde{I}[2,0]=\widetilde{I}[6,0]=\widetilde{I}[7,0]=1 & \text{not merged,}\\ \widetilde{I}[0,1]=\widetilde{I}[3,2]=1 & \text{merged.}\end{cases}$$

Note again that in accordance with (6.11)*, we have that* $\sum_{n=0}^{N-1}\sum_{r=0}^{N-1}\widetilde{I}[n,r](r+1)=3+2+3=8$*.*

6.4.2 Superposition Systems

We now employ the above construction to arrive at a variable-resolution TF analysis via superposition windows. To this end, let the set $\mathscr{F}$ be defined as a function of any Gabor system $\mathscr{G}(w,a,b)$ on $\mathbb{C}^L$ as follows:

$$\mathscr{F}\triangleq\bigcup_{r\in\mathbb{Z}_N}\mathscr{G}(w_r,a,b_L),$$

where the frequency lattice $b_L\mathbb{Z}$ encompasses all possible Gabor systems on $\mathbb{C}^L$ for a fixed choice of integral M:

$$b_L\mathbb{Z};\ M_L \triangleq \mathrm{lcm}(\{1, 2\ldots, \max(L, M)\}),\ b_L \triangleq L/M_L.$$

Elements of $\mathscr{F}$ may then be defined in analogy to (6.2) as

$$\phi_{m,n,r} \triangleq \mathcal{M}_{mb_L}\mathcal{T}_{na}w_r = \mathcal{M}_{mb_L}\big(\textstyle\sum_{n'=0}^{r}\mathcal{T}_{(n'+n)a}w\big),$$

and in turn give rise to *superposition systems*, defined as appropriately chosen subsets of $\mathscr{F}$.

Definition 9 (Superposition Systems and Admissibility)**.** *Fix an ordered partition function $\widetilde{I}[n,r]$ and a function $M[n,r] : \mathbb{Z}_N \times \mathbb{Z}_N \to \mathbb{Z}_{M_L}$. We call any $I[m,n,r] : \mathbb{Z}_{M_L} \times \mathbb{Z}_N \times \mathbb{Z}_N \to \{0,1\}$ an* admissible selection function *on $\mathscr{F} = \cup_r \mathscr{G}(w_r, a, b_L)$ if it satisfies the following two properties:*

$$\begin{aligned} &I[0,n,r] = \widetilde{I}[n,r] \quad \forall\,(n,r) \in \mathbb{Z}_N \times \mathbb{Z}_N, \\ &I[0,n,r] = 1 \Rightarrow I[\tfrac{M_L}{M[n,r]}m, n, r] = 1,\ \forall\, m \in \mathbb{Z}_{M[n,r]}. \end{aligned} \tag{6.12}$$

Furthermore, we call the induced set of elements a superposition system *$\mathscr{F}(I)$:*

$$\phi_{m,n,r} \in \mathscr{F}(I) \Leftrightarrow I[m,n,r] = 1.$$

It follows from (6.12) that the first $M[n,r]$ modulates of each selected superposition window are included in $\mathscr{F}(I)$, and thus we later suppress the dependence of I on frequency bin index m when possible, by abbreviating $I[\cdot,n,r]$ as $I[n,r]$.

6.5 Superposition Frames: Main Results

Starting from a Gabor system $\mathscr{G}(w,a,b)$, we see that any superposition system $\mathscr{F}(I) \subset \cup_r \mathscr{G}(w_r, a, b_L)$ effectively yields a "variable-resolution" subsampled short-time Fourier transform, defined for all $m \in \mathbb{Z}_{M_L}$ and $n, r \in \mathbb{Z}_N$ as

$$X[m,n,r] \triangleq \begin{cases} \langle x, \phi_{m,n,r}\rangle & \text{if } \phi_{m,n,r} \in \mathscr{F}(I), \\ 0 & \text{otherwise.} \end{cases} \tag{6.13}$$

Consequently, we now establish conditions under which superposition systems $\mathscr{F}(I)$ form frames for $\mathbb{C}^L$, in analogy to the relation between a Gabor frame and its corresponding fixed-resolution short-time Fourier transform.

6.5.1 General Case: Sufficiency

To begin our analysis, consider the case of an admissible selection function $I[m,n,r]$ for which $M[n,r] = M_{\mathrm{g}}$ for all $n, r \in \mathbb{Z}_N$, corresponding to the notion of a *global* frequency lattice of arbitrary resolution: $b_{\mathrm{g}}\mathbb{Z}$ with $b_{\mathrm{g}} = L/M_{\mathrm{g}}$. Our first result, proved in Appendix 6.A, ensures that the induced superposition system $\mathscr{F}(I)$ is a frame for $\mathbb{C}^L$ if the following test condition holds.

Theorem 1 (Sufficiency Condition, Superposition Frames)**.** *Fix a Gabor system $\mathscr{G}(w,a,\cdot)$ on $\mathbb{C}^L$, with $N=L/a$, w real and nonnegative, and define for $s,t\in\mathbb{Z}_L$, $n,r\in\mathbb{Z}_N$, the term*

$$\beta_{nr}(s,t) \triangleq w_r[t-na]w_r[t-na-s].$$

Let $I[n,r]$ be any *admissible selection function for which $M[n,r]=M_g$ for some $M_g\in\{1,2,\ldots,L\}$. Then, for index term $k\in\{\lceil (t-(L-1))/M_g\rceil,\ldots,\lfloor t/M_g\rfloor\}$, the condition*

$$\forall t\in\mathbb{Z}_L,\ \sum_{n=0}^{N-1}\sum_{r=0}^{N-1} I[n,r]\Big(\beta_{nr}(0,t)-\sum_{k\neq 0}\beta_{nr}(kM_g,t)\Big)>0 \tag{6.14}$$

is sufficient to guarantee that the superposition system $\mathscr{F}(I)\subset\cup_r\mathscr{G}(w_r,a,b_L)$ is a frame for $\mathbb{C}^L$.

Satisfying the criterion of (6.14) implies that the underlying frame operator is *strictly diagonally dominant*—a sufficient condition for strict positive definiteness. This is a popular criterion in the literature (see, e.g., [209, Corollary 6], [210, Theorem 8.4.4]) and, as can be seen from (6.14), takes a particularly simple form in the superposition setting.

6.5.2 Superposition Frames and Frame Bounds

In Theorem 1 above, we considered a general class of superposition frames associated with an arbitrary frequency lattice $b_{\mathrm{g}}\mathbb{Z}$. In Theorems 2 and 3 below, we study two distinct classes of superposition systems using non-uniform (local) and uniform (global) modulation structures defined as follows.

Definition 10 (Admissible Selection Functions I^{l} and I^{g})**.** *Fix a Gabor system $\mathscr{G}(w,a,L/M)$, associate to it any ordered partition function $\widetilde{I}[n,r]$, and define*

$$M_r \triangleq \max\left(\mathrm{len}(w_r),M\right); \quad b_r\triangleq L/M_r, \tag{6.15}$$

$$M_g \triangleq \max_{r\,:\,\widetilde{I}[\cdot,r]=1} M_r; \qquad b_g\triangleq L/M_g. \tag{6.16}$$

These quantities induce, via $M[n,r]=M[\cdot,r]=M_r$ or $M[n,r]=M_g$ constant, respective classes of local *and* global *admissible selection functions $I^{\mathrm{l}}[m,n,r]$ and $I^{g}[m,n,r]$. Note that when no (N-part partition) or all (1-part partition) windows have been merged, the constants of* (6.15) *and* (6.16) *are equal.*

We now show that superposition systems $\mathscr{F}(I^{\mathrm{g}})$ and $\mathscr{F}(I^{\mathrm{l}})$ are frames for $\mathbb{C}^L$. Later, we will verify that such frames admit *diagonal* frame operators. This special structure leads not only to fast reconstruction algorithms, but also to the preservation of lower frame bounds as stated in the theorem below and proved in Appendix 6.A.

Theorem 2 (Local and Global Superposition Frames)**.** *Let $\mathscr{G}(w,a,b)$ be a Gabor frame for $\mathbb{C}^L$, with w real and nonnegative. Then for any choice of admissible selection functions I^{l} and I^{g}, the local and global superposition systems $\mathscr{F}(I^{\mathrm{l}})$ and $\mathscr{F}(I^{g})$ are also frames for $\mathbb{C}^L$.*

Thus we see that for every Gabor system on $\mathbb{C}^L$ and any associated ordered partition function, setting $M[n,r]$ in accordance with M_r or M_{g} will yield local or global superposition frames. Moreover, as we detail later, the iterative arguments employed above suggest precise algorithmic constructions.

We now proceed to establish the important property that, for all local and global $I[m,n,r]$ of Definition 10, *superposition frames preserve lower frame bounds*, thus ensuring numerical stability of the resultant representation. The following result, proved in Appendix 6.A, also formulates the corresponding minimax-optimal superposition frame bounds.

Theorem 3 (Superposition Frame Bound Properties)**.** *Let $\mathscr{G}(w,a,b)$ be a frame for $\mathbb{C}^L$, with associated maximal lower frame bound $A>0$. Then for any admissible I^l and I^g:*

1. *The quantity A remains a valid lower frame bound for both $\mathscr{F}(I^l)$ and $\mathscr{F}(I^g)$.*
2. *The* minimum *maximal lower superposition frame bound over all admissible I^l and I^g is*
$$A_{\mathrm{opt}} = \frac{L}{b^{\max}} \cdot \min_{t\in\mathbb{Z}_L} \left(\sum_{n=0}^{N-1} |\mathcal{T}_{na} w[t]|^2 \right),$$
with $b^{max} = \min(b, L/\operatorname{len}(w))$. It is attained in the absence of merging: $\mathscr{F}(I^l) = \mathscr{F}(I^g) = \mathscr{G}(w,a,b^{\max})$.
3. *The* maximum *minimal upper superposition frame bound over all admissible I^l and I^g is*
$$B_{\mathrm{opt}} = \frac{L}{b^{\min}} \cdot \max_{t\in\mathbb{Z}_L} \left(\Big| \sum_{n=0}^{N-1} \mathcal{T}_{na} w[t] \Big|^2 \right),$$
with $b^{\min} = \min(b,1)$. It is attained when all translates of w have been merged: $\mathscr{F}(I^l) = \mathscr{F}(I^g) = \mathscr{G}(w_{N-1}, L, b^{\min})$, with $w_{N-1} = \sum_{n=0}^{N-1} \mathcal{T}_{na} w[t]$.

6.6 Fast Reconstruction via Superposition Frames

We now show how the special structure of our superposition construction gives rise to a number of efficient reconstruction procedures. For any signal of interest $x \in \mathbb{C}^L$, the superposition frame analysis coefficients $X[m,n,r] = I[m,n,r]\langle x, \phi_{m,n,r}\rangle$ can be computed via fast Fourier transform (FFT) once an admissible selection function has been specified. Superposition frames also enable *fast* (FFT-based) *reconstruction* from the corresponding analysis coefficients, in contrast to the general case of $\mathcal{O}(L^3)$ complexity for frame-based reconstruction via inversion of the frame operator.

We first provide a fast constant-overlap-add reconstruction method, which obviates the need for canonical dual frames. We next show that reconstruction via the canonical dual can also proceed by way of a pointwise modification of each superposition window $\phi_{0,n,r}$, followed by the application of FFTs, as in the case of general nonstationary Gabor frames [208]. Third, we show that in settings reminiscent of lapped orthogonal transforms, calculation of canonical dual windows is possible *independently* of any I^{g}—in contrast to

the typical signal-adaptive setting, where the structure of the frame operator is a function of the instantiated signal adaptation. Last, we compare the computational complexity of these procedures.

6.6.1 Reconstruction via the Constant Overlap-Add Method

The classical "overlap-add" approach to signal reconstruction from short-time Fourier coefficients proceeds as follows [212]. Recall the covering condition of (6.6) which is necessary for a Gabor system $\mathscr{G}(w,a,b)$ to form a frame for $\mathbb{C}^L$, and also sufficient if $M = L/b \geq \text{len}(w)$. Clearly, this covering condition holds if translates $\{\mathcal{T}_{na}w : n \in \mathbb{Z}_N\}$ form a partition of unity on $\mathbb{C}^L$ (see, e.g., the top panel of Figure 6.1(b)), and to this end we obtain the following definition, long popular in the signal processing literature.

Definition 11 (Constant Overlap-Add Window Constraint)**.** *Fix a Gabor system $\mathscr{G}(w,a,\cdot)$ on $\mathbb{C}^L$. Then, noting the discrete Fourier transform evaluation $\widehat{w}[0] = \sum_{t=0}^{L-1} w[t]$, the window w is said to satisfy the* constant overlap-add *constraint if*

$$\forall\, t \in \mathbb{Z}_L,\ \sum_{n=0}^{N-1} w[t-na] = \frac{\widehat{w}[0]}{a}. \tag{6.17}$$

To clarify the role of this overlap-add constraint in fast reconstruction, consider a Gabor frame $\mathscr{G}(w,a,b)$ on $\mathbb{C}^L$ for which w satisfies (6.17), with $M = L/b$ chosen such that $M \geq \text{len}(w)$. The associated short-time analysis coefficients $\{X[m,n] : m \in \mathbb{Z}_M, n \in \mathbb{Z}_N\}$ are obtained as inner products of any $x \in \mathbb{C}^L$ according to (6.1), and it is easy to show that

$$x[t] = \frac{a}{\widehat{w}[0]} \sum_{n=0}^{N-1} \left(\frac{1}{M} \sum_{m=0}^{M-1} X[m,n] e^{2\pi i m b t/L} \right). \tag{6.18}$$

The constraint thus admits a reconstruction procedure based on the overlapping additions of a sequence of discrete Fourier transforms on $\mathbb{C}^M$.

Remark 7 (Superposition Windows Preserve Overlap-Add)**.** *By their* linear *construction, superposition windows w_r preserve the constant overlap-add constraint of Definition 11 for any Gabor system $\mathscr{G}(w_r,a,\cdot)$. To see this, note that the discrete Poisson summation formula on $\mathbb{C}^L$, for $N = L/a$, is given by $\sum_{n=0}^{N-1} \mathbb{I}_0[t-na] = a^{-1} \sum_{k=0}^{a-1} e^{2\pi i t k N/L}$. Applying this expression to $\{\mathcal{T}_{na}w : n \in \mathbb{Z}_N\}$ yields the relation*

$$\sum_{n=0}^{N-1} w[t-na] = \frac{1}{a} \sum_{k=0}^{a-1} e^{2\pi i t k N/L} \widehat{w}[kN],$$

and it follows that the constraint of (6.17) *holds (for a given time lattice constant a) if the Fourier transform $\widehat{w}$ satisfies*

$$\widehat{w}[kN] = 0,\ \forall\, k \in \{1,\ldots,a-1\}.$$

Since $\text{supp}(\widehat{w_r}) \subseteq \text{supp}(\widehat{w})$*, in accordance with the argument of Remark 5, it follows that if* (6.17) *holds for a given $\mathscr{G}(w,a,\cdot)$, then it will also hold for any $\mathscr{G}(w_r,a,\cdot)$, for $r \in \mathbb{Z}_N$.*

The popularity of the overlap-add constraint of Definition 11 is due in large part to its simplicity, coupled with the efficiency of evaluating (6.18). An important property of our superposition construction is that it preserves this constraint not only for Gabor frames $\mathscr{G}(w_r, a, b)$, but also for all induced superposition frames $\mathscr{F}(I^{\mathrm{g}})$ and $\mathscr{F}(I^l)$ as shown in Appendix 6.A.

Theorem 4 (Superposition Frames Preserve Overlap-Add)**.** *Consider a Gabor frame $\mathscr{G}(w, a, b)$ on $\mathbb{C}^L$ satisfying the constant overlap-add constraint of* (6.17)*. The following statements hold for any I^g, and also for any I^l, with M_g, b_g replaced by M_r, b_r.*

1. *The superposition frame $\mathscr{F}(I^g)$ satisfies the following* generalized *overlap-add constraint:*
$$\sum_{n=0}^{N-1}\sum_{r=0}^{N-1} I^g[n,r]\mathcal{T}_{na}w_r[t] = \frac{\widehat{w}[0]}{a}. \tag{6.19}$$

2. *Each $\mathscr{F}(I^g)$ satisfies the overlap-add reconstruction property that, for any $x \in \mathbb{C}^L$ with corresponding frame coefficients $\{X[m,n,r]\}$ defined by* (6.13)*,*
$$x[t] = \frac{a}{\widehat{w}[0]}\sum_{n=0}^{N-1}\sum_{r=0}^{N-1}\left(\frac{1}{M_g}\sum_{m=0}^{M_g-1} X[\tfrac{M_L}{M_g}m, n, r]e^{2\pi i m b_g t/L}\right). \tag{6.20}$$

6.6.2 Reconstruction via Canonical Dual Superposition Frames

We next develop the reconstruction properties of our superposition families in a frame-theoretic context, noting that they qualify as "painless nonorthogonal expansions" [213], and that the development below also holds for more general nonstationary Gabor frames [208]. In analogy to Definition 5, we associate a superposition frame operator $S_I : \mathbb{C}^L \to \mathbb{C}^L$ through its action on any $x \in \mathbb{C}^L$ as $S_I x = \sum_{\phi\in\mathscr{F}(I)}\langle x, \phi_{m,n,r}\rangle\phi_{m,n,r}$.

Definition 12 (Superposition Frame Operator, Walnut Form)**.** *The discrete Walnut representations of superposition frame operators S_{I^g} and S_{I^l} are respectively given by the $L \times L$ positive semi-definite matrices with entries*

$$S_{I^g}[t,t'] \triangleq M_g\,\mathbb{I}_{M_g\setminus(t-t')}[t-t']\sum_{n,r=0}^{N-1} I^g[n,r]\mathcal{T}_{na}w_r[t]\overline{\mathcal{T}_{na}w_r[t']},$$

$$S_{I^l}[t,t'] \triangleq \sum_{n,r=0}^{N-1} I^l[n,r]M_r\,\mathbb{I}_{M_r\setminus(t-t')}[t-t']\mathcal{T}_{na}w_r[t]\overline{\mathcal{T}_{na}w_r[t']}.$$

Theorem 2 implies that any Gabor frame $\mathscr{G}(w, a, b)$ and admissible I^{g} or I^l together give rise to a superposition frame, and hence the corresponding superposition frame operators are of full rank. Thus, to each global superposition frame $\mathscr{F}(I^{\mathrm{g}}) = \{\phi_{m,n,r}\}$ corresponds a unique canonical dual frame $\{\widetilde{\phi}_{m,n,r}\}$, whose elements are obtained in turn as

$\widetilde{\phi}_{m,n,r} \triangleq S_{I^{\mathrm{g}}}^{-1}\phi_{m,n,r}$. Accordingly, when $S_{I^{\mathrm{g}}}$ is diagonal we may index elements of $\{\widetilde{\phi}_{m,n,r}\}$ by the same admissible I^{g}, and we obtain the following reconstruction property:

$$\forall x \in \mathbb{C}^L, t \in \mathbb{Z}_L, \;\; x[t] = \sum_{m,n,r:I^{\mathrm{g}}[m,n,r]=1} \langle x, \phi_{m,n,r}\rangle \widetilde{\phi}_{m,n,r}[t],$$

with the above also holding for local I^l by Theorem 2. We thus denote by $\widetilde{\mathscr{F}(I^{\mathrm{g}})}$ or $\widetilde{\mathscr{F}(I^l)}$ the corresponding dual frames, and observe the following consequence of the Walnut representation of Definition 12 above.

Theorem 5 (Fast Inversion via Canonical Dual)**.** *For any $\mathscr{F}(I^g)$, $\mathscr{F}(I^l)$ derived from a Gabor frame $\mathscr{G}(w,a,b)$ on $\mathbb{C}^L$, the corresponding operators S_{I^g} and S_{I^l} are* diagonal*, and each canonical dual frame element takes the form*

$$\widetilde{\phi}_{m,n,r}[t] = \frac{\mathcal{M}_{mb_L}\mathcal{T}_{na}w_r[t]}{M_g\sum_{n'=0}^{N-1}\sum_{r'=0}^{N-1} I^g[n',r']\,|\mathcal{T}_{n'a}w_{r'}[t]|^2} \tag{6.21}$$

for $\mathscr{F}(I^g)$, and similarly for $\mathscr{F}(I^l)$ with respect to each M_r.

Note that the corresponding formula for the nonstationary Gabor frames of [208] in the diagonal case is similar to (6.21). However, in the superposition frame setting, the constraints on the window structure not only preserve lower frame bounds and yield fast inversion via the constant overlap-add method, but also enable signal-independent evaluation of the canonical dual in certain cases, as we now show.

6.6.3 Adaptive Lapped Superposition Frames

Reconstruction via the canonical dual $\widetilde{\mathscr{F}(I^{\mathrm{g}})}$ according to (6.21) requires knowledge of the admissible selection function $I^{\mathrm{g}}[n,r]$ corresponding to a given signal adaptation. This stands in contrast not only to the usual Gabor setting, wherein the form of the canonical dual frame can be obtained immediately, but also to the constant overlap-add approach described in Section 6.6.1, which avoids computation of the canonical dual entirely. However, by coupling our superposition construction with the following *neighbor overlap* condition, we are able to compute $\widetilde{\mathscr{F}(I^{\mathrm{g}})}$ *prior* to adaptation—that is, without knowledge of which ordered partition function will be used in subsequent signal analysis.

Definition 13 (Neighbor Overlap Condition)**.** *Let $\mathscr{G}(w,a,\cdot)$ be a Gabor system on $\mathbb{C}^L$, with $N = L/a$. It is said to satisfy the* neighbor overlap *condition if, for all $n, n' \in \mathbb{Z}_N$,*

$$\mathrm{supp}\,(\mathcal{T}_{na}w) \cap \mathrm{supp}\,(\mathcal{T}_{n'a}w) = \emptyset \text{ if } |n-n'| > 1. \tag{6.22}$$

Any admissible selection function preserves the neighbor overlap property, leading to the following notion of *lapped superposition frames*, whose properties we develop below.

Definition 14 (Adaptive Lapped Superposition Frames)**.** *Let $\mathscr{G}(w,a,\cdot)$ be a Gabor frame on $\mathbb{C}^L$ that* simultaneously *satisfies the overlap-add constraint of* (6.17) *and the neighbor-overlap condition of* (6.22)*. Then for any admissible I^g, we call $\mathscr{F}(I^g)$ an adaptive* lapped superposition frame.

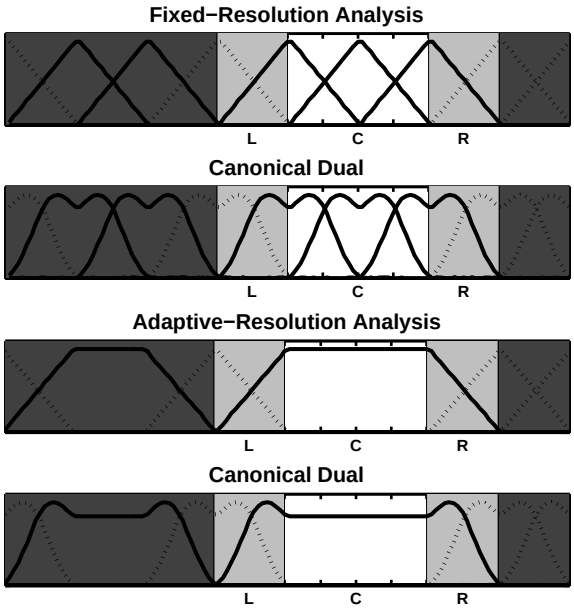


Figure 6.3: Illustration of adaptive, lapped superposition frames. Translates $\{\mathcal{T}_{na}w\}$ from a Gabor frame $\mathscr{G}(w,a,\cdot)$ constructed from triangular windows with 50% overlap (top panel), shown with translates of its canonical dual window $\widetilde{w}$ (second panel); remaining panels repeat this sequence for an induced superposition frame $\mathscr{F}(I^{\mathrm{g}})$. On the left and right sets L and R, the canonical dual windows of $\widetilde{\mathscr{F}(I^{\mathrm{g}})}$ agree pointwise with those of $\mathscr{G}(\widetilde{w},a,\cdot)$; on the center set C, they are constant.

Note that the overlap-add constraint of (6.17) ensures a partition of unity by window translates, while the neighbor overlap condition of (6.22) is also required in the case of lapped *orthogonal* transforms (see, e.g., [185]). While our construction retains the flavor of time-varying lapped transforms [214, 215], we emphasize that the resultant frames can avoid the lack of translation invariance inherent in the orthogonal setting, while still ensuring fast reconstruction.

We show below that if $\mathscr{F}(I^{\mathrm{g}})$ is a lapped superposition frame derived from a Gabor frame $\mathscr{G}(w,a,b)$, then its canonical dual frame elements may be pre-computed. This situation is illustrated in Figure 6.3, where the support sets of a window w, its canonical dual $\widetilde{w}$, and their immediate neighbors are partitioned into subsets labeled L, C, and R. Since $\mathscr{F}(I^{\mathrm{g}})$ must inherit the neighbor-overlap condition from $\mathscr{G}(w,a,b)$, it follows that whenever $\mathscr{F}(I^{\mathrm{g}})$ admits a diagonal frame operator, the corresponding canonical dual windows of $\widetilde{\mathscr{F}(I^{\mathrm{g}})}$ are *constant* on the center set C, as shown in Figure 6.3. Moreover, the highlighted dual superposition window is pointwise equal to $\widetilde{w}$ on the sets L and R (see bottom two panels of Figure 6.3).

To formalize this intuition, define for each r the sets

$$\begin{cases} L_r \triangleq \mathrm{supp}(w_r) \cap \mathrm{supp}(\mathcal{T}_{-(r+1)a}w_r), \\ R_r \triangleq \mathrm{supp}(w_r) \cap \mathrm{supp}(\mathcal{T}_{(r+1)a}w_r), \\ C_r \triangleq \mathrm{supp}(w_r) \setminus (L_r \cup R_r), \end{cases} \tag{6.23}$$

and note that, for any global selection function I^{g}, $\phi_{\cdot,n,r} \in \mathscr{F}(I^{\mathrm{g}})$ and its dual $S_{I^{\mathrm{g}}}^{-1}\phi_{\cdot,n,r} \in \widetilde{\mathscr{F}(I^{\mathrm{g}})}$ are *both* supported exclusively on the set $\mathcal{T}_{na}(L_r \cup C_r \cup R_r)$. Then the following theorem, proved in Appendix 6.A, establishes our main result: the canonical dual frame of any $\mathscr{F}(I^{\mathrm{g}})$ can be computed independently of any ordered partition function, and, therefore, can be computed *prior* to observing any data.

Theorem 6 (Canonical Duals of Adaptive Lapped Frames)**.** *Let $\mathscr{F}(I^g)$ arise from a Gabor frame $\mathscr{G}(w,a,\cdot)$ for $\mathbb{C}^L$ satisfying* (6.17) *and* (6.22)*. Then every $\widetilde{\phi}_{m,n,r} \in \widetilde{\mathscr{F}(I^g)}$ can be constructed by modulations $\mathcal{M}_{mb_L}$ and translations $\mathcal{T}_{na}$ of lapped windows corresponding to each r as follows:*

$$\widetilde{\phi}_{0,0,r}[t] \triangleq \frac{1}{M_g} \cdot \begin{cases} \frac{w[t]}{\sum_{n=0}^{N-1} |w[t-na]|^2} & \text{if } t \in L_r, \\ \frac{a}{\widehat{w}[0]} & \text{if } t \in C_r, \\ \frac{w[t-(r+1)a]}{\sum_{n=0}^{N-1} |w[t-na]|^2} & \text{if } t \in R_r. \end{cases}$$

Here the sets L_r, C_r, R_r are defined in (6.23)*, and we note that only the expression for C_r is to be employed when $\mathscr{F}(I^g)$ is comprised entirely of modulations of w_{N-1}.*

We note that many popular Gabor systems $\mathscr{G}(w,a,\cdot)$ satisfy the requirements of this theorem—including triangular, Hamming, and raised-cosine windows w at 50% overlap—thus enabling a variety of new adaptive, lapped superposition frame families that all admit fast reconstruction.

6.6.4 Adaptive Dyadic Superposition Frames

Here we construct a class of so-called *dyadic* superposition frames that admit offline canonical dual construction even when the overlap-add constraint of (6.17) is *not* satisfied. We base this construction on the notion of *dyadic* ordered partition functions, which may be thought of as indexing binary trees.

Definition 15 (Dyadic Admissible Selection Functions)**.** *Let the number of translates N of w in $\mathscr{G}(w,a,b)$ be a power of two, and define the set $\mathscr{H} \triangleq \{0,1,\ldots,\log_2 N\}$. An ordered partition function $\widetilde{I}^d[n,r]$ is* dyadic *if it satisfies the conditions of Definition 8 and*

$$\widetilde{I}^d[n,r] = 1 \text{ only } \textit{if } r = 2^h - 1 \textit{ for some } h \in \mathscr{H}.$$

We denote by $I^d[m,n,r]$ a dyadic *admissible selection function induced by $\widetilde{I}^d[n,r]$ and a global frequency lattice constant $b_g = L/M_g$, with M_g defined according to* (6.16)*.*

Viewing the N translates of w in $\mathscr{G}(w,a,b)$ as leaves of a binary tree of height $\log_2 N$, a dyadic ordered partition function selects windows corresponding to some tree level h.

Definition 16 (Dyadic Gabor and Superposition Frames)**.** *Fix an initial Gabor frame $\mathscr{G}(w,a,b)$, such that $N = L/a$ is a power of two, and let $h \in \mathscr{H}$ index* height *in a binary tree. Now restrict $r \in \mathbb{Z}_n$ to the index set $\mathscr{R} \triangleq \{2^h - 1\}$, and fix for each $r \in \mathscr{R}$ a time lattice constant $a_r \triangleq a(r+1)$, and a dyadic admissible selection function I^d with associated global frequency lattice constant b_g. Then:*

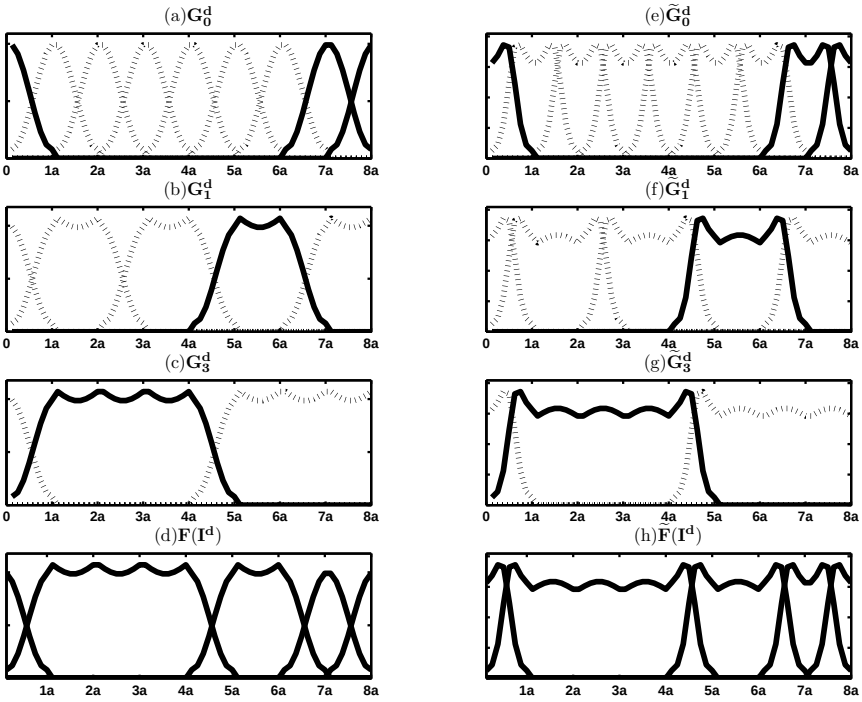


Figure 6.4: Illustration of adaptive dyadic superposition frames. Repeated superposition merges of Hamming windows, following the structure of a binary tree, are shown in panels (a-c), along with the canonical dual windows associated to each corresponding Gabor system (e-g). A dyadic superposition frame $\mathscr{F}(I^d)$ can be formed from the selected unmodulated elements of (d), which in turn will admit the corresponding canonical dual windows shown in (h), computed according to (6.21).

1. *We define* dyadic superposition Gabor frame $G_r^d(b_g)$ *for all* $r \in \mathscr{R}$*, and their union* $G_{\cup r}^d(b_g)$*, as follows:*

$$G_r^d(b_g) \triangleq \mathscr{G}(w_r, a_r, b_g) \subseteq \mathscr{G}(w_r, a, b_L),$$
$$G_{\cup r}^d(b_g) \triangleq \cup_{r\in\mathscr{R}} G_r^d(b_g) \subseteq \cup_{r=1}^{N-1} \mathscr{G}(w_r, a, b_L).$$

2. *We call* $\mathscr{F}(I^d) \subset G_{\cup r}^d(b_g)$ *a* dyadic superposition frame.

The fact that $\mathscr{F}(I^d)$ and every dyadic $G_r^d(b_\mathrm{g}) = \mathscr{G}(w_r, a_r, b_\mathrm{g})$ are frames for $\mathbb{C}^L$ follows from the assumption that $\mathscr{G}(w, a, b)$ is a frame, coupled with the result of Theorem 2.

An example of this construction is illustrated in Figure 6.4: The left panel shows Hamming windows $w \equiv w_0$ at 50% overlap, along with the corresponding dyadically-indexed superposition windows w_1 and w_3, and an example dyadic superposition frame $\mathscr{F}(I^d)$ at bottom. The right panel shows canonical duals associated to each dyadic superposition Gabor frame $G_r^d(\cdot)$, for $r \in \{0, 1, 3\}$, along with the corresponding canonical dual $\widetilde{\mathscr{F}(I^d)}$ at bottom right. The fact that $\widetilde{\mathscr{F}(I^d)}$ contains elements from each of these individual dual Gabor frames $\widetilde{G_r^d(\cdot)}$ is verified by the following theorem.

Theorem 7 (Canonical Duals of Adaptive Dyadic Frames)**.** *Let a dyadic superposition frame $\mathscr{F}(I^d) \subset G^d_{\cup r}(b_g)$ arise from a Gabor frame $\mathscr{G}(w,a,b)$ for $\mathbb{C}^L$ satisfying the neighbor-overlap condition of* (6.22)*. Then its canonical dual $\widetilde{\mathscr{F}(I^d)}$ can be computed by either path of the following commutative diagram, where S_{I^d} is the frame operator associated to $\mathscr{F}(I^d)$, and S^d_r is that associated to each dyadic Gabor frame G^d_r:*

$$\begin{array}{ccc} G^d_{\cup r}(b_g) & \xrightarrow{(S^d_r)^{-1}, r\in\mathscr{R}} & \cup_{r\in\mathscr{R}}\widetilde{G^d_r(b_g)} \\ \Big\downarrow I^d & & \Big\downarrow I^d \\ \mathscr{F}(I^d) & \xrightarrow{(S_{I^d})^{-1}} & \widetilde{\mathscr{F}(I^d)} \end{array} . \tag{6.24}$$

Observe that computing the $\widetilde{\mathscr{F}(I^d)}$ via direct inversion of its (diagonal) frame operator S_{I^d} corresponds to the down-and-right path in (6.24), and hence requires knowledge of I^d. However, Theorem 7 implies that all elements in $\widetilde{\mathscr{F}(I^d)}$ can be pre-computed by instead following the right-and-down path. As in the case of lapped superposition frames in Section 6.6.3, the neighbor overlap condition of (6.22) plays a key role.

6.6.5 Computational Complexity

We now address the relative computational complexity of the various reconstruction methods presented above. Recall that these algorithms all yield a diagonal frame operator, implying that only $\mathcal{O}(L)$ operations are required for its inversion, compared to $\mathcal{O}(L^3)$ in the worst case of a non-diagonal frame operator lacking any special structure. While in some circumstances this complexity can be reduced (see, e.g., [209]), such schemes remain super-linear in L, rendering them impractical for use in applications with $L \gg 1$.

Recall that when no windows are merged, we recover a Gabor frame $\mathscr{G}(w,a,b)$ containing NM elements, with $Na = Mb = L$. If the associated frame operator is diagonal, an FFT-based approach requires $NM(1+\log_2 M)$ complex multiplications to compute the short-time analysis coefficients $X[m,n]$ from $x \in \mathbb{C}^L$, with NM of these needed to obtain the N individual short-time segments $\{\mathcal{T}_{na}w[t]x[t]\mathbb{I}_{\mathrm{supp}(\mathcal{T}_{na}w)}[t]\}$. When reconstruction proceeds via inversion of the frame operator, then $NM\log_2 M$ operations are required to compute the necessary inverse FFTs, plus NM operations to multiply each resultant segment by the appropriate dual window. In the overlap-add setting, this window is the identity, and hence these latter MN operations are avoided. An in-depth discussion of the general (non-diagonal) case, including methods based on the Zak transform, is given in [209] and [216].

In the case that windows are merged, note that a global superposition frame $\mathscr{F}(I^{\mathrm{g}})$ derived from $\mathscr{G}(w,a,b)$ comprises $M_{\mathrm{g}} = L/b_{\mathrm{g}} \geq M$ modulations of $N_{\mathrm{g}} \leq N$ windows:

$$N_{\mathrm{g}} = \sum_{n=0}^{N-1}\sum_{r=0}^{N-1} I^{\mathrm{g}}[n,r]. \tag{6.25}$$

It follows that $N_{\mathrm{g}}M_{\mathrm{g}}(\log_2 M_{\mathrm{g}} + 1)$ complex multiplications are required to compute the analysis coefficients $X[m,n,r]$ via N_{g} FFTs, followed by $N_{\mathrm{g}}M_{\mathrm{g}}\log_2 M_{\mathrm{g}}$ operations required for the N_{g} inverse FFTs required for reconstruction.

Table 6.1: Computational complexity orders $\mathcal{O}(\cdot)$ of various analysis and synthesis algorithms considered in this chapter

	No Adaptation		Adaptation		
Analysis	$NM(1+\log_2 M)$		$N_{\mathrm{g}}M_{\mathrm{g}}(1+\log_2 M_{\mathrm{g}})$		
Synth. Method	Overlap-Add	Canon. Dual	Section 6.6.1	Section 6.6.2	Sections 6.6.3
Synthesis	$NM\log_2 M$	$NM(1+\log_2 M)$	$N_{\mathrm{g}}M_{\mathrm{g}}(\log_2 M_{\mathrm{g}})$	$N_{\mathrm{g}}M_{\mathrm{g}}(3+\log_2 M_{\mathrm{g}})$	$N_{\mathrm{g}}M_{\mathrm{g}}(1+\log_2 M_{\mathrm{g}})$

If $\mathscr{F}(I^{\mathrm{g}})$ arises from the special case described in Theorem 6, then the elements of $\widetilde{\mathscr{F}(I^{\mathrm{g}})}$ can be pre-computed, and only $N_{\mathrm{g}}M_{\mathrm{g}}$ extra operations are necessary for multiplication by the requisite canonical dual windows. In the general case, elements of $\widetilde{\mathscr{F}(I^{\mathrm{g}})}$ must be computed directly via (6.21), as a function of the chosen selection function I^{g}. Appealing to (6.25), we see that this computation can be accomplished using another $2N_{\mathrm{g}}M_{\mathrm{g}}$ calculations, leading to a total of $N_{\mathrm{g}}M_{\mathrm{g}}(3+\log_2 M_{\mathrm{g}})$ complex multiplications. This analysis is also an upper bound for the worst-case complexity for any $\mathscr{F}(I^l)$ having the same ordered partition function as $\mathscr{F}(I^{\mathrm{g}})$.

The various analysis and synthesis complexities discussed above are summarized in Table 6.1. In practice, the complexity of the signal adaptation procedure must also be taken into account. This complexity depends both on the method for searching among ordered partition functions $\widetilde{I}[n,r]$, and the cost function used to compare them. Since it is clearly infeasible to compare all 2^{N-1} ordered partition functions by exhaustive search, we next consider greedy and dynamic-programming-based approaches below. In both cases, the complexity of evaluating the associated cost functions increases with window length, and therefore it is advisable in practice to set an upper bound on the maximal number of window merges.

6.7 Signal Adaptation Algorithms and Examples

Signal-adaptive modification of an initial Gabor frame $\mathscr{G}(w,a,b)$ on $\mathbb{C}^L$ via superposition can produce any one of the possible 2^{N-1} superposition frames whose properties we characterized in Sections 6.5 and 6.6. We now detail two instances of a broad class of signal adaptation algorithms, any of which can be used to select a superposition frame for subsequent signal analysis. We propose both greedy and dynamic programming approaches in Section 6.7.1, and illustrate their performance with two brief examples in Section 6.7.2.

6.7.1 Signal-Adaptive Superposition Frame Selection

The first of the two Gabor frame adaptation algorithms we describe is a simple greedy approach that can be implemented by "growing" a given window forward in time through successive attempts to merge it with its subsequent neighboring translates [153]. Whenever a proposed merge fails, the procedure resets and repeats, halting when the end of the data stream is reached (or, equivalently in our cyclic setting, when the initial window is once again encountered).

A decision whether or not to merge adjacent windows can be made based on any suitable cost function. As one example, we employ the TF concentration measure appearing in the popular work of [180, 182] on adaptive optimal-kernel TF representations.

Specifically, consider a short-time segment $x_{n,r}[t] \triangleq \{\mathcal{T}_{na}w[t]x[t]\mathbb{I}_{\mathrm{supp}(\mathcal{T}_{na}w)}[t]\}$, and define its time-frequency concentration in the manner of [153, 180–182]

$$C(x_{n,r}) \triangleq \frac{\sum_{m=0}^{M_r-1} |\langle x, \mathcal{M}_{mb_r}\mathcal{T}_{na}w_r\rangle|^4}{\left(\sum_{m=0}^{M_r-1} |\langle x, \mathcal{M}_{mb_r}\mathcal{T}_{na}w_r\rangle|^2\right)^2}, \tag{6.26}$$

with M_r, b_r defined via (6.15). This ratio of powers of norms of short-time Fourier coefficients is suggestive of an "empirical spectral kurtosis," and has also been used in minimum entropy deconvolution [217]; other choices are also possible [192].

As shown in [180], maximizing (6.26) favors short-time segments that *concentrate* local signal energy within the smallest regions of the TF plane. Indeed, below we obtain similar results on an example akin to the one employed in [180]: the resultant superposition frames comprise shorter windows near time-localized transients, and longer windows near oscillatory signal portions. The resulting procedure requires $\mathcal{O}(N)$ iterations and is summarized in Algorithm 6.1.

Algorithm 6.1 Greedy Signal-Adaptive Superposition Frame Selection [153]

Initialization

- Fix input data $x \in \mathbb{C}^L$ and a Gabor frame $\mathscr{G}(w, a, b)$
- Set $(p, n_p) = (0, 0)$ and initialize $\widetilde{I}[n, r]$ to be the N-part ordered partition function of Example 1

Greedy Selection: For $n = 0, 1, \ldots, N-1$,

- Compute a merged window
$$\mathcal{T}_{n_p a}w_{p+1} = \mathcal{T}_{n_p a}w_p + \mathcal{T}_{na}w$$
and $C(x_{n_p,p+1})$, $C(x_{n_p,p})$ and $C(x_{n,0})$ via (6.26)
- If $C(x_{n_p,p+1}) \leq \max(C(x_{n_p,p}), C(x_{n,0}))$, reject the proposed merge: set (p, n_p) as $(p, n+p+1)$, and leave $\widetilde{I}[n, r]$ unchanged
- Otherwise accept the proposed merge: set (p, n_p) as $(p+1, n_p)$ and update $\widetilde{I}[n, r]$ as
$$\begin{aligned} &\widetilde{I}[n_p, p+1] = 1 && \text{(add: } \mathcal{T}_{n_p a}w_{p+1}\text{)}, \\ &\widetilde{I}[n_p, p] = \widetilde{I}[n, 0] = 0 && \text{(remove: } \mathcal{T}_{n_p a}w_p, \mathcal{T}_{na}w\text{)} \end{aligned}$$

Output: Return the set of variable-length windows induced by $\widetilde{I}[n, r]$

The second algorithm we present is based on the dynamic programming approach to adaptive segmentation popular in the audio coding literature [198–202]. The basic idea is to fix an *additive* cost function $J(\cdot)$, and find an optimal ordered partition function $\widetilde{I}^*[n, r]$ in the sense that it minimizes the sum of individual segment costs $J(x_{n,r}[t])$:

$$\widetilde{I}^*[n, r] \triangleq \underset{\widetilde{I}[n,r]}{\mathrm{argmin}} \sum_{n,r:\widetilde{I}[n,r]=1} \widetilde{I}[n, r]J(x_{n,r}[t]).$$

Many choices for $J(\cdot)$ are possible, including rate-distortion cost functions [199], sparsity-inducing measures [192], and the well-known entropy cost of [196], which we employ below.

To formalize our approach, define J_n^* as the minimum cost among ordered partition functions on $\{0, 1, \ldots, na-1\}$, and let $J_{n,r} \triangleq J(x_{n,r}[t])$ represent the cost associated to covering the region $\{na, na+1, \ldots, (n+r)a\}$. The resulting dynamic program requires $\mathcal{O}(N^2)$ iterations and is summarized in Algorithm 6.2.

Note that to preserve cost additivity in the presence of overlapping, non-orthogonal windows, Algorithm 6.2 evaluates $J(x_{n,r}[t])$ on regions *smaller* than those covered by the corresponding windows, in a manner which recovers the approach of [199] in the block-Fourier case.

Once a set of variable-length windows is obtained via any selection procedure returning an ordered partition function, a local or global modulation structure can be chosen via (6.15) or (6.16), respectively, in order to obtain a signal-adaptive superposition frame. In practice, application-specific considerations are likely to play a role in superposition frame selection, and to this end we note that a variety of other algorithms and approaches are possible. In particular, we note that the merging criteria based on parametric and, especially, nonparametric hypothesis tests for stationarity described in Chapter 5 are natural approaches to signal adaptation (see, e.g., [117, 155, 218]).

Algorithm 6.2 Signal-Adaptive Superposition Frame Selection via Dynamic Programming [198]

Initialization

- Fix input data $x \in \mathbb{C}^L$, a Gabor frame $\mathscr{G}(w, a, b)$ and initialize the cost function $J_0^* = 0$
- For each $\mathcal{T}_{na}w \in \{T_{na}w : n \in \mathbb{Z}_N\}$ calculate the support set

$$\mathscr{D}_n \triangleq \{t : \mathcal{T}_{na}w[t] > \mathcal{T}_{n'a}w[t], n \neq n' \in \mathbb{Z}_N\}$$

Dynamic Program

- For $n = 0, 1, \ldots, N-1$, compute sequentially the nth segmental cost and associated boundary by

$$J_n^* = \min_{0 \leq r < n} \left(J_r^* + J_{n,r}\right) \qquad b_n^* = \arg\min_{0 \leq r < n} \left(J_r^* + J_{n,r}\right),$$

  with $J_{n,r}$ calculated using signal data supported on $\cup_{k=n}^{n+r} \mathscr{D}_k$
- Compute the optimal selection function $\widetilde{I}^*[n, r]$ using $\{b_n^* : n \in \mathbb{Z}_N\}$ via the standard "backtracking" procedure [198]

Output: Return the set of variable-length windows induced by $\widetilde{I}^*[n, r]$

6.7.2 Illustrative Examples

To conclude our investigation of superposition frames, we now consider two illustrative examples that combine Algorithms 6.1 and 6.2 with the analysis and reconstruction procedures presented earlier. These examples—a stylized synthetic waveform akin to the

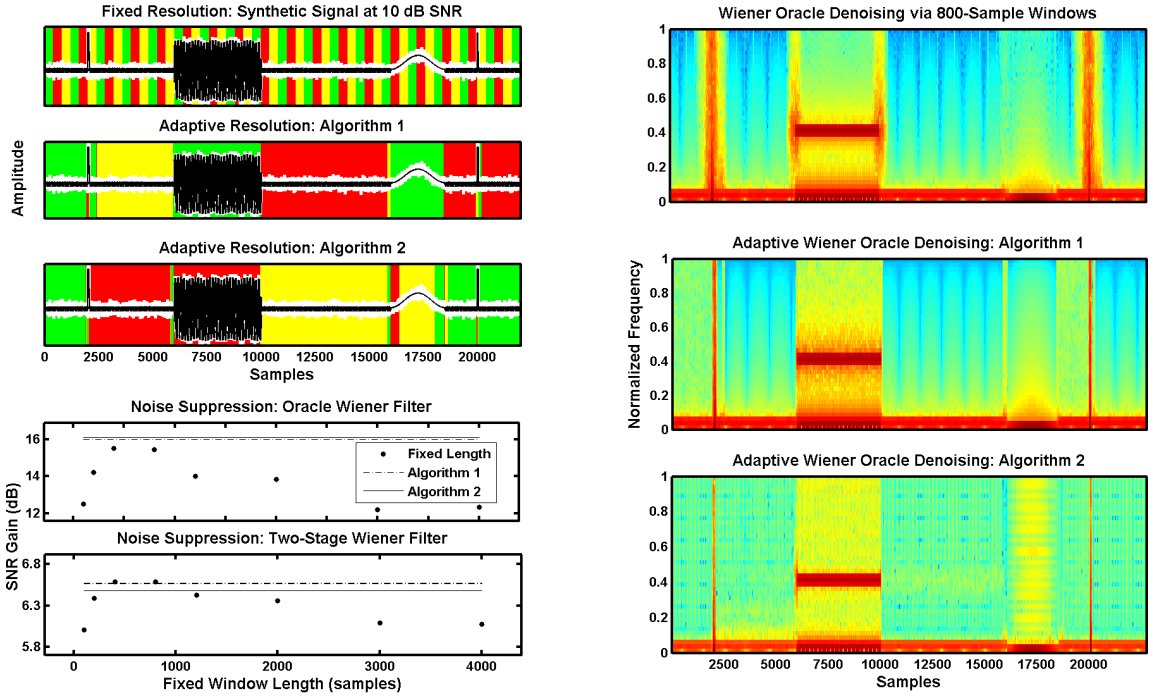


Figure 6.5: Adaptive analysis-synthesis of a noisy synthetic test signal (10 dB SNR) via superposition frames. Top left: Adaptive segmentations of the *noisy* signal (white, with clean version superimposed in black); background rectangles highlight the temporal extent of selected superposition windows. Bottom left: Results of a noise suppression experiment using an oracle (resp. two-stage) Wiener filter, averaged over 50 trials. Standard deviations range from 0.13–0.27 dB (oracle) and from 0.03–0.06 dB (two-stage). Right: spectrograms formed from the oracle-denoised signals using 80 sample Hanning windows with 50% overlap.

example employed in [180] and a phonetically balanced speech utterance from the TIMIT corpus [219]—both exhibit varying TF structure, which in turn motivates signal-adaptive analysis and reconstruction.

Our first example signal x (Figure 6.5, top left) comprises a local and global sinusoidal term, two impulses, and a bump function. We conducted a variety of experiments in which varying levels of white Gaussian noise n were added to x, and Algorithms 6.1 and 6.2 were then applied to $y \triangleq x + n$ to obtain signal-adaptive frame analysis coefficients $Y[m, n, r]$ on a global frequency lattice ($M_{\mathrm{g}} = 6000$). Using these as well as fixed-resolution analyses for a range of window lengths, we then applied to $Y[m, n, r]$ both an "oracle" Wiener suppression rule (local signal spectrum estimated by $|X[m, n, r]|^2$) and a two-stage Wiener suppression rule (by appropriately soft-thresholding $|Y[m, n, r]|^2$), and obtained a time-domain reconstruction $\widehat{x}$ via the corresponding canonical dual superposition frame.

The remainder of Figure 6.5 reports the results of a typical run at 10 dB signal-to-noise ratio (SNR), with the superposition system of Algorithm 6.1 derived from an initial Gabor system $\mathscr{G}(w, a, \cdot)$ comprising a 100-sample Hamming window w and time lattice constant $a = 50$, and that of Algorithm 6.2 based on a 65-sample Hamming window with

$a = 32$. Although we have observed Algorithm 6.2 to be more noise-robust in practice, and to yield better performance with somewhat shorter initial windows, it may be seen that both algorithms yield broadly similar analyses with respect to dominant signal features at 10 dB SNR. Moreover, over a range of noise levels and fixed-resolution analyses, we have observed improved SNR gains $20 \log_{10}(\|y - x\|/\|\widehat{x} - x\|)$ in both the oracle and two-stage cases, as shown in the bottom-left panels of Figure 6.5.

Reconstruction spectrograms $20 \log_{10} |\widehat{X}[m, n, r]|$—based on the *oracle* denoising for visual clarity—are shown in the right-hand panel of Figure 6.5. They indicate that, in comparison to an *a priori* well chosen fixed-resolution analysis using 800-sample Hamming windows with $a = 400$, the onsets and offsets of localized TF features are better preserved by superposition frames. Since the best fixed-resolution window length is not known *a priori* in practice, the adaptive approach remains attractive, despite the lessening SNR gains obtained in the simple two-stage denoising approach. These results suggest the investigation of more sophisticated denoising schemes, also bearing in mind that in the case of nonstationary noise, the best adaptive analysis may well be SNR-dependent.

We repeated the same battery of tests with our second example signal x (Figure 6.6, bottom), a portion of the phonetically-balanced TIMIT speech waveform corresponding to the phrase "... [eye]d and amazed" (/train/dr1/fsah0/si1244.wav). This utterance, chosen to illustrate time-varying spectral content typical of speech, contains two plosives ([eye]*d*, amaze*d*), two steady vowels (*a*nd, *a*mazed), and a time-varying diphthong (am*a*zed). The exact phonetic TIMIT segmentation (si1244.phn) is shown between the two spectrogram panels of Figure 6.6, which correspond respectively to reconstructions based on an *a priori* well chosen fixed-resolution (top, 30 ms) and adaptive-resolution (middle, Algorithm 6.2, starting from 3 ms Hamming windows with $a = 1.5$ ms) oracle-Wiener denoising, respectively, at 10 dB SNR. The bottom panel of Figure 6.6 illustrates the corresponding adaptive analysis, which is seen to agree well with major features of the given TIMIT segmentation; Algorithm 6.1 also yielded a similar analysis.

Following the same experimental procedure as in the case of Figure 6.5, we observed broadly similar results—though with lower overall SNR gains obtained for this less stylized example. Importantly, however, superposition windows are seen to better preserve vowel onsets and plosives; see in particular the boxed regions of the fixed-resolution spectrogram in the top panel of Figure 6.6, corresponding to the two plosives and initial vowel-diphthong onsets in the word "amazed."

6.7.3 Informal Listening Tests

In addition, we conducted a series of informal listening tests with ten trained listeners at MIT Lincoln Laboratory to evaluate whether an enhancement system based on a signal-adaptive TF analysis reduces the amount of musical noise in the enhanced waveform as compared to a fixed-resolution system. Specifically, we report the results using the superposition system of Algorithm 6.1 derived from an initial Gabor system $\mathscr{G}(w, a, \cdot)$ comprising a 10 ms triangular window w and a time lattice constant corresponding to 5 ms. These results are compared to an *a priori* well chosen fixed-resolution analysis using 20 ms triangular windows with 10 ms overlap.

Listeners were presented with the voiced sentences: "Why were you away a year

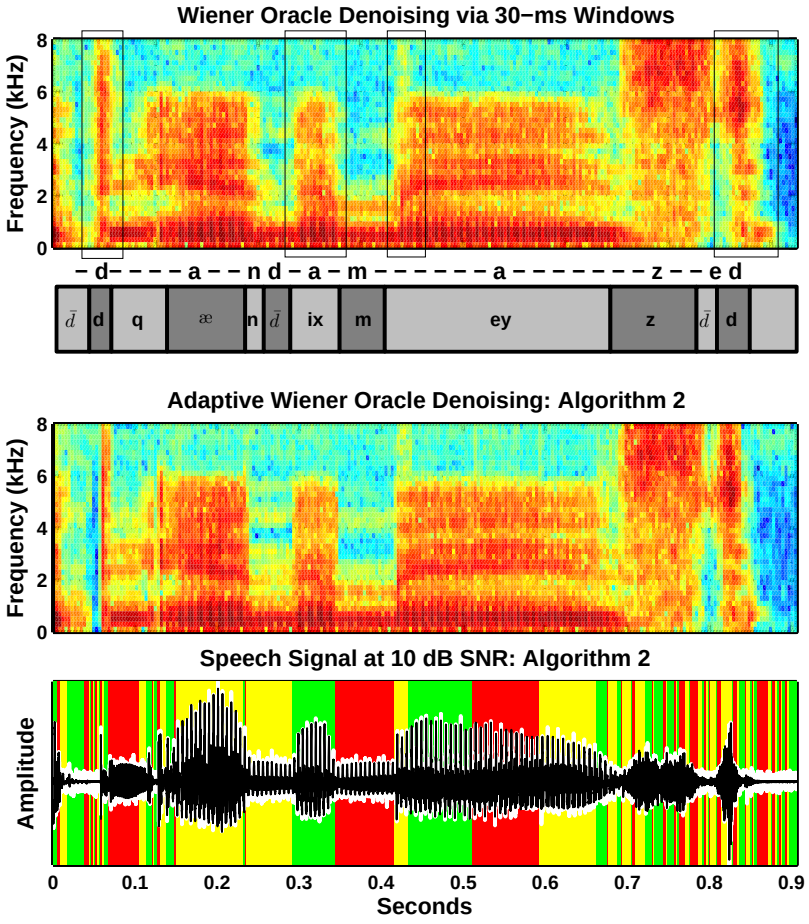


Figure 6.6: Adaptive analysis-synthesis of a noisy speech signal (10 dB SNR). Bottom: adaptive segmentation via Algorithm 6.2, shown with spectrogram of adaptive oracle-Wiener-denoised version. Top: spectrogram of fixed-resolution (30-ms) denoised version, shown with orthographic and phonetic transcriptions; boxes highlight temporal smearing of plosives and vowel onsets. Both spectrograms were formed using 5 ms Hanning windows with 50% overlap.

Roy?" and "Nanny may know my meaning," each spoken by two male and two female speakers, at 0 dB and 5 dB SNR, for a total of 16 utterances. Each listener was also presented with four TIMIT sentences (2 male speakers, 2 female speakers) for a total of 8 utterances. For each presented waveform the listener heard the clean and the noisy samples followed by two repetitions of the enhanced sentences by the adaptive- and fixed-resolution systems in a random order. At the end of the presentation the listener was asked which of the enhanced waveforms, if any, had musical noise. The results are summarized in Table 6.2. Among the voiced data, a large majority of responses (74.4%) indicated that less musical noise was present in the waveforms enhanced by the adaptive system—less than 5% preferred the fixed resolution system. The results using the phonetically-balanced TIMIT utterances are more modest, but nonetheless promising. We suspect that the performance gap between the two cases can be closed through the incorporation of a probabilistic model of speech presence. Overall, however, the results show that the adaptation can help reduce

Table 6.2: Summary of listener preferences from tests for the enhancement scheme resulting in least musical noise. The results are averaged over the responses of all ten listeners for all voiced and TIMIT utterances.

	Adaptive-Resolution System	No Preference	Fixed-Resolution System
Voiced	74.3%	21.3%	4.4%
TIMIT	33.8%	53.7%	12.5%

the musical noise artifact.

In this manner we see that superposition frames, when coupled with appropriate waveform adaptation criteria, show strong potential for use in a variety of signal-adaptive analysis-synthesis settings. For signal enhancement applications, a natural next step would be to extend the approach of [155], in which adaptive segmentation is used to estimate the local signal spectrum for enhancement purposes, but reconstruction is done using a fixed-resolution TF lattice. A variety of other multi-stage or iterative approaches suggest themselves, given the additional flexibility engendered by the overcomplete, signal-adaptive superposition frames presented in this chapter.

6.8 Summary

In this chapter we have introduced a new set of tools for nonstationary signal analysis. Specifically, we developed a broad family of adaptive, linear time-frequency representations termed superposition frames, and showed that they admit a host of desirable properties, including preservation of the lower frame bound in the adaptation process and fast overlap-add reconstruction akin to standard short-time Fourier techniques. We introduced online and offline algorithms for selecting the best superposition frame for adaptive analysis of a signal of interest. Through a discussion of these signal adaptation criteria and multiple examples, the resultant analysis-synthesis systems were shown to provide an effective and practical method for realizing signal-adaptive TF analysis coupled with fast reconstruction. We illustrated our approach with speech enhancement examples, which, together with informal listening tests, indicated the benefit of our construction.

6.A Appendix: Theorem Proofs

Proof of Lemma 2. To establish (6.9) for all $w \in \mathbb{C}^L$ and $M \geq \mathrm{len}(w)$, expand the left-hand side of (6.9) as

$$\sum_{m=0}^{M-1} |\langle x, \mathcal{M}_{mb} w\rangle|^2 = \sum_{m=0}^{M-1} \langle x, \mathcal{M}_{mb} w\rangle \overline{\langle x, \mathcal{M}_{mb} w\rangle} = \sum_{t=0}^{L-1}\sum_{t'=0}^{L-1} x[t]\overline{x[t']}\,\overline{w[t]}w[t'] \sum_{m=0}^{M-1} e^{-2\pi i m b(t-t')/L}$$
$$= M \sum_{t=0}^{L-1}\sum_{t'=0}^{L-1} \mathbb{I}_{M\backslash(t-t')}[t-t']\, x[t]\overline{x[t']}\,\overline{w[t]}w[t']. \tag{6.27}$$

Now consider all $M \geq \mathrm{len}(w)$ that divide $t - t'$; since $\overline{w[t]}w[t'] = 0$ for all $|t - t'| \geq \mathrm{len}(w)$, we need only consider the case $t - t' = 0$, whereupon we recover from (6.27) the right-hand side of (6.9). □

Proof of Lemma 3. To prove sufficiency, assume that $M_0 = M$, and hence $b_0 = b$. From (6.9) of Lemma 2, it follows that the right-hand side of (6.10) may be expanded as

$$M\sum_{t=0}^{L-1}|x[t]|^2\Big(|w_p[t]|^2+|w_q{}'[t]|^2+2\mathrm{Re}\{\overline{w_p[t]}w_q{}'[t]\}\Big), \tag{6.28}$$

with the rightmost term nonnegative by Definition 7. Dropping this term from (6.28) and applying (6.9) again, this time in the reverse direction, shows that (6.10) holds for any $M_0 \in \{\max(\mathrm{len}(w_p),\mathrm{len}(w_q)),\ldots,M\}$, thus proving sufficiency.

To prove necessity, assume to the contrary and consider a setting in which $M < \mathrm{len}(w_p + w_q{}') = L$. Noting that (6.10) may be stated as $\langle S^{(p,q)}x,x\rangle \le \langle S^{(p+q)}x,x\rangle$ for positive semi-definite frame operators $S^{(p,q)}$ and $S^{(p+q)}$, assume that elements of the former span $\mathbb{C}^L$ and thus form a frame; hence $\langle S^{(p,q)}x,x\rangle > 0$ for all nonzero $x \in \mathbb{C}^L$. However, since $M < L$, the M elements of the latter cannot span $\mathbb{C}^L$. Hence there exists at least one nonzero $x \in \mathbb{C}^L$ such that $\langle S^{(p+q)}x,x\rangle = 0$, thus contradicting the stated inequality. □

Proof of Theorem 1. To establish the theorem we directly bound the quantity $\sum_{\phi\in\mathscr{F}(I)}|\langle x,\phi_{m,n,r}\rangle|^2$ from below. To begin, observe that

$$\begin{aligned}\sum_{\phi\in\mathscr{F}(I)}|\langle x,\phi_{m,n,r}\rangle|^2 &= \sum_{m=0}^{M_\mathrm{g}-1}\sum_{n,r=0}^{N-1} I[n,r]\left|\langle x,\mathcal{M}_{mb_\mathrm{g}}\mathcal{T}_{na}w_r\rangle\right|^2\\ &= M_\mathrm{g}\sum_{n,r=0}^{N-1} I[n,r]\sum_{t,t'=0}^{L-1} x[t]\overline{x[t']}\mathcal{T}_{na}(\overline{w_r[t]}w_r[t'])\mathbb{I}_{M_\mathrm{g}\setminus(t-t')}[t-t'],\end{aligned}$$

with the latter expression above obtained by the expansion of (6.27). Now, if M_g divides $t-t'$, then $t' = t - k(t)M_\mathrm{g}$ for some integer $k(t)$, whose domain is deduced by observing that $0 \le t - k(t)M_\mathrm{g} \le L-1$ and $0 \le t' \le L-1$ together imply that $k(t) \in \mathscr{K} \triangleq \{\lceil (t-(L-1))/M_\mathrm{g}\rceil,\ldots,\lfloor t/M_\mathrm{g}\rfloor\}$. Then, a change of variable for t' yields the simplification

$$M_\mathrm{g}\sum_{t=0}^{L-1}\sum_{n,r=0}^{N-1}\sum_{k(t)\in\mathscr{K}} x[t]\overline{x[t-kM_\mathrm{g}]}I[n,r]\mathcal{T}_{na}(\overline{w_r[t]}w_r[t-kM_\mathrm{g}]).$$

For $k=0$, this quantity can be bounded from below as

$$M_\mathrm{g}\textstyle\sum_{t=0}^{L-1}|x[t]|^2\cdot\sum_{n,r=0}^{N-1}I[n,r]\,|\mathcal{T}_{na}w_r[t]|^2 \ge M_\mathrm{g}\|x\|^2\cdot\min_{t\in\mathbb{Z}_L}\sum_{n,r=0}^{N-1}I[n,r]\,|\mathcal{T}_{na}w_r[t]|^2,$$

with the remaining terms in $\mathscr{K}\setminus\{0\}$ handled as follows. Invoking the assumption of a real, nonnegative window w to simplify the corresponding expression, observe that the terms $\mathcal{T}_{na}(\overline{w_r[t]}w_r[t-kM_\mathrm{g}])$ are then everywhere nonnegative. The sum of the remaining terms can hence be bounded below by

$$\begin{aligned}M_\mathrm{g}\min_{t\in\mathbb{Z}_L}\sum_{n,r=0}^{N-1}I[n,r]\sum_{k(t)\in\mathscr{K}\setminus\{0\}}(\mathcal{T}_{na}(w_r[t]w_r[t-kM_\mathrm{g}]))\cdot\sum_{t''=0}^{L-1}x[t'']\overline{x[t''-kM_\mathrm{g}]}&\\ \ge -M_\mathrm{g}\|x\|^2\min_{t\in\mathbb{Z}_L}\sum_{n,r=0}^{N-1}I[n,r]\sum_{k(t)\in\mathscr{K}\setminus\{0\}}\mathcal{T}_{na}(w_r[t]w_r[t-kM_\mathrm{g}]),&\end{aligned}$$

where the second inequality follows by observing that $\sum_{t=0}^{L} x[t+s]\overline{x[t]} \geq -\|x\|^2$ for any $f \in \mathbb{C}^L$ and $s \in \mathbb{Z}_L$. Thus, we obtain the claimed results since $\sum_{\phi\in\mathscr{F}(I)} |\langle x, \phi_{m,n,r}\rangle|^2$ is bounded from below by $M_{\rm g}\|x\|^2$ times

$$\min_{t\in\mathbb{Z}_L} \sum_{n,r=0}^{N-1} I[n,r]\Big(|\mathcal{T}_{na} w_r[t]|^2 - \sum_{k\in\mathscr{K}\setminus\{0\}} \mathcal{T}_{na}(w_r[t] w_r[t-kM_{\rm g}])\Big).$$

□

Proof of Theorem 2. As our finite-dimensional setting implies the existence of an upper frame bound for any admissible I, only the existence of a lower frame bound need be established. The proof proceeds via Lemma 3 and an iterative argument.

To begin, consider the admissible selection function $I_0^{\rm g}[m,n,r] = I_0^l[m,n,r]$ induced by an N-part ordered partition, which is associated to the event that *no merging* of windows in $\mathscr{G}(w,a,b)$ occurs, and hence $\mathscr{F}(I_0^{\rm g}) = \mathscr{G}(w,a,b_{\rm g})$, with $b_{\rm g} = L/M_{\rm g}$ and $M_{\rm g} = \max(\mathrm{len}(w), M)$ according to (6.16). Lemma 1 then ensures that $\mathscr{F}(I_0^{\rm g}) = \mathscr{F}(I_0^l)$ is a frame for $\mathbb{C}^L$, with maximal lower frame bound $M_{\rm g} = \min_{t\in\mathbb{Z}} \sum_{n=0}^{N-1} |\mathcal{T}_{na} w[t]|^2 > 0$.

Next consider any admissible selection function $I_1^{\rm g}[m,n,r]$ induced by an $(N-1)$-part ordered partition, corresponding to the case that *exactly one* pair of windows $w \equiv w_0$ from the initial Gabor frame $\mathscr{G}(w,a,b)$ is merged via the superposition sum of (6.8). In this case, there exists one $n^* \in \mathbb{Z}_N$ such that $I_1^{\rm g}[0,n^*,1] = 1$, and so (6.12) implies that $\mathscr{F}(I_1^{\rm g})$ contains the elements $\{\mathcal{M}_{mb_{\rm g}}\mathcal{T}_{n^*a} w_1 : m \in \mathbb{Z}_{M_{\rm g}}\}$. Each of these elements can in turn be decomposed into the following sum:

$$\mathcal{M}_{mb_{\rm g}}\mathcal{T}_{n^*a} w_1 = \mathcal{M}_{mb_{\rm g}}\mathcal{T}_{n^*a} w_0 + \mathcal{M}_{mb_{\rm g}}\mathcal{T}_{(n^*+1)a} w_0. \tag{6.29}$$

Since (6.16) implies $M_{\rm g} \geq \mathrm{len}(w_1) = \mathrm{len}(w_0 + \mathcal{T}_a w_0)$, we obtain by (6.29) and the superadditivity property of Lemma 3:

$$\sum_{m=0}^{M_{\rm g}-1} |\langle x, \mathcal{M}_{mb_{\rm g}}\mathcal{T}_{n^*a} w_1\rangle|^2 \;\geq\; \sum_{m=0}^{M_{\rm g}-1} |\langle x, \mathcal{M}_{mb_{\rm g}}\mathcal{T}_{n^*a} w_0\rangle|^2 + \sum_{m=0}^{M_{\rm g}-1} |\langle x, \mathcal{M}_{mb_{\rm g}}\mathcal{T}_{(n^*+1)a} w_0\rangle|^2, \tag{6.30}$$

for all $x \in \mathbb{C}^L$. Next, noting that by the decomposition of (6.29), we have

$$\mathscr{F}(I_1^{\rm g}) = \mathscr{G}(w,a,b_{\rm g}) \cup \{\mathcal{M}_{mb_{\rm g}}\mathcal{T}_{n^*a} w_1\} \setminus \left(\{\mathcal{M}_{mb_{\rm g}}\mathcal{T}_{n^*a} w_0\} \cup \{\mathcal{M}_{mb_{\rm g}}\mathcal{T}_{(n^*+1)a} w_0\}\right), \tag{6.31}$$

we see that (6.30) and (6.31) together imply that for all $x \in \mathbb{C}^L$,

$$\sum_{\phi\in\mathscr{F}(I_1^{\rm g})} |\langle x, \phi_{m,n,r}\rangle|^2 \geq \sum_{\phi\in\mathscr{G}(w,a,b_{\rm g})} |\langle x, \phi_{m,n}\rangle|^2. \tag{6.32}$$

Since $\mathscr{G}(w,a,b_{\rm g}) = \mathscr{F}(I_0^{\rm g})$ is a frame, the existence of a lower frame bound for $\mathscr{F}(I_1^{\rm g})$ is guaranteed by (6.32).

Now consider the general case in which $\mathscr{F}(I^{\rm g})$ contains multiple merges. Since our construction ensures that all admissible selection functions can be obtained by iterative

partitioning in the manner above, we can always recover the inequality of (6.32) for $\mathscr{F}(I^{\mathrm{g}})$ and any b_{g} according to (6.16), by linearity of superposition and repeated application of Lemma 3.

A similar iterative argument holds for $\mathscr{F}(I_1^l)$, with b_{g} replaced by b_0 from (6.15). In place of (6.31) we obtain

$$\mathscr{F}(I_1^l) = (\mathscr{G}(w,a,b_0) \cup \{\mathcal{M}_{mb_1}\mathcal{T}_{n^*a}w_1\}) \setminus \left(\{\mathcal{M}_{mb_0}\mathcal{T}_{n^*a}w_0\} \cup \{\mathcal{M}_{mb_0}\mathcal{T}_{(n^*+1)a}w_0\}\right), \quad (6.33)$$

with $b_1 = L/M_1$ according to (6.15). Note that

$$M_1 \geq \mathrm{len}(w_1) = \mathrm{len}(w_0 + \mathcal{T}_a w_0), \qquad \text{and} \qquad M_0 \geq \mathrm{len}(w_0), \text{ and } M_1 \geq M_0;$$

thus, we may apply Lemma 3 to the latter three terms of (6.33), yielding the required result for $\mathscr{F}(I_1^l)$: for all $x \in \mathbb{C}^L$,

$$\sum_{\phi \in \mathscr{F}(I_1^l)} |\langle x, \phi_{m,n,r}\rangle|^2 \geq \sum_{\phi \in \mathscr{G}(w,a,b_0)} |\langle x, \phi_{m,n,r}\rangle|^2.$$

As above, the proof for general $\mathscr{F}(I^l)$ then follows. □

Proof of Theorem 3. Let S be the frame operator associated to $\mathscr{G}(w,a,b)$, with smallest eigenvalue A, and observe that $\min_{t\in\mathbb{Z}_L} S[t,t] \geq A$ by the Schur-Horn convexity theorem. Now, for any admissible I^{g} or I^l, the proof of Theorem 2 shows that the maximal lower frame bound of $\mathscr{F}(I^{\mathrm{g}})$ or $\mathscr{F}(I^l)$ is bounded from below by that of some $\mathscr{G}(w,a,b')$, with b' denoting respectively b_{g} or b_0. Therefore, 1) will follow if we can show the maximal lower frame bound of $\mathscr{G}(w,a,b')$ to be no less than $\min_{t\in\mathbb{Z}_L} S[t,t]$. To do so, note that Lemma 1 implies that the frame operator S' of $\mathscr{G}(w,a,b')$ is diagonal, with smallest eigenvalue $\min_{t\in\mathbb{Z}_L} S'[t,t] = (M'/M)\min_{t\in\mathbb{Z}_L} S[t,t]$; (6.15) and (6.16) then yield $M' \geq M$.

Next recall that superposition frames are induced from a Gabor frame $\mathscr{G}(w,a,b)$ by merging $n \in \mathbb{Z}_N$ neighboring window translates $\mathcal{T}_{na}w$. By the above argument, $\min_{t\in\mathbb{Z}_L} S'[t,t]$ is itself a maximal lower frame bound for the case $\mathscr{G}(w,a,b')$ attained whenever *no* window translates are merged; likewise, the merging of *all* translates yields $\mathscr{G}(w_{N-1},L,b'')$ for some unique b'', with minimal upper frame bound $\max_{t\in\mathbb{Z}_L} S''[t,t]$.

To show that these cases are in fact extremal as claimed, we appeal to the same iterative argument used to prove Theorem 2. There, the superadditivity property of Lemma 3 was invoked to show that for any admissible I^{g} or I^l, merging superposition windows *cannot decrease* the overall energy of the resultant frame coefficients. Thus, the case of $\mathscr{G}(w,a,b')$ considered above represents attainment of the *minimum* maximal lower superposition frame bound, and moreover $M' = \max(M, \mathrm{len}(w))$ via (6.15) and (6.16). Likewise, $\mathscr{G}(w_{N-1},L,b'')$ yields the *maximum* minimal upper superposition frame bound, with $M'' = \max(M, \mathrm{len}(w_{N-1}))$. Lemma 2 then establishes the bound directly:

$$\begin{aligned}\sum_{\phi \in \mathscr{G}(w_{N-1},L,b'')} |\langle x, \phi_{m,n,r}\rangle|^2 &= \textstyle\sum_{m=0}^{M''-1} |\langle x, \mathcal{M}_{mb''}w_{N-1}\rangle|^2 \\ &= M'' \textstyle\sum_{t=0}^{L-1} |x[t]|^2 |w_{N-1}[t]|^2 \leq \|x\|^2 M'' \max_{t\in\mathbb{Z}_L} |w_{N-1}[t]|^2,\end{aligned}$$

and the proof is completed by noting that as $\mathscr{G}(w,a,b)$ is assumed a Gabor frame for $\mathbb{C}^L$, the covering condition of Remark 2 implies that $|\operatorname{supp}(w_{N-1})| = \operatorname{len}(w_{N-1}) = L$, and hence $M'' = \max(M,L)$ as claimed. □

Proof of Theorem 4. First consider $I_0^{\mathrm{g}}[n,r]$, the global selection function associated to the event that no windows are merged. In this case, the generalized overlap-add constraint of (6.19) is satisfied by hypothesis, since it follows directly from (6.17) that

$$\sum_{n=0}^{N-1}\sum_{r=0}^{N-1} I_0^{\mathrm{g}}[n,r]\mathcal{T}_{na}w_r[t] = \sum_{n=0}^{N-1}\mathcal{T}_{na}w[t] = \frac{\widehat{w}[0]}{a}.$$

Now consider the global selection function $I_1^{\mathrm{g}}[n,r]$, corresponding to the case that *exactly one* pair of windows $w \equiv w_0$ is merged. In this case, there exists some $n^* \in \mathbb{Z}_N$ such that $I_1^{\mathrm{g}}[n^*,1] = 1$. Thus, the frame $\mathscr{F}(I_1^{\mathrm{g}})$ contains the element $\{\mathcal{T}_{n^*a}w_1\}$, which can be decomposed according to (6.8) as

$$\mathcal{T}_{n^*a}w_1 = \mathcal{T}_{n^*a}w_0 + \mathcal{T}_{(n^*+1)a}w_0. \tag{6.34}$$

Admissibility of I_1^{g} implies that $I_1^{\mathrm{g}}[n^*,1] = 1$, but that $I_1^{\mathrm{g}}[n^*,0] = I_1^{\mathrm{g}}[n^*{+}1,0] = 0$. Therefore, by (6.34) we have

$$\sum_{n=0}^{N-1}\sum_{r=0}^{N-1} I_1^{\mathrm{g}}[n,r]\mathcal{T}_{na}w_r[t] = \mathcal{T}_{n^*a}w_1[t] + \sum_{n\in\mathbb{Z}_N\setminus\{n^*,n^*+1\}} I_1^{\mathrm{g}}[n,0]\mathcal{T}_{na}w_0[t] = \sum_{n=0}^{N-1}\mathcal{T}_{na}w_0[t] = \frac{\widehat{w}[0]}{a},$$

and we see that (6.19) holds for $\mathscr{F}(I_1^{\mathrm{g}})$. Naturally, the selection function I^{g} may index many merged windows—not just one, as in the case of I_1^{g}. However, repeated application of the above argument shows that $\mathscr{F}(I^{\mathrm{g}})$ satisfies (6.19) for any I^{g}.

To prove Statement 2, note first that $M_{\mathrm{g}} \geq \operatorname{len}(\phi_{0,n,r})$ for each $\phi_{0,n,r} \in \mathscr{F}(I^{\mathrm{g}})$, in accordance with (6.16). Recalling that $X[m,n,r] = I^{\mathrm{g}}[m,n,r]\langle x, \phi_{m,n,r}\rangle$ by (6.13), we observe that the innermost summation of (6.20) is recognizable as an inverse discrete Fourier transform on $\mathbb{C}^{M_{\mathrm{g}}}$, with $M_{\mathrm{g}} = L/b_{\mathrm{g}}$. For fixed n and $t - na \in \mathbb{Z}_{M_{\mathrm{g}}}$, this term evaluates to $I^{\mathrm{g}}[n,r]\mathcal{T}_{na}w_r[t]x[t]$:

$$\frac{1}{M_{\mathrm{g}}}\sum_{m=0}^{M_{\mathrm{g}}-1} X[\tfrac{M_L}{M_{\mathrm{g}}}m,n,r]e^{2\pi i m b_{\mathrm{g}} t/L} \underset{t-na\in\mathbb{Z}_{M_{\mathrm{g}}}}{=} I^{\mathrm{g}}[n,r]\mathcal{T}_{na}w_r[t]x[t].$$

Perfect reconstruction then follows from the generalized OLA constraint of (6.19), as

$$\frac{a}{\widehat{w}[0]}\sum_{n=0}^{N-1}\sum_{r=0}^{N-1} I^{\mathrm{g}}[n,r]\mathcal{T}_{na}w_r[t]x[t] = \frac{a}{\widehat{w}[0]}\frac{\widehat{w}[0]}{a}x[t] = x[t].$$

For the case of a local selection function I^l, note that the generalized overlap-add constraint of (6.19) is still implied by (6.17), since the argument for the case of admissible I^{g} holds independently of the modulation structure employed. Consequently, by substituting M_r, b_r for $M_{\mathrm{g}}, b_{\mathrm{g}}$ and noting that $M_r \geq \operatorname{len}(\phi_{0,n,r})$ by (6.15) for each $\phi_{0,n,r} \in \mathscr{F}(I^l)$, we see that the result of (6.20) also holds for all $\mathscr{F}(I^l)$. □

Proof of Theorem 6. To establish the result, first note that $S_{I^{\mathrm{g}}}$ is by hypothesis diagonal, and hence by (6.21), we have

$$\widetilde{\phi}_{m,n,r}[t] = \frac{\mathcal{M}_{mb_L}\mathcal{T}_{na}w_r[t]}{M_{\mathrm{g}}\sum_{n'=0}^{N-1}\sum_{r'=0}^{N-1} I^{\mathrm{g}}[n',r']\,|\mathcal{T}_{n'a}w_{r'}[t]|^2}. \tag{6.35}$$

If all windows $\{\mathcal{T}_{na}w : n \in \mathbb{Z}_N\}$ have been merged to yield a single superposition window w_{N-1} whose modulates comprise the superposition frame $\mathscr{F}(I^{\mathrm{g}})$ of interest, then the constant overlap-add constraint of (6.17) applied to (6.35) immediately implies the result, as both its numerator and denominator yield constants, whose ratio is in turn $a/(M_{\mathrm{g}}\widehat{w}[0])$, with $M_{\mathrm{g}} = \max(L, M)$. Therefore, assume that this is not the case.

To begin, note that (6.35), together with the neighbor-overlap condition of (6.22), implies that $\mathrm{supp}(\widetilde{\phi}_{m,n,r}) \subseteq \mathcal{T}_{na}(L_r \cup C_r \cup R_r)$. Since these sets are mutually disjoint, we proceed by showing that (6.35) agrees with

$$\begin{cases} \widetilde{\phi}_{m,n,0}[t] & \text{if } t \in \mathcal{T}_{na}L_r, \\ \frac{1}{M_{\mathrm{g}}}\frac{a}{\widehat{w}[0]}e^{2\pi i m b_L t/L} & \text{if } t \in \mathcal{T}_{na}C_r, \\ \widetilde{\phi}_{m,n+r+1,0}[t] & \text{if } t \in \mathcal{T}_{na}R_r, \end{cases} \tag{6.36}$$

where

$$\widetilde{\phi}_{m,n,0}[t] \triangleq \frac{\mathcal{M}_{mb_L}\mathcal{T}_{na}w[t]}{M_{\mathrm{g}}\sum_{n'=0}^{N-1}|w[t-n'a]|^2}.$$

We now proceed to show that (6.35) evaluates to (6.36). First, we have that the numerator of (6.35) evaluates on $\mathcal{T}_{na}C_r$ to

$$\begin{aligned} \mathcal{M}_{mb_L}\mathcal{T}_{na}w_r[t]\,\mathbb{I}_{\mathcal{T}_{na}C_r}[t] &= \mathcal{M}_{mb_L}\left(\textstyle\sum_{n'=0}^{r}\mathcal{T}_{(n+n')a}w[t]\right)\mathbb{I}_{\mathcal{T}_{na}C_r}[t] \\ &= \mathcal{M}_{mb_L}\left(\textstyle\sum_{n'=0}^{N-1}\mathcal{T}_{(n+n')a}w[t]\right)\mathbb{I}_{\mathcal{T}_{na}C_r}[t] \\ &= e^{2\pi i m b_L t/L}\left(\tfrac{\widehat{w}[0]}{a}\right)\mathbb{I}_{\mathcal{T}_{na}C_r}[t], \end{aligned}$$

where the second equality follows from the neighbor overlap condition of (6.22), and the third by the overlap-add constraint of (6.17) together with the definition of the set C_r. The denominator of (6.35) evaluates on this same set $\mathcal{T}_{na}C_r$ to

$$\begin{aligned} &\left(M_{\mathrm{g}}\textstyle\sum_{n'=0}^{N-1}\sum_{r'=0}^{N-1} I^{\mathrm{g}}[n',r']\,|\mathcal{T}_{n'a}w_{r'}[t]|^2\right)\mathbb{I}_{\mathcal{T}_{na}C_r}[t] \\ &\quad = \left(M_{\mathrm{g}}|\mathcal{T}_{na}w_r[t]|^2\right)\mathbb{I}_{\mathcal{T}_{na}C_r}[t] = \left(M_{\mathrm{g}}\tfrac{\widehat{w}^2[0]}{a^2}\right)\mathbb{I}_{\mathcal{T}_{na}C_r}[t], \end{aligned}$$

with the first equality following from the fact that no windows other than $\mathcal{T}_{na}w_r$ are supported on $\mathcal{T}_{na}C_r$, and the second from (6.17). Hence we have equality of (6.35) and (6.36) on $\mathcal{T}_{na}C_r$.

Applying next the neighbor-overlap condition of (6.22) and the definition of L_r, we observe that the corresponding numerator term of (6.35) evaluates to

$$\mathcal{M}_{mb_L}\mathcal{T}_{na}w_r[t]\,\mathbb{I}_{\mathcal{T}_{na}L_r}[t] = \mathcal{M}_{mb_L}\left(\textstyle\sum_{n'=0}^{r}\mathcal{T}_{(n+n')a}w[t]\right)\mathbb{I}_{\mathcal{T}_{na}L_r}[t] = \mathcal{M}_{mb_L}\mathcal{T}_{na}w[t]\,\mathbb{I}_{\mathcal{T}_{na}L_r}[t].$$

Evaluating the denominator of (6.35) on $\mathcal{T}_{na}L_r$ yields $(M_{\mathrm{g}}\sum_{n'=0}^{N-1}\sum_{r'=0}^{N-1} I^{\mathrm{g}}[n',r']|\mathcal{T}_{n'a}w_{r'}[t]|^2)\mathbb{I}_{\mathcal{T}_{na}L_r}[t]$, which may be split into three parts according to index n, including the term $\mathcal{T}_{na}w_r[t]$ as:

$$\begin{aligned}
&M_{\mathrm{g}}\left(\textstyle\sum_{r'=0}^{N-1}\sum_{n'=0}^{N-1} I^{\mathrm{g}}[n',r']|\mathcal{T}_{n'a}w_{r'}[t]|^2\right)\mathbb{I}_{\mathcal{T}_{na}L_r}[t]\\
&= M_{\mathrm{g}}\left(\textstyle\sum_{r'=0}^{N-1}\sum_{n'=0}^{n-1} I^{\mathrm{g}}[n',r']|\mathcal{T}_{n'a}w_{r'}[t]|^2 + |\mathcal{T}_{na}w_r[t]|^2\right.\\
&\qquad\left. + \textstyle\sum_{r'=0}^{N-1}\sum_{n'=n+1}^{N-1} I^{\mathrm{g}}[n',r']|\mathcal{T}_{n'a}w_{r'}[t]|^2\right)\mathbb{I}_{\mathcal{T}_{na}L_r}[t].
\end{aligned}$$

The middle expression of $|\mathcal{T}_{na}w_r[t]|^2$ stems from the fact that $I^{\mathrm{g}}[n,r]=1$ whenever $\widetilde{\phi}_{\cdot,n,r}\in\widetilde{\mathscr{F}(I^{\mathrm{g}})}$, and correspondingly $I^{\mathrm{g}}[n,r']=0$ whenever $r'\neq r$, since I^{g} is an admissible selection function. Moreover, coupled with the assumed neighbor-overlap condition of (6.22), this same property implies that exactly *two* superposition windows are supported on $\mathcal{T}_{na}L_r$, one of which is the superposition window $\mathcal{T}_{na}w_r[t]$ isolated in the sum above.

By the superposition construction, it must be the case that the portion of $\mathcal{T}_{na}w_r[t]$ supported on $\mathcal{T}_{na}L_r$ takes the form $\mathcal{T}_{na}w[t]$, whereas the portion of the remaining window supported on $\mathcal{T}_{na}L_r$ takes the form $\mathcal{T}_{(n-1)a}w[t]$. Thus

$$\begin{aligned}
&M_{\mathrm{g}}\left(\textstyle\sum_{r'=0}^{N-1}\sum_{n'=0}^{N-1} I^{\mathrm{g}}[n',r']|\mathcal{T}_{n'a}w_{r'}[t]|^2\right)\mathbb{I}_{\mathcal{T}_{na}L_r}[t]\\
&\quad= M_{\mathrm{g}}\left(|\mathcal{T}_{(n-1)a}w[t]|^2+|\mathcal{T}_{na}w[t]|^2\right)\mathbb{I}_{\mathcal{T}_{na}L_r}[t] = M_{\mathrm{g}}\textstyle\sum_{n'=0}^{N-1}|w[t-n'a]|^2\mathbb{I}_{\mathcal{T}_{na}L_r}[t],
\end{aligned}\tag{6.37}$$

and (6.35) is seen to equal (6.36) on the set $\mathcal{T}_{na}L_r$. The case of $\mathcal{T}_{na}R_r$ proceeds by an identical argument, thereby confirming that (6.35) agrees separately on $\mathcal{T}_{na}(L_r,C_r,R_r)$ with the quantities of (6.36), as claimed.

Finally, to complete the proof, observe that the cyclic group setting of $\mathbb{Z}_L$ implies the relation $\widetilde{\phi}_{m,n,0}[t]=\mathcal{M}_{mb_L}\mathcal{T}_{na}\widetilde{\phi}_{0,0,0}[t]$, since for any integer n, $\sum_{n'=0}^{N-1}|\mathcal{T}_{n'a}w[t]|^2=\sum_{n'=0}^{N-1}|\mathcal{T}_{(n'+n)a}w[t]|^2$. Applying this relation to (6.36), we obtain the theorem as stated. Note that for *local* I^l, the sequence of equalities analogous to those in (6.37) requires the local frequency lattice, which may not be known prior to observing the signal. □

Proof of Theorem 7. The dyadic superposition frame $\mathscr{F}(I^d)$ represents the set of elements selected from $G^d_{\cup r}(b_{\mathrm{g}})$. Here we denote its the canonical dual by $\widetilde{\mathscr{F}_{I^d}}(G^d_{\cup r})$, reflecting the explicit dependence on $G^d_{\cup r}$, and likewise define $\mathscr{F}_{I^d}(\cup\widetilde{G^d_r})$, the set of elements selected by I^d from $\cup_{r\in\mathscr{R}}\widetilde{G^d_r}$. To establish the result, we must verify the claimed equality

$$\widetilde{\mathscr{F}_{I^d}}(G^d_{\cup r}) = \mathscr{F}_{I^d}(\cup\widetilde{G^d_r}).\tag{6.38}$$

We proceed to establish the equality of (6.38) elementwise, noting first that the number of modulates of some $T_{na}w_r$ in $\widetilde{\mathscr{F}_{I^d}}(G^d_{\cup r})$ is given by $M_{\mathrm{g}}=L/b_{\mathrm{g}}$ and, by construction, is equal to the number of modulates of the same shifted window in each $G^d_r(b_{\mathrm{g}})=\mathscr{G}(w_r,a,b_{\mathrm{g}})$. We therefore fix $m\in\mathbb{Z}_{M_{\mathrm{g}}}$ and $r\in\mathscr{R}$ for the remainder of the proof.

Since the dyadic superposition frame operator S_{I^d} is diagonal, the superposition Walnut formula of (6.21) implies that we may write each $\widetilde{\phi}_{m,n,r}\in\widetilde{\mathscr{F}_{I^d}}(G^d_{\cup r})$ as

$$\widetilde{\phi}_{m,n,r}[t]=\frac{\mathcal{M}_{mb_L}\mathcal{T}_{na}w_r[t]}{M_{\mathrm{g}}\sum_{n'=0}^{N-1}\sum_{r'=0}^{N-1}I^d[n',r']|\mathcal{T}_{n'a}w_{r'}[t]|^2}.\tag{6.39}$$

Each element $\widehat{\phi}_{m,n'',r} \in \widetilde{G_r^d}$ can be likewise written as:

$$\widehat{\phi}_{m,n'',r}[t] = \frac{\mathcal{M}_{mb_L}\mathcal{T}_{n''a_r}w_r[t]}{M_{\mathrm{g}}\sum_{n^*=0}^{N_r-1}|\mathcal{T}_{n^*a_r}w_r[t]|^2}, \tag{6.40}$$

with $a_r = a(r+1)$, $N_r = L/a_r$, and $0 \leq n'' < N_r$, in accordance with Definition 16. Note that we ordinarily index modulations of Gabor frame elements by $mb_{\mathrm{g}}, m \in \mathbb{Z}_{M_{\mathrm{g}}}$, but in (6.40) we adopt the indexing scheme mb_L for appropriate $m \in \mathbb{Z}_{M_L}$, in order to facilitate its direct comparison to (6.39).

In order to establish the equality of sets in (6.38), we need to show that if $\widetilde{\phi}_{m,n,r} \in \widetilde{\mathscr{F}_{I^d}}(G^d_{\cup r})$, then the expressions in (6.39) and (6.40) are equivalent; i.e.,

$$I^d[m,n,r] = 1 \;\Rightarrow\; \widetilde{\phi}_{m,n,r} = \widehat{\phi}_{m,n,r}. \tag{6.41}$$

To establish the implication of (6.41), we first show that the condition $I^d[m,n,r] = 1$ implies that the numerators of (6.39) and (6.40) agree. Since m and r are fixed, this means that for all n such that $I^d[m,n,r] = 1$, there must exist an $0 \leq n'' < N_r$ satisfying

$$\mathcal{M}_{mb_L}\mathcal{T}_{na}w_r = \mathcal{M}_{mb_L}\mathcal{T}_{n''a_r}w_r. \tag{6.42}$$

Equality in (6.42) is achieved when $na = n''a_r = n''(r+1)a$, which clearly holds if $r+1$ divides n. But since $I^d[m,n,r]$ selects elements from $G_r^d = \mathscr{G}(w_r, a(r+1), b_{\mathrm{g}})$, then $r+1$ divides n by construction, and (6.42) follows.

The argument for agreement of the denominators is more delicate, because it is *not* true that for all $t \in \mathbb{Z}_L$,

$$\textstyle\sum_{n'=0}^{N-1}\sum_{r'=0}^{N-1} I^d[n',r']|\mathcal{T}_{n'a}w_{r'}[t]|^2 = \sum_{n^*=0}^{N_r-1}|\mathcal{T}_{n^*a_r}w_r[t]|^2. \tag{6.43}$$

Instead, we show that (6.43) holds for all $t \in \operatorname{supp}(\mathcal{T}_{na}w_r)$—which, together with (6.42), is sufficient to establish (6.41), and consequently our claimed result.

Let $S_{n,r} \triangleq \operatorname{supp}(\mathcal{T}_{na}w_r)$, with $\mathbb{I}_{S_{n,r}}[t]$ the corresponding indicator function. Using the same arguments as in the penultimate portion of the proof of Theorem 6, observe that the left-hand side of (6.43) can be decomposed as follows:

$$\begin{aligned}
&\left(\textstyle\sum_{r'=0}^{N-1}\sum_{n'=0}^{N-1} I^d[n',r']|\mathcal{T}_{n'a}w_{r'}[t]|^2\right)\mathbb{I}_{S_{n,r}}[t] \\
&= \left(\textstyle\sum_{r'=0}^{N-1}\sum_{n'=0}^{n-1} I^d[n',r']|\mathcal{T}_{n'a}w_{r'}[t]|^2 + \mathcal{T}_{na}w_r^2[t] + \sum_{r'=0}^{N-1}\sum_{n'=n+1}^{N-1} I^d[n',r']|\mathcal{T}_{n'a}w_{r'}[t]|^2\right)\mathbb{I}_{S_{n,r}}[t] \\
&= \left(|\mathcal{T}_{(n-1)a}w[t]|^2 + |\mathcal{T}_{na}w_r[t]|^2 + |\mathcal{T}_{(n+r+1)a}w[t]|^2\right)\mathbb{I}_{S_{n,r}}[t].
\end{aligned}$$

Applying the neighbor-overlap requirement of (6.22) to the right-hand side of (6.43) yields

$$\left(\textstyle\sum_{n^*=0}^{N_r-1}|\mathcal{T}_{n^*a_r}w_r[t]|^2\right)\mathbb{I}_{S_{n,r}}[t] = \left(|\mathcal{T}_{(n-1)a}w[t]|^2 + |\mathcal{T}_{na}w_r[t]|^2 + |\mathcal{T}_{(n+r+1)a}w[t]|^2\right)\mathbb{I}_{S_{n,r}}[t],$$

thus establishing the equality of (6.43), and hence the result. □

Chapter 7

Source Modeling

In the last four chapters we developed a number of statistical methods for non-stationary signal analysis and applied them to modeling temporal variation of the vocal tract in the context of speech processing. Formant tracking, time-varying linear prediction, and adaptive short-time Fourier analysis all capture the time-varying covariance structure of the speech waveform, which, according to the source-filter model, arises from the time-varying characteristics of vocal tract. Here, we present a number of contributions to the complementary problem of modeling the temporal variation in the source waveform during voicing.

It is well known that the classical autoregressive model fails to take into account the quasi-periodic nature of the source waveform typical of voiced speech. Here, we augment the traditional linear prediction framework to incorporates an estimate of the source waveform through the use of flexible basis function expansions via nonparametric wavelet regression. In addition to studying properties of the resultant estimators when the functional expansion is known *a priori*, we propose signal-adaptive methods of selecting an appropriate subspace for modeling the source waveform.

We show that the resultant family of models may be applied to a number of problems in speech analysis ranging from vocal tract and source-harmonics-to-noise ratio estimation to inverse filtering and voicing detection. For instance, we demonstrate that our framework not only allows for improved estimation of the autoregressive coefficients parameterizing the vocal tract in a manner that is robust to pitch variation, but also precludes the need for nonlinear optimization procedures typically required in glottal waveform estimation. Our approach is illustrated through a variety of experiments using both synthesized and real speech waveforms.

7.1 Introduction

Autoregressive (AR) modeling forms the basis of many successful speech analysis algorithms to date. Arising naturally from the source-filter view of speech production, AR models represent the transfer function of the vocal tract (filter), which is excited by the glottal airflow (source). However, algorithms commonly used to estimate parameters of the underlying AR model, including the covariance and autocorrelation methods, assume a

white Gaussian noise model for the source waveform (see e.g., Section 2.4.6 and [3]), which is inconsistent with its quasi-periodic nature during voicing and leads to biased estimates of the AR coefficients [220].

The primary way of accounting for this model mismatch has been to incorporate a model of the glottal airflow volume velocity (or its derivative) in the linear prediction framework and estimate the source waveform jointly with the vocal tract parameters. Parametric modeling of the source is the most popular approach, with the Rosenberg-Klatt [13,18] and LF [14–17] models used by [41] and [43, 221], respectively. An alternative is the modeling approach of [20–22] whereby the glottal volume velocity $u_g[n]$ over a *single pitch period* is modeled using a polynomial expansion. These models allow for simpler estimators than those for the Rosenberg or LF models, and are more flexible than the associated rigid parametric forms. However, estimates of the pitch period and glottal closure timings are still required—leading to difficult nonlinear optimization problems.

To address these issues, we propose an extension of the AR(p) model, whereby a time-varying mean $\mu[n]$ is introduced to explicitly capture the quasi-periodic nature of the glottal flow resulting in a so-called ARX (AR with eXogenous input) model [91]. Following on from the work of [222] and [223], we model $\mu[n]$ by way of a basis function expansion with a particular focus on wavelet-based expansions. A key aspect of our approach is to model the glottal flow over the length of *multiple* pitch periods, in contrast to fitting a single-pitch-period model and convolving it with a pulse train, as is typical in the epoch extraction literature [43]. As a result, our approach not only allows for *linear* estimation of vocal tract parameters, but also does not require knowledge of speech parameters, such as pitch period and glottal closure timings, and is robust to pitch variation.

We introduce the proposed model in Section 7.2, derive the corresponding linear maximum likelihood (ML) and maximum a posteriori (MAP) estimators in Section 7.3, and study properties of the ML estimator and the associated Cramer-Rao bounds in Section 7.4. Next, in Section 7.5, we propose a number of online, data-dependent approaches to selecting the functional basis for representing the time-varying mean $\mu[n]$, and address the question of hypothesis testing for the presence of $\mu[n]$ in Section 7.6. We first evaluate the proposed techniques in Section 7.7.1, using synthetic speech data, and show that the resultant model produces more accurate estimates of the vocal tract transfer function than classical linear prediction. Experiments with recorded speech, described in Section 7.7.2, suggest that the model can better explain variability in voiced speech and can, potentially, be used to estimate the relative amounts of voicing and aspiration energy in each waveform, which is of interest in the clinical setting and many other applications. We conclude with a brief discussion in Section 7.9.

7.2 Model Formulation

In Section 2.4.6, we have already discussed that the classical AR(p) model for speech:

$$\text{AR}(p): \quad x[n] = \sum_{i=1}^{p} a_i x[n-i] + \sigma w[n], \tag{7.1}$$

where $w[n]$ is a white Gaussian sequence scaled by a gain $\sigma > 0$, cannot adequately capture the quasi-periodic time-varying nature of the glottal airflow. Instead, we propose to adapt to the speech setting a more *flexible* autoregressive model with *exogenous* or external variables:

$$\text{ARX}(p): \quad x[n] = \sum_{i=1}^{p} a_i x[n-i] + \mu[n] + \sigma w[n]. \tag{7.2}$$

The discrete-time difference equation of (7.2) comprises an all-pole LTI system driven by a Gaussian process with constant variance σ^2 and a *time-varying mean* $\mu[n]$, which we use to capture the quasi-periodic nature of the glottal airflow. Specifically, the sequence $w[n]$ is a zero-mean white Gaussian process with unit variance scaled by σ, and the mean $\mu[n]$ is defined according to:

$$\mu[n] \triangleq \sum_{k=1}^{r} \beta_k g_k[n], \tag{7.3}$$

where the r functions $\{g_1[n], g_2[n], \ldots, g_r[n]\}$ are specified before any data are observed. Certain waveforms motivate letting σ^2 depend on time [24], though we don't pursue this extension here due to the complexity of the resultant estimation procedures.

It is easy to see that the ARX(p) model specified by (7.2) and (7.3) reduces to the classical AR(p) process when $\beta_k = 0$ for all $1 \leq k \leq r$. Indeed, our method of modeling the temporal variation in the source waveform, resembles the approach taken in modeling the TVAR coefficient temporal trajectories in Chapters (4) and (5). We discuss ways of selecting the functions in the expansion of (7.3) approaches in Section 7.5 below.

We note that the application of the ARX(p) model specified by (7.2) and (7.3) to source estimation is new, but the model itself has already appeared in other literatures. To wit, its study was initiated in [25, 224–226] both from a theoretical perspective (i.e., estimation and asymptotic analysis) and as it applied to econometrics. It has since been used in control theory [91] and signal processing [222], but to our knowledge has not yet been applied in the context of speech processing. A number of authors in the speech processing literature (e.g., [227]) have referred to (7.2) as an ARX(p) model, but instead of the functional representation of (7.3) they assume that $\mu[n]$ is parameterized according to one of the standard source waveform models such as the Liljencrantz-Fant model described, earlier, in Section 2.2.

7.3 Parameter Estimation

Here we consider the question of parameter estimation for the ARX(p) model. We first derive conditional maximum likelihood estimators in Section 7.3.1, and then approach estimation from the Bayesian perspective in Section 7.3.2.

7.3.1 Maximum Likelihood

The ARX(p) model of (7.2) and (7.3) is specified by a vector of AR coefficients $\boldsymbol{a} \triangleq \begin{pmatrix} a_1 & a_2 & \cdots & a_p \end{pmatrix}^T \in \mathbb{R}^{p \times 1}$, a vector of expansion coefficients $\boldsymbol{\beta} = \begin{pmatrix} \beta_1 & \beta_2 & \cdots & \beta_r \end{pmatrix}^T \in \mathbb{R}^{r \times 1}$, and the noise variance σ^2. In order to derive the maximum likelihood (ML) estimator of

the model parameters $\boldsymbol{\theta} \triangleq (\boldsymbol{a}^T, \boldsymbol{\beta}^T, \sigma^2)^T$, we partition N samples of the process according to:

$$\boldsymbol{x} = (\boldsymbol{x}_p \,|\, \boldsymbol{x}_{N-p})^T \triangleq \big(x[0] \cdots x[p-1] \;\mid\; x[p] \cdots x[N-1]\big)^T,$$

and note that the joint probability density function of the data given $\boldsymbol{\theta}$ is given by:

$$p(\boldsymbol{x}\,;\boldsymbol{\theta}) = p(\boldsymbol{x}_{N-p} \,|\, \boldsymbol{x}_p\,;\boldsymbol{\theta}) p(\boldsymbol{x_p}\,;\boldsymbol{\theta}). \tag{7.4}$$

When $N \gg p$, the second term can be safely ignored and the unconditional likelihood of (7.4) can be approximated by conditional likelihood $p(\boldsymbol{x}_{N-p} \,|\, \boldsymbol{x}_p\,;\boldsymbol{\theta})$. Gaussianity of $w[n]$ implies that this conditional likelihood is given by:

$$p(\boldsymbol{x}_{N-p} \,|\, \boldsymbol{x}_p\,;\boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^{(N-p)/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{n=p}^{N-1} e^2[n]\right), \tag{7.5}$$

where

$$e[n] \triangleq x[n] - \left(\sum_{i=1}^{p} a_i x[n-i] + \sum_{k=1}^{r} \beta_k g_k[n]\right)$$

is the prediction error. Clearly, the maximizer of (7.5) is also the least-squares solution to the following linear regression problem:

$$\boldsymbol{x}_{N-p} = \boldsymbol{X}\boldsymbol{a} + \boldsymbol{G}\boldsymbol{\beta} + \sigma\boldsymbol{w}_{N-p} = (\boldsymbol{X} \,|\, \boldsymbol{G}) \begin{pmatrix}\boldsymbol{a}\\ \boldsymbol{\beta}\end{pmatrix} + \sigma\boldsymbol{w}_{N-p}, \tag{7.6}$$

where $\boldsymbol{w}_{N-p} \triangleq \big(w[p] \;\; w[p+1] \;\; \cdots \;\; w[N-1]\big)^T \in \mathbb{R}^{(N-p)\times 1}$, and the matrices $\boldsymbol{X} \in \mathbb{R}^{(N-p)\times p}$ and $\boldsymbol{G} \in \mathbb{R}^{(N-p)\times r}$ are defined by:

$$\boldsymbol{X} \triangleq \begin{pmatrix} x[p-1] & \cdots & x[0] \\ x[p] & \cdots & x[1] \\ \vdots & \ddots & \vdots \\ x[N-2] & \cdots & x[N-p+1] \end{pmatrix} \quad \text{and} \quad \boldsymbol{G} \triangleq \begin{pmatrix} g_1[p] & \cdots & g_r[p] \\ g_1[p+1] & \cdots & g_r[p+1] \\ \vdots & \ddots & \vdots \\ g_1[N-1] & \cdots & g_r[N-1] \end{pmatrix}. \tag{7.7}$$

It is easy to see from (7.6) that the least-squares estimate of $\boldsymbol{a}$ and $\boldsymbol{\beta}$ is given by the pseudo-inverse of $(\boldsymbol{X} \,|\, \boldsymbol{G})$ as in:

$$\begin{pmatrix}\widehat{\boldsymbol{a}}\\ \widehat{\boldsymbol{\beta}}\end{pmatrix} = (\boldsymbol{X} \,|\, \boldsymbol{G})^{\#}\, \boldsymbol{x}_{N-p} = \begin{pmatrix}\boldsymbol{X}^T\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{G} \\ \boldsymbol{G}^T\boldsymbol{X} & \boldsymbol{G}^T\boldsymbol{G}\end{pmatrix}^{-1} \begin{pmatrix}\boldsymbol{X}^T \\ \boldsymbol{G}^T\end{pmatrix} \boldsymbol{x}_{N-p}, \tag{7.8}$$

where $\boldsymbol{M}^{\#} \triangleq (\boldsymbol{M}^T\boldsymbol{M})^{-1}\boldsymbol{M}^T$ denotes the pseudoinverse of a matrix $\boldsymbol{M}$. Note that if all the basis functions g_k were everywhere equal to 0, then the model of (7.2) reduces to the classical AR model, and the least-squares estimator of (7.8) reduces to the covariance method of linear prediction described earlier in Section 2.4.1. The submatrices $\boldsymbol{X}^T\boldsymbol{X}$, $\boldsymbol{X}^T\boldsymbol{G}$ and $\boldsymbol{G}^T\boldsymbol{G}$ provide estimates of the autocorrelation of the data, the correlation between the data and basis functions and among the basis functions, respectively. If the functions $\{g_1, \ldots, g_r\}$ are orthonormal on $\mathbb{R}^{(N-p)\times 1}$, then $\boldsymbol{G}^T\boldsymbol{G} = \boldsymbol{I}$, which simplifies the block structure in (7.8).

Next, consider (7.8) from a geometrical perspective. Let $\text{span}(\boldsymbol{M})$ denote the column span of $\boldsymbol{M}$, $P_{\boldsymbol{M}} \triangleq \boldsymbol{M}\boldsymbol{M}^{\#}$ the orthogonal projection onto $\text{span}(\boldsymbol{M})$, and $P_{\boldsymbol{M}}^{\perp} \triangleq \boldsymbol{I} - P_{\boldsymbol{M}}$ the orthogonal projection onto $\text{span}(\boldsymbol{M})^{\perp}$—the orthogonal complement of $\text{span}(\boldsymbol{M})$. Then the estimators of (7.8) can be written according to:

$$\widehat{\boldsymbol{a}} = \left(P_{\boldsymbol{G}}^{\perp}\boldsymbol{X}\right)^{\#} \boldsymbol{x}_{N-p}, \tag{7.9}$$

$$\widehat{\boldsymbol{\beta}} = \left(P_{\boldsymbol{X}}^{\perp}\boldsymbol{G}\right)^{\#} \boldsymbol{x}_{N-p}. \tag{7.10}$$

An easy way to see this is to take derivatives of $\log(p(\boldsymbol{x}_{N-p} \,|\, \boldsymbol{x}_p\,; \boldsymbol{\theta}))$ with respect to $\boldsymbol{a}$ and $\boldsymbol{\beta}$, and set the result equal to zero to obtain the following set of normal equations that $\widehat{\boldsymbol{a}}$ and $\widehat{\boldsymbol{\beta}}$ must satisfy (see e.g., Speckman [228] for a similar development in the context of kernel regression):

$$\left(\boldsymbol{X}^T\boldsymbol{X}\right)\widehat{\boldsymbol{a}} = \boldsymbol{X}^T\left(\boldsymbol{x}_{N-p} - \boldsymbol{G}\widehat{\boldsymbol{\beta}}\right) \tag{7.11}$$

$$\left(\boldsymbol{G}^T\boldsymbol{G}\right)\widehat{\boldsymbol{\beta}} = \boldsymbol{G}^T\left(\boldsymbol{x}_{N-p} - \boldsymbol{X}\widehat{\boldsymbol{a}}\right). \tag{7.12}$$

Multiplying both sides of (7.12) by $\boldsymbol{G}\left(\boldsymbol{G}^T\boldsymbol{G}\right)^{-1}$ and substituting the resultant expression into (7.11) leads to

$$\begin{aligned}\left(\boldsymbol{X}^T\boldsymbol{X}\right)\widehat{\boldsymbol{a}} &= \boldsymbol{X}^T\left(\boldsymbol{x}_{N-p} - \boldsymbol{P}_{\boldsymbol{G}}\left(\boldsymbol{x}_{N-p} - \boldsymbol{X}\widehat{\boldsymbol{a}}\right)\right)\\ &= \boldsymbol{X}^T\boldsymbol{x}_{N-p} - \boldsymbol{X}^T\boldsymbol{P}_{\boldsymbol{G}}\boldsymbol{x}_{N-p} + \boldsymbol{X}^T\boldsymbol{P}_{\boldsymbol{G}}\boldsymbol{X}\widehat{\boldsymbol{a}}.\end{aligned}$$

Rearranging and factoring out common terms leads to:

$$\left(\boldsymbol{X}^T\left(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{G}}\right)\boldsymbol{X}\right)\widehat{\boldsymbol{a}} = \boldsymbol{X}^T\left(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{G}}\right)\boldsymbol{x}_{N-p},$$

which immediately yields (7.9) since

$$\begin{aligned}\widehat{\boldsymbol{a}} &= \left(\boldsymbol{X}^T\left(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{G}}\right)\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\left(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{G}}\right)\boldsymbol{x}_{N-p}\\ &= \left(\boldsymbol{X}^T\boldsymbol{P}_{\boldsymbol{G}}^{\perp}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{P}_{\boldsymbol{G}}^{\perp}\boldsymbol{x}_{N-p} = \left(\boldsymbol{P}_{\boldsymbol{G}}^{\perp}\boldsymbol{X}\right)^{\#}\boldsymbol{x}_{N-p}.\end{aligned}$$

The formula of (7.10) may be obtained similarly by substituting (7.11) into (7.12). An alternative approach to deriving (7.9) and (7.10) consists of applying the matrix inversion lemma and the Woodbury identity to the least-squares estimator of (7.8).

Therefore, by (7.9) and (7.10), we can reconstruct $\boldsymbol{x}_{N-p}$ in the $(p+r)$-dimensional subspace of $\mathbb{R}^{(N-p)\times 1}$ corresponding to the column space of the matrix $(\boldsymbol{X} \,|\, \boldsymbol{G})$ according to:

$$\widehat{\boldsymbol{x}}_{N-p} = \boldsymbol{X}\widehat{\boldsymbol{a}} + \boldsymbol{G}\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}\left(P_{\boldsymbol{G}}^{\perp}\boldsymbol{X}\right)^{\#} + \boldsymbol{G}\left(P_{\boldsymbol{X}}^{\perp}\boldsymbol{G}\right)^{\#}\right)\boldsymbol{x}_{N-p}.$$

Note that $\text{span}(\boldsymbol{X})$ and $\text{span}(\boldsymbol{G})$ are not necessarily orthogonal—the angle between these subspaces depends on the selected basis functions as well as the data. Of interest to note in our setting, is that the first set of regressors (columns of $\boldsymbol{X}$) depends on the data, whereas the second set of regressors (columns of $\boldsymbol{G}$) does not. In addition, note that if $r = N - p$ and

the columns of $\boldsymbol{G}$ are a basis for $\mathbb{R}^{N-p}$, then $\text{span}(\boldsymbol{X}) \subseteq \text{span}(\boldsymbol{G}) = \mathbb{R}^{N-p}$ and $\boldsymbol{a}$ cannot be estimated via (7.9) since $P_{\boldsymbol{G}}^{\perp}\boldsymbol{X} = \boldsymbol{0}$.

Finally, we note that after estimates of the regression coefficients $\boldsymbol{a}$ and $\boldsymbol{\beta}$ are obtained via (7.8), or equivalently via (7.9) and (7.10), the conditional maximum likelihood estimate of σ^2 is given, as a function of the prediction error by:

$$\widehat{\sigma^2} = \frac{1}{N-p}\|\boldsymbol{x}_{N-p} - \boldsymbol{X}\widehat{\boldsymbol{a}} - \boldsymbol{G}\widehat{\boldsymbol{\beta}}\|^2. \tag{7.13}$$

7.3.2 Bayesian Approach

In order to obtain additional insight into the ARX(p) model of (7.2) and the associated conditional maximum likelihood estimators of Section 7.3.1 above, we recast (7.2) in a hierarchical Bayesian framework [229], derive the associated estimators and compare them with the conditional maximum likelihood estimators of (7.8).

The conditional likelihood of the parameters $\boldsymbol{\theta}$ given the observation sequence follows from (7.5) as:

$$p(\boldsymbol{x}_{N-p} \,|\, \boldsymbol{x}_p \,;\, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{X}\boldsymbol{a} + \boldsymbol{G}\boldsymbol{\beta}, \sigma^2),$$

and suggests a standard hierarchical Bayesian model. We detail a fully-conjugate model which allows for analytically integrating out nuisance parameters—we may wish to average over the time-varying mean coefficients $\boldsymbol{\beta}$ if estimation of the all-pole parameters $\boldsymbol{a}$ is of primary concern, and vice-versa.

We specify prior distributions for all model parameters, including the vectors of autoregressive coefficients $\boldsymbol{a}$, time-varying mean coefficients $\boldsymbol{\beta}$, and the noise variance σ^2. In particular, we assume that $\boldsymbol{a} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{\Lambda}_a)$ and $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{\Lambda}_\beta)$, where $\boldsymbol{\Lambda}_a \in \mathbb{R}^{p\times p}$ and $\boldsymbol{\Lambda}_\beta \in \mathbb{R}^{r\times r}$ are positive-definite matrices. A standard inverse-Gamma conjugate prior $\mathcal{IG}(a_0, b_0)$ is adopted for the noise variance σ^2. Because of sensitivity to hyperparameters, we follow standard practice and adopt a Gamma hyperprior on b_0 for robustness by assuming that $b_0 \sim \mathcal{G}(\kappa, \eta)$.

The posterior probability distribution of the model parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2, b_0)$, conditioned on the observations $\boldsymbol{x}$ and the fixed model parameters $\boldsymbol{\psi} \triangleq (\boldsymbol{\Lambda}_a, \boldsymbol{\Lambda}_\beta, a_0, \kappa, \eta)$ is therefore given by:

$$\begin{aligned}
&p\left(\boldsymbol{a}, \boldsymbol{\beta}, \sigma^2, b_0 \,|\, \boldsymbol{x} \,;\, \boldsymbol{\psi}\right) \propto p\left(\boldsymbol{x} \,|\, \boldsymbol{a}, \boldsymbol{\beta}, \sigma^2\right) p(\boldsymbol{a} \,|\, \sigma^2; \boldsymbol{\Lambda}_a) p(\boldsymbol{\beta} \,|\, \sigma^2; \boldsymbol{\Lambda}_\beta) p(\sigma^2 \,|\, b_0; a_0) p(b_0; \kappa, \eta) \\
&= \sigma^{-N} \exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{x}_{N-p} - \boldsymbol{X}\boldsymbol{a} - \boldsymbol{G}\boldsymbol{\beta}\right)^T\left(\boldsymbol{x}_{N-p} - \boldsymbol{X}\boldsymbol{a} - \boldsymbol{G}\boldsymbol{\beta}\right)\right) \\
&\cdot \sigma^{-(p+r)} \left|\boldsymbol{\Lambda}_a\right|^{-1/2} \left|\boldsymbol{\Lambda}_b\right|^{-1/2} \exp\left(-\frac{1}{2\sigma^2}\boldsymbol{a}^T\boldsymbol{\Lambda}_a^{-1}\boldsymbol{a}\right) \exp\left(-\frac{1}{2\sigma^2}\boldsymbol{\beta}^T\boldsymbol{\Lambda}_\beta^{-1}\boldsymbol{\beta}\right) \\
&\cdot \sigma^{-2(a_0-1)} e^{-b_0/\sigma^2} b_0^{\kappa-1} e^{-b_0/\eta}.
\end{aligned} \tag{7.14}$$

The fully-conjugate structure of posterior distribution of (7.14) enables us to in-

tegrate out the expansion coefficients $\boldsymbol{\beta}$. The resulting marginal density takes the form:

$$p\left(\boldsymbol{a}, \sigma^2, b_0 \,|\, \boldsymbol{x}\,;\, \boldsymbol{\psi}\right) = \sigma^{-(N+p+r+a_0-1)} e^{-b_0/\sigma^2} b_0^{\kappa-1} e^{-b_0/\eta} \left|\boldsymbol{\Lambda}_a\right|^{-1/2} \left|\boldsymbol{\Lambda}_b\right|^{-1/2} \exp\left(-\frac{1}{2\sigma^2}\boldsymbol{a}^T\boldsymbol{\Lambda}_a^{-1}\boldsymbol{a}\right)$$
$$\cdot \left|\boldsymbol{\Gamma}_1\right|^{-1/2} \exp\left(\frac{1}{2\sigma^2}\boldsymbol{\Gamma}_2^T\boldsymbol{\Gamma}_1^{-1}\boldsymbol{\Gamma}_2\right) \exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{x}_{N-p} - \boldsymbol{X}\boldsymbol{a}\right)^T\left(\boldsymbol{x}_{N-p} - \boldsymbol{X}\boldsymbol{a}\right)\right),$$

where the matrices $\boldsymbol{\Gamma_1} \in \mathbb{R}^{r\times r}$ and $\boldsymbol{\Gamma_2} \in \mathbb{R}^{1\times r}$ are defined according to:

$$\boldsymbol{\Gamma}_1 \triangleq \boldsymbol{G}^T\boldsymbol{G} + \boldsymbol{\Lambda}_b^{-1} \qquad \text{and} \qquad \boldsymbol{\Gamma}_2 \triangleq \left(\boldsymbol{x}_{N-p} - \boldsymbol{X}\boldsymbol{a}\right)^T\boldsymbol{G}. \tag{7.15}$$

Collecting all the terms with the variable $\boldsymbol{a}$ yields:

$$p\left(\boldsymbol{a}, \sigma^2, b_0 \,|\, \boldsymbol{x}\,;\, \boldsymbol{\psi}\right) = \sigma^{-(N+p+r+a_0-1)} e^{-b_0/\sigma^2} b_0^{\kappa-1} e^{-b_0/\eta} \left|\boldsymbol{\Lambda}_a\right|^{-1/2} \left|\boldsymbol{\Lambda}_b\right|^{-1/2} \left|\boldsymbol{\Gamma}_1\right|^{-1/2}$$
$$\cdot \exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{a}^T\boldsymbol{\Lambda}_a^{-1}\boldsymbol{a} - \boldsymbol{\Gamma}_2^T\boldsymbol{\Gamma}_1^{-1}\boldsymbol{\Gamma}_2 + \left(\boldsymbol{x}_{N-p} - \boldsymbol{X}\boldsymbol{a}\right)^T\left(\boldsymbol{x}_{N-p} - \boldsymbol{X}\boldsymbol{a}\right)\right)\right),$$

and substituting (7.15) into the exponential term leads to:

$$p\left(\boldsymbol{a}, \sigma^2, b_0 \,|\, \boldsymbol{x}\,;\, \boldsymbol{\psi}\right) = \sigma^{-(N+p+r+a_0-1)} e^{-b_0/\sigma^2} b_0^{\kappa-1} e^{-b_0/\eta} \left|\boldsymbol{\Lambda}_a\right|^{-1/2} \left|\boldsymbol{\Lambda}_b\right|^{-1/2} \left|\boldsymbol{\Gamma}_1\right|^{-1/2}$$
$$\cdot \exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{a}^T\boldsymbol{\Lambda}_a^{-1}\boldsymbol{a} + \left(\boldsymbol{x}_{N-p} - \boldsymbol{X}\boldsymbol{a}\right)^T\left(\boldsymbol{I} - \boldsymbol{G}^T\left(\boldsymbol{G}^T\boldsymbol{G} + \boldsymbol{\Lambda}_b^{-1}\right)\boldsymbol{G}^T\right)^{-1}\left(\boldsymbol{x}_{N-p} - \boldsymbol{X}\boldsymbol{a}\right)\right)\right). \tag{7.16}$$

The marginal density of (7.16) may be used to calculate marginal minimum-mean-squared error (MMSE) and maximum a posteriori (MAP) estimates of $\boldsymbol{a}$, σ^2 and b_0.

There is an important connection between the marginal density of (7.16) and the maximum likelihood estimators of Section 7.3.1. If instead of putting a Normal prior on $\boldsymbol{a}$, we were to use an improper prior proportional to a constant everywhere on its support (effectively assuming that $\boldsymbol{a}$ was unknown, but nonrandom), then the marginal density of (7.16) reduces to:

$$p\left(\boldsymbol{a}, \sigma^2, b_0 \,|\, \boldsymbol{x}\,;\, \boldsymbol{\psi}\right) = \sigma^{-(N+r+a_0-1)} e^{-b_0/\sigma^2} b_0^{\kappa-1} e^{-b_0/\eta} \left|\boldsymbol{\Lambda}_b\right|^{-1/2} \left|\boldsymbol{\Gamma}_1\right|^{-1/2}$$
$$\cdot \exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{x}_{N-p} - \boldsymbol{X}\boldsymbol{a}\right)^T\left(\boldsymbol{I} - \boldsymbol{G}^T\left(\boldsymbol{G}^T\boldsymbol{G} + \boldsymbol{\Lambda}_b^{-1}\right)\boldsymbol{G}^T\right)^{-1}\left(\boldsymbol{x}_{N-p} - \boldsymbol{X}\boldsymbol{a}\right)\right). \tag{7.17}$$

Maximizing (7.17) with respect to $\boldsymbol{a}$ yields the following marginal MAP estimator:

$$\widehat{\boldsymbol{a}}_{\text{MMAP}} = \left(\boldsymbol{X}^T\left(\boldsymbol{I} - \boldsymbol{G}^T\left(\boldsymbol{G}^T\boldsymbol{G} + \boldsymbol{\Lambda}_b^{-1}\right)\boldsymbol{G}^T\right)\boldsymbol{X}\right)^{-1}\left(\boldsymbol{I} - \boldsymbol{G}^T\left(\boldsymbol{G}^T\boldsymbol{G} + \boldsymbol{\Lambda}_b^{-1}\right)\boldsymbol{G}^T\right)^{-1}\boldsymbol{X}^T\boldsymbol{x}_{N-p}. \tag{7.18}$$

If, in addition, we were to allow $\boldsymbol{\Lambda}_b^{-1}$ to approach $\boldsymbol{0}$ (corresponding to a diffuse, or in the limit noninformative, prior on $\boldsymbol{\beta}$) in the marginal MAP estimator of (7.18), then we obtain:

$$\left(\boldsymbol{I} - \boldsymbol{G}^T\left(\boldsymbol{G}^T\boldsymbol{G} + \boldsymbol{\Lambda}_b^{-1}\right)^{-1}\boldsymbol{G}^T\right) \to \left(\boldsymbol{I} - \boldsymbol{G}^T\left(\boldsymbol{G}^T\boldsymbol{G}\right)^{-1}\boldsymbol{G}^T\right) = P_{\boldsymbol{G}}^{\perp}. \tag{7.19}$$

Consequently, the marginal MAP estimator of (7.18) converges to the ML estimator of (7.9) since substituting (7.19) into (7.18) yields

$$\widehat{\boldsymbol{a}}_{\mathrm{MMAP}} = \left(\boldsymbol{X}^T P_{\boldsymbol{G}}^{\perp} \boldsymbol{X}\right)^{-1} P_{\boldsymbol{G}}^{\perp} \boldsymbol{X}^T \boldsymbol{x}_{N-p} = \left(\boldsymbol{X}^T P_{\boldsymbol{G}}^{\perp} P_{\boldsymbol{G}}^{\perp} \boldsymbol{X}\right)^{-1} P_{\boldsymbol{G}}^{\perp} \boldsymbol{X}^T \boldsymbol{x}_{N-p} = \left(P_{\boldsymbol{G}}^{\perp} \boldsymbol{X}\right)^{\#} \boldsymbol{x}_{N-p}.$$

Therefore the marginal MAP estimator of (7.18) is, in essence, a regularized version of the ML estimator of (7.9). A similar relationship can be established for the ML estimator of $\boldsymbol{\beta}$ given by (7.10) and an appropriate marginal MAP estimator derived in analogy to the above treatment.

7.4 Asymptotic Analysis

In this section, we study the asymptotic behavior of the maximum likelihood estimators of (7.9) and (7.10) presented in Section 7.3.1. First in Section 7.4.1, we collect existing results [25] and state the regularity conditions required for consistency and asymptotic normality. Next, we derive the associated Cramer-Rao bounds in Section 7.4.2, in part following [222] and show how they yield important insights into the case when the AR spectrum overlaps with that of the time-varying mean $\mu[n]$. The formulas for the Fisher information matrix derived here will be also needed in our development of hypothesis testing for the ARX(p) model in Section 7.6.

7.4.1 Consistency and Asymptotic Normality

The asymptotic properties of the maximum likelihood estimator of (7.8) for the ARX(p) model specified by (7.2) and (7.3) were studied by [224, 225] in a more general multivariate setting. Following [25], we first rewrite the difference equation of (7.2) as a first-order *vector* autoregressive process according to:

$$\boldsymbol{x}_n + \boldsymbol{A}\boldsymbol{x}_{n-1} + \boldsymbol{B}\boldsymbol{g}_n = \boldsymbol{w}_n, \tag{7.20}$$

where $\boldsymbol{x}_n \triangleq \begin{pmatrix} x[n] & x[n-1] & \cdots & x[n-p+1] \end{pmatrix}^T \in \mathbb{R}^{p\times 1}$, $\boldsymbol{g}_n \triangleq \begin{pmatrix} g_1[n] & g_2[n] & \cdots & g_r[n] \end{pmatrix}^T \in \mathbb{R}^{r\times 1}$, $\boldsymbol{w}_n \triangleq \begin{pmatrix} w[n] & 0 & \cdots & 0 \end{pmatrix}^T \in \mathbb{R}^{p\times 1}$, and the matrices $\boldsymbol{A} \in \mathbb{R}^{p\times p}$ and $\boldsymbol{B} \in \mathbb{R}^{r\times r}$ are defined by:

$$\boldsymbol{A} = \begin{pmatrix} a_1 & a_2 & \cdots & a_p \\ -1 & 0 & \cdots & 0 \\ 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}, \qquad \boldsymbol{B} = \begin{pmatrix} \beta_1 & \beta_2 & \cdots & \beta_r \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}.$$

Next, note that the estimator of (7.8) is equivalent to:

$$\begin{pmatrix} \displaystyle\sum_{n=p}^{N-1} \boldsymbol{x}_{n-1}\boldsymbol{x}_{n-1}^T & \displaystyle\sum_{n=p}^{N-1} \boldsymbol{x}_{n-1}\boldsymbol{g}_n^T \\ \displaystyle\sum_{n=p}^{N-1} \boldsymbol{g}_n\boldsymbol{x}_{n-1}^T & \displaystyle\sum_{n=p}^{N-1} \boldsymbol{g}_n\boldsymbol{g}_n^T \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{a}} \\ \widehat{\boldsymbol{\beta}} \end{pmatrix} = \begin{pmatrix} \displaystyle\sum_{n=p}^{N-1} \boldsymbol{x}_{n-1}\boldsymbol{x}_n^T \\ \displaystyle\sum_{n=p}^{N-1} \boldsymbol{g}_n\boldsymbol{x}_n^T \end{pmatrix}. \tag{7.21}$$

The form of equation (7.21) motivates considering the existence of the following limits:

$$\boldsymbol{H} = \lim_{N\to\infty} \frac{1}{N}\sum_{n=p}^{N} \boldsymbol{v}_n \boldsymbol{v}_n^T, \quad \boldsymbol{L} = \lim_{N\to\infty} \frac{1}{N}\sum_{n=p}^{N} \boldsymbol{v}_{n-1} \boldsymbol{g}_n^T, \quad \text{and} \quad \boldsymbol{M} = \lim_{N\to\infty} \frac{1}{N}\sum_{n=p}^{N} \boldsymbol{g}_n \boldsymbol{g}_n^T, \tag{7.22}$$

where the vector $\boldsymbol{v}_n \in \mathbb{R}^{p\times 1}$ is defined according to:

$$\boldsymbol{v}_n \triangleq \sum_{m=p}^{n-1} (-\boldsymbol{A})^m \begin{pmatrix} \boldsymbol{\beta}^T \boldsymbol{g}_{n-m} \\ \boldsymbol{0}_{p-1\times 1} \end{pmatrix}.$$

In addition, we need to define the following matrix

$$\boldsymbol{F} = \sum_{m=p}^{\infty} \boldsymbol{A}^m \boldsymbol{\Sigma} \left(\boldsymbol{A}^T\right)^m,$$

where $\boldsymbol{\Sigma} \triangleq \sigma^2 \boldsymbol{e}_1 \boldsymbol{e}_1^T$ and $\boldsymbol{e}_1 \triangleq \begin{pmatrix} 1 & \boldsymbol{0}_{1\times p-1} \end{pmatrix}$.

We are now ready to state the precise conditions that imply the consistency and asymptotic normality of (7.21) (or equivalently (7.8)). The proofs can be found in [25].

Theorem 8 (Consistency [25])**.** *If $x[n]$ is defined by (7.2) and (7.3), with $w[n]$ a zero-mean white Gaussian sequence with $\mathbb{E}\big(w^2[n]\big) = \sigma^2 > 0$, if the characteristic roots of $-\boldsymbol{A}$ are less than 1 in absolute value, if all the limits in (7.22) exist, if $\boldsymbol{g}_n^T \boldsymbol{g}_n < C, n = 1, 2, \ldots$ for some constant C, if $\boldsymbol{F}$ is positive definite and $\boldsymbol{M}$ is nonsingular, then as $N \to \infty$ we have:*

$$\lim_{N\to\infty} \widehat{\boldsymbol{a}} = \boldsymbol{a} \qquad \textit{and} \qquad \lim_{N\to\infty} \widehat{\boldsymbol{\beta}} = \boldsymbol{\beta},$$

with the limits converging in probability.

Theorem 9 (Asymptotic Normality [25])**.** *If $x[n]$ is defined by (7.2) and (7.3), with $w[n]$ a zero-mean white Gaussian sequence with $\mathbb{E}\big(w^2[n]\big) = \sigma^2 > 0$, if the characteristic roots of $-\boldsymbol{A}$ are less than 1 in absolute value, if $\boldsymbol{g}_n^T \boldsymbol{g}_n < C, n = 1, 2, \ldots$ for some constant C, if all the limits in (7.22) exist, if $\boldsymbol{g}_n^T \boldsymbol{g}_n < C, n = 1, 2, \ldots$ for some constant C, if $\boldsymbol{F}$ and $\boldsymbol{M}$ are positive definite, then $\sqrt{N}\left((\widehat{\boldsymbol{a}} - \boldsymbol{a}) \;\; (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right)^T$ has a limiting Normal distribution with mean $\boldsymbol{0}_{(p+r)\times 1}$ and covariance matrix:*

$$\sigma^2 \begin{pmatrix} \boldsymbol{F} + \boldsymbol{H} & \boldsymbol{L} \\ \boldsymbol{L}^T & \boldsymbol{M} \end{pmatrix}^{-1}. \tag{7.23}$$

Note that requiring the magnitude of the characteristic roots of $-\boldsymbol{A}$ to be less than 1 in absolute value is equivalent to requiring the poles of the associated autoregressive process to lie inside the unit circle.

the condition for

7.4.2 Cramér-Rao Bounds

In order to calculate the Cramér-Rao lower bound (CRLB), we need to compute entries of the Fisher information matrix (FIM) $I_{\boldsymbol{\theta\theta}}$ defined by:

$$I_{\boldsymbol{\theta\theta}} \triangleq -\mathbb{E}\left(\frac{\partial^2 \ln p(\boldsymbol{x}_{N-p} \,|\, \boldsymbol{x}_p \,;\, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}\right) \tag{7.24}$$

The necessary first derivatives are given by:

$$\frac{\partial \ln p(\boldsymbol{x}_{N-p} \,|\, \boldsymbol{x}_p \,;\, \boldsymbol{\theta})}{\partial a_i} = -\frac{1}{\sigma^2}\sum_{n=p}^{N-1} e[n]x[n-i]$$

$$\frac{\partial \ln p(\boldsymbol{x}_{N-p} \,|\, \boldsymbol{x}_p \,;\, \boldsymbol{\theta})}{\partial \beta_k} = -\frac{1}{\sigma^2}\sum_{n=p}^{N-1} e[n]g_k[n]$$

$$\frac{\partial \ln p(\boldsymbol{x}_{N-p} \,|\, \boldsymbol{x}_p \,;\, \boldsymbol{\theta})}{\partial \sigma^2} = -\frac{N-p}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{n=p}^{N-1} e^2[n]$$

and the second derivatives are given by:

$$\frac{\partial^2 \ln p(\boldsymbol{x}_{N-p} \,|\, \boldsymbol{x}_p \,;\, \boldsymbol{\theta})}{\partial a_i \partial a_{i'}} = -\frac{1}{\sigma^2}\sum_{n=p}^{N-1} x[n-i]x[n-i']$$

$$\frac{\partial^2 \ln p(\boldsymbol{x}_{N-p} \,|\, \boldsymbol{x}_p \,;\, \boldsymbol{\theta})}{\partial a_i \partial \beta_k} = -\frac{1}{\sigma^2}\sum_{n=p}^{N-1} g_k[n]x[n-i]$$

$$\frac{\partial^2 \ln p(\boldsymbol{x}_{N-p} \,|\, \boldsymbol{x}_p \,;\, \boldsymbol{\theta})}{\partial a_i \partial \sigma^2} = \frac{1}{\sigma^4}\sum_{n=p}^{N-1} e[n]x[n-i]$$

$$\frac{\partial^2 \ln p(\boldsymbol{x}_{N-p} \,|\, \boldsymbol{x}_p \,;\, \boldsymbol{\theta})}{\partial \beta_k \partial \beta_{k'}} = -\frac{1}{\sigma^2}\sum_{n=p}^{N-1} g_k[n]g_{k'}[n]$$

$$\frac{\partial^2 \ln p(\boldsymbol{x}_{N-p} \,|\, \boldsymbol{x}_p \,;\, \boldsymbol{\theta})}{\partial \beta_k \partial \sigma^2} = \frac{1}{\sigma^4}\sum_{n=p}^{N-1} e[n]g_k[n]$$

$$\frac{\partial^2 \ln p(\boldsymbol{x}_{N-p} \,|\, \boldsymbol{x}_p \,;\, \boldsymbol{\theta})}{\partial \sigma^2 \partial \sigma^2} = \frac{N-p}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{n=p}^{N-1} e^2[n].$$

Prior to computing the necessary expectations, we will need to calculate $m[n] \triangleq \mathbb{E}(x[n])$ since (7.2) is not a zero-mean process. Let us suppose that the first p means $m[0], m[1], \ldots, m[p-1]$ are known. (This assumption is consistent with our *conditional* maximum likelihood setting). Using (7.2), it is then easy to see that $m[n]$ can be recursively computed for any $p \leq n \leq N-1$ according to:

$$m[n] = \sum_{i=1}^{p} a_i m[n-i] + \mu[n]. \tag{7.25}$$

In other words, the time-varying mean of $x[n]$ is just the time-varying mean of the innovations sequence colored by an all-pole filter. Now, we are ready to compute the necessary expectations.

$$\begin{aligned}
-\mathbb{E}\left(\frac{\partial^2 \ln p(\boldsymbol{x}_{N-p} \,|\, \boldsymbol{x}_p\,;\,\boldsymbol{\theta})}{\partial a_i \partial a_{i'}}\right) &= \frac{1}{\sigma^2}\sum_{n=p}^{N-1}\mathbb{E}\left(x[n-i]x[n-i']\right) \\
&= \frac{1}{\sigma^2}\sum_{n=p}^{N-1}\mathbb{E}\left((x[n-i]-m[n-i])(x[n-i']-m[n-i'])\right) + \sum_{n=p}^{N-1} m[n-i]m[n-i'] \\
&= (N-p)r_{xx}[i-i'] + \sum_{n=p}^{N-1} m[n-i]m[n-i']
\end{aligned}$$

$$-\mathbb{E}\left(\frac{\partial^2 \ln p(\boldsymbol{x}_{N-p} \,|\, \boldsymbol{x}_p\,;\,\boldsymbol{\theta})}{\partial a_i \partial \beta_k}\right) = \frac{1}{\sigma^2}\sum_{n=p}^{N-1} g_k[n]\mathbb{E}\left(x[n-i]\right) = \frac{1}{\sigma^2}\sum_{n=p}^{N-1} g_k[n]m[n-i]$$

$$-\mathbb{E}\left(\frac{\partial^2 \ln p(\boldsymbol{x}_{N-p} \,|\, \boldsymbol{x}_p\,;\,\boldsymbol{\theta})}{\partial a_i \partial \sigma^2}\right) = -\frac{1}{\sigma^4}\sum_{n=p}^{N-1}\mathbb{E}\left(e[n]x[n-i]\right) = 0$$

$$-\mathbb{E}\left(\frac{\partial^2 \ln p(\boldsymbol{x}_{N-p} \,|\, \boldsymbol{x}_p\,;\,\boldsymbol{\theta})}{\partial \beta_k \partial \beta_{k'}}\right) = \frac{1}{\sigma^2}\sum_{n=p}^{N-1} g_k[n]g_{k'}[n]$$

$$-\mathbb{E}\left(\frac{\partial^2 \ln p(\boldsymbol{x}_{N-p} \,|\, \boldsymbol{x}_p\,;\,\boldsymbol{\theta})}{\partial \beta_k \partial \sigma^2}\right) = \frac{1}{\sigma^4}\sum_{n=p}^{N-1}\mathbb{E}(e[n])g_k[n] = 0$$

$$-\mathbb{E}\left(\frac{\partial^2 \ln p(\boldsymbol{x}_{N-p} \,|\, \boldsymbol{x}_p\,;\,\boldsymbol{\theta})}{\partial \sigma^2 \partial \sigma^2}\right) = -\frac{N-p}{2\sigma^4} + \frac{1}{\sigma^6}\sum_{n=p}^{N-1}\mathbb{E}(e^2[n]) = -\frac{N-p}{2\sigma^4} + \frac{1}{\sigma^6}\sum_{n=p}^{N-1}\sigma^2 = \frac{N-p}{2\sigma^4}.$$

We may simplify the above expressions through the introduction of the vectors $\boldsymbol{m}_i \in \mathbb{R}^{(N-p)\times 1}$ for $1 \leq i \leq p$ defined via

$$\boldsymbol{m}_i \triangleq \begin{pmatrix} m[p-i] & m[p+1-i] & \cdots & m[N-1-i] \end{pmatrix}^T,$$

and the matrix $\boldsymbol{M} \in \mathbb{R}^{(N-p)\times p}$ defined by

$$\boldsymbol{M} \triangleq \begin{pmatrix} \boldsymbol{m}_p & \boldsymbol{m}_{p-1} & \cdots & \boldsymbol{m}_1 \end{pmatrix}.$$

Then the Fisher information matrix takes the following form:

$$\boldsymbol{I}(\boldsymbol{\theta}) \triangleq \begin{pmatrix} \boldsymbol{I}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}(\boldsymbol{\theta}) & \boldsymbol{I}_{\boldsymbol{\alpha}\boldsymbol{\beta}}(\boldsymbol{\theta}) & \boldsymbol{I}_{\boldsymbol{\alpha}\sigma^2}(\boldsymbol{\theta}) \\ \boldsymbol{I}_{\boldsymbol{\beta}\boldsymbol{\alpha}}(\boldsymbol{\theta}) & \boldsymbol{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\theta}) & \boldsymbol{I}_{\boldsymbol{\beta}\sigma^2}(\boldsymbol{\theta}) \\ \boldsymbol{I}_{\sigma^2\boldsymbol{\alpha}}(\boldsymbol{\theta}) & \boldsymbol{I}_{\sigma^2\boldsymbol{\beta}}(\boldsymbol{\theta}) & \boldsymbol{I}_{\sigma^2\sigma^2}(\boldsymbol{\theta}) \end{pmatrix} = \frac{1}{\sigma^2}\begin{pmatrix} \widetilde{\boldsymbol{R}} + \boldsymbol{M}^T\boldsymbol{M} & \boldsymbol{M}^T\boldsymbol{G} & \boldsymbol{0}_{p\times 1} \\ \boldsymbol{G}^T\boldsymbol{M} & \boldsymbol{G}^T\boldsymbol{G} & \boldsymbol{0}_{r\times 1} \\ \boldsymbol{0}_{1\times p} & \boldsymbol{0}_{1\times r} & (N-p)/2\sigma^2 \end{pmatrix}, \tag{7.26}$$

where $\widetilde{\boldsymbol{R}} = (N-p)\boldsymbol{R}$ and $\boldsymbol{R}$ is the symmetric Toeplitz matrix constructed from the autocorrelation sequence $(r_{xx}[0], r_{xx}[1], \ldots, r_{xx}[p-1])$.

The CRLB is therefore given by

$$\mathrm{Cov}(\widehat{\boldsymbol{\theta}}) \geq \boldsymbol{I}^{-1}(\boldsymbol{\theta}),$$

where a matrix inequality $\boldsymbol{A} \geq \boldsymbol{B}$ means that $\boldsymbol{A} - \boldsymbol{B}$ is positive semidefinite. Using the matrix inversion lemma, we may extract lower bounds on subblocks of $\boldsymbol{\theta}$ corresponding to $\boldsymbol{a}$, $\boldsymbol{\beta}$ and σ^2. The lower bound on $\boldsymbol{a}$ is given by:

$$\mathrm{Cov}(\widehat{\boldsymbol{a}}) \geq \frac{1}{\sigma^2}\left(\widetilde{\boldsymbol{R}} + \boldsymbol{M}^T\boldsymbol{M} - \boldsymbol{M}^T\boldsymbol{G}\left(\boldsymbol{G}^T\boldsymbol{G}\right)^{-1}\boldsymbol{G}^T\boldsymbol{M}\right)^{-1} = \frac{1}{\sigma^2}\left(\widetilde{\boldsymbol{R}} + \boldsymbol{M}^T P_{\boldsymbol{G}}^{\perp}\boldsymbol{M}\right)^{-1}. \tag{7.27}$$

The lower bound on $\boldsymbol{\beta}$ is given by:

$$\begin{aligned}
\mathrm{Cov}(\widehat{\boldsymbol{\beta}}) &\geq \frac{1}{\sigma^2}\left(\boldsymbol{G}^T\boldsymbol{G} - \boldsymbol{G}^T\boldsymbol{M}\left(\widetilde{\boldsymbol{R}} + \boldsymbol{M}^T\boldsymbol{M}\right)^{-1}\boldsymbol{M}^T\boldsymbol{G}\right)^{-1} \\
&= \frac{1}{\sigma^2}\left(\boldsymbol{G}^T\left(\boldsymbol{I}_{N-p} - \boldsymbol{M}\left(\widetilde{\boldsymbol{R}} + \boldsymbol{M}^T\boldsymbol{M}\right)^{-1}\boldsymbol{M}^T\right)\boldsymbol{G}\right)^{-1} \\
&= \frac{1}{\sigma^2}\left(\boldsymbol{G}^T\left(\boldsymbol{I}_{N-p} - \boldsymbol{M}\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{M}^T\left(\boldsymbol{I}_{N-p} + \boldsymbol{M}\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{M}^T\right)\right)^{-1}\boldsymbol{G}\right)^{-1} \\
&= \frac{1}{\sigma^2}\left(\boldsymbol{G}^T\left(\boldsymbol{I}_{N-p} + \boldsymbol{M}\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{M}^T\right)^{-1}\boldsymbol{G}\right)^{-1}.
\end{aligned} \tag{7.28}$$

Here the second-to-last equality follows from the Searle identity $\left(\boldsymbol{A} + \boldsymbol{B}\boldsymbol{B}^T\right)^{-1}\boldsymbol{B} = \boldsymbol{A}^{-1}\boldsymbol{B}\left(\boldsymbol{I} + \boldsymbol{B}^T\boldsymbol{A}\boldsymbol{B}\right)^{-1}$. Finally, the CRLB on $\widehat{\sigma^2}$ is given by:

$$\mathrm{Var}(\widehat{\sigma^2}) \geq \frac{2\sigma^4}{N-p}.$$

Now consider (7.27) and (7.28) when entries of the matrix of means $\boldsymbol{M}$ are large. This happens when the spectral energy of the columns of $\boldsymbol{G}$ (and consequently the mean signal $\mu[n] = \boldsymbol{G}\boldsymbol{\beta}$) and the all-pole spectrum associated to $\boldsymbol{a}$ concurrently take large values over the same set of frequencies. In this case, it is evident from (7.27) and (7.28) that the trace of the CRLB for the AR coefficients $\boldsymbol{a}$ *decreases*, while the trace of the CRLB of the expansion coefficients $\boldsymbol{\beta}$ *increases*.

Since at first glance this seems to be a somewhat counterintuitive result, we illustrate it through two examples based on the following ARX(2) process:

$$x[n] = a_1 x[n-1] + a_2 x[n-2] + \beta_1 \cos[2\pi n\omega/f_s] + w[n]. \tag{7.29}$$

In the first experiment, we set a_1 and a_2 so that the second-order resonator associated to the induced all-pole model has a center frequency of 2 kHz with $f_s = 16$ kHz and consider the CRLB for $\widehat{a}$ and $\widehat{\beta}_1$ as a function of ω as it is varied from 0 to 8 kHz. Clearly, the largest overlap in the frequency content of the exogenous sinusoid and the autoregressive filter should occur when ω takes on values in the neighborhood of 2 kHz. This is confirmed in the results shown in the left three panels of Figure 7.1 that demonstrate the expected behavior of the CRLB for the two AR coefficients $\boldsymbol{a}$ and the expansion coefficient β_1.

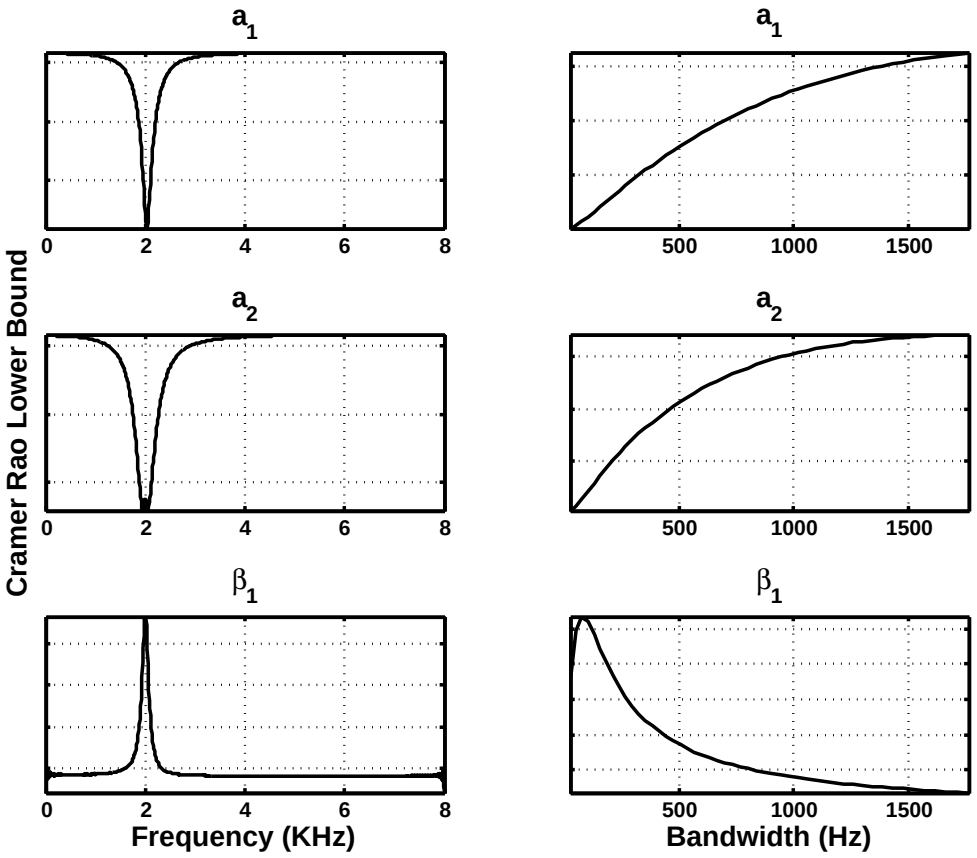


Figure 7.1: Overlap of spectral energy in the exogenous variable and AR filter for an ARX(2) model and the Cramer-Rao lower bound. Left: CRLB is shown for a_1, a_2, β_1 as a function of the frequency of the exogenous sinusoid in (7.29). Right: CRLB is shown for a_1, a_2, β_1 as a function of bandwidth of the all-pole filter associated to (7.29) with center frequency fixed and equal to the frequency of the exogenous sinusoid.

In the second experiment, we fixed the center frequency of the exogenous sinusoid to $\omega = 2$ kHz and varied a_1 and a_2 so that the center frequency of the associated second-order digital was *also* 2 kHz, while its bandwidth varied from 0 to 1500 Hz. Note that as the bandwidth approaches 0, the frequency response of the all-pole filter sharpens around 2 kHz. In fact, when the bandwidth is equal to 0 the AR(2) model converges to the exogenous sinusoid exactly. However, since in this extremal case the covariance matrix $\boldsymbol{R}_{xx}$ is singular as the process becomes deterministic, the CRLB cannot be computed. The results, shown in the right three panels of Figure 7.1, confirm that as the bandwidth decreases the entries of the matrix $\boldsymbol{M}$ decrease, resulting in the predicted increases and decreases in value of the CRLB for β_1 and $\boldsymbol{a}$, respectively.

There are two takeaways from this analysis. First, the coefficients of exogenous waveform $\mu[n]$ in an ARX model (e.g., the sinusoid in (7.29)) will be more difficult to estimate if the spectral content of $\mu[n]$ significantly overlaps with that of the autoregressive filter. This is not too surprising from the point of view of deconvolution. For instance,

in the context of speech processing, with $\mu[n]$ representing a glottal source waveform, this result implies that it is harder to estimate the spectral characteristics of the source if they match those of the filter, as is often the case with high-pitched speakers [3].

Second, the decrease in the CRLB for the AR coefficients is a well-known effect in the system identification literature. Indeed, much effort in the controls community has been focused on the design of probing signals for inferring parameters of unknown linear systems [230]. In fact, one of the classic criteria for probing signal design to the maximize the trace of the Fisher information matrix [230] associated to the system parameters. Thus, in our ARX(2) example the best probing signal among sinusoids is the one whose frequency matches that of the AR(2)-derived second-order resonator.

An interesting speech example that directly reflects the ARX(2) model of (7.29) is that of opera singers who attempt to align the first formant of their vocal tract with the fundamental frequency of sustained vowels [3]. The above discussion implies that while it is harder to identify the source waveform of an opera singer it is easier to identify the parameters of the vocal tract.

7.5 Subspace Selection

In this section we turn to the question of model selection and discuss two distinct approaches to selecting basis functions for the expansion of (7.3)—their choice determines the r-dimensional subspace $\boldsymbol{R}^{N-p}$ in which the estimated time-varying mean $\mu[n]$ will lie. The first method, discussed in Section 7.5.1, relies on selecting r functions prior to observing any data. On the other hand, the second approach enables online, signal-adaptive function selection as described in Section 7.5.2.

7.5.1 Offline Approaches

The simplest approach is to select r elements from some basis for $\mathbb{R}^{N-p}$. A variety of natural choices for the basis functions exist; for instance, a small number of low-order polynomials are used in [222]. In the context of speech processing, a very natural choice might be to use Fourier series as $\mu[n]$ is often modeled as periodic, especially for sustained vowels. However, all $N-p$ sinusoids cannot be used because columns of $\boldsymbol{X}$ would then lie in the column span of $\boldsymbol{G}$, which implies that the maximum likelihood estimator of (7.9) cannot be used since $P^{\perp}_{\boldsymbol{G}}\boldsymbol{X} = \boldsymbol{0}$. Thus, since only $r << N-p$ functions may be used, a method for selecting which frequencies to model would be necessary, but hard to design since source waveforms are not necessarily sparse in a Fourier series expansion (i.e., many terms may be necessary for an accurate representation). Moreover, an accurate estimate of the pitch period would be required, and the approach would not be robust to natural pitch variation and irregular phonation (e.g., pitch jitter) due to the *global* support of the Fourier series (i.e., over multiple pitch periods).

Some of these issues may be avoided by using *time-localized* functions such as wavelets. Specifically, assuming that $N-p$ is a power of 2, we model $\mu[n]$ according to the

following dyadic wavelet expansion:

$$\mu[n] = \sum_{j=0}^{L} \sum_{k=0}^{2^j-1} \phi_{j,k} 2^{-j/2} \phi[2^{-j}n - k] + \sum_{j=L+1}^{M} \sum_{k=0}^{2^j-1} \psi_{j,k} 2^{-j/2} \psi[2^{-j}n - k], \tag{7.30}$$

where $\psi[n]$ is the mother wavelet, $\phi[n]$ is the associated scaling function, and L is an integer such that 2^L denotes the number of scaling functions in the wavelet decomposition. Due to the *local* nature of the representation of (7.30), the energy of the signal is concentrated in fewer coefficients overall; wavelet representations are known to be sparse for broad classes of signals, which is one of the primary reasons for their popularity [231]. In practice, we have found that the r lowest-resolution wavelets, which includes all scaling functions when $r \geq 2^L$, yields reasonable performance as demonstrated in Section 7.7.

Another interesting offline approach would be to learn a functional basis from previously-recorded speech signals using inverse filtering together with principal components analysis as in the recent work of [232].

7.5.2 Online Approaches

We now describe a number of approaches to estimating the autoregressive coefficients $\boldsymbol{a}$ and the time-varying mean $\mu[n]$ in the ARX(p) model of (7.2) that do not rely on an *apriori* specification of the basis functions to be used in the expansion of (7.3). Instead, given a time series of observations, an appropriate subset of a wavelet basis for $\mathbb{R}^{N-p}$ is to be selected *online* in a signal-adaptive manner. The proposed methods fall into a broader class of algorithms based on linear sparse signal representations, see e.g., the recent textbook [231] and references therein. A related approach for estimating partially linear models in the context of functional magnetic resonance imaging is presented in [233], though the employed model has no autoregressive component.

To begin, we assume that $N - p$ is a power of two and let columns of the matrix $\boldsymbol{W} \in \mathbb{R}^{(N-p)\times(N-p)}$ contain the elements of some wavelet basis for $\mathbb{R}^{N-p}$. Thus, $\boldsymbol{W}$ is the orthonormal discrete wavelet transform (DWT) matrix so that $\boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{I}$. The precise choices of the underlying mother wavelet and the assumed boundary conditions do not need to be made at this point. Setting $\boldsymbol{G} = \boldsymbol{W}$ in (7.6) we obtain:

$$\boldsymbol{x}_{N-p} = \boldsymbol{X}\boldsymbol{a} + \boldsymbol{W}\boldsymbol{\beta} + \boldsymbol{w}_{N-p}. \tag{7.31}$$

However, as in the case of Fourier series, the maximum likelihood estimator of (7.9) cannot be used because the columns of $\boldsymbol{X}$ lie in the column span of $\boldsymbol{W}$. An appealing alternative is to select a subspace of dimension $r << N-p$ in order to model the time-varying mean, but selecting the best such subspace (in the sense of maximizing the conditional likelihood of the data) is impractical since it requires the examination of combinatorially many solutions.

To avoid these difficulties our approach is to consider iterative, not joint, estimators of $\boldsymbol{a}$ and $\boldsymbol{\beta}$ as is typical in inference methods for semi-parametric regression and, in particular, for partially linear models [228, 234]. To this end, we define the vector $\boldsymbol{x}_w \triangleq \boldsymbol{W}^T\boldsymbol{x}_{N-p}$ and the matrix $\boldsymbol{X}_w \triangleq \boldsymbol{W}^T\boldsymbol{X}$ as the discrete wavelet transforms of the data $\boldsymbol{x}_{N-p}$ and the

design matrix $\boldsymbol{X}$, respectively. Next, observe that the normal equations of (7.11) and (7.12) may be rewritten in the wavelet domain as follows:

$$\begin{aligned}\boldsymbol{a} &= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T(\boldsymbol{x}_{N-p} - \boldsymbol{W}\boldsymbol{\beta}) \\ &= \left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{W}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{W}^T(\boldsymbol{x}_{N-p} - \boldsymbol{W}\boldsymbol{\beta}) \\ &= \left((\boldsymbol{W}^T\boldsymbol{X})^T(\boldsymbol{W}^T\boldsymbol{X})\right)^{-1}(\boldsymbol{W}^T\boldsymbol{X})^T\left(\boldsymbol{W}^T\boldsymbol{x}_{N-p} - \boldsymbol{W}^T\boldsymbol{W}\boldsymbol{\beta}\right) \\ &= \left(\boldsymbol{X}_w^T\boldsymbol{X}_w\right)^{-1}\boldsymbol{X}_w^T(\boldsymbol{x}_w - \boldsymbol{\beta})\end{aligned} \tag{7.32}$$

$$\boldsymbol{\beta} = \left(\boldsymbol{W}^T\boldsymbol{W}\right)^{-1}\boldsymbol{W}^T(\boldsymbol{x} - \boldsymbol{X}\boldsymbol{a}) = (\boldsymbol{x}_w - \boldsymbol{X}_w\boldsymbol{a}), \tag{7.33}$$

Since it is desirable that (7.32) and (7.33) should be satisfied by estimators of $\boldsymbol{a}$ and $\boldsymbol{\beta}$, a natural iterative estimation approach would be to initialize $\boldsymbol{\beta}$ to $\boldsymbol{0}$ and to iteratively update estimates of $\boldsymbol{a}$ and $\boldsymbol{\beta}$ using (7.32) and (7.33).

We will see in Section 7.7.1.2, however, that this approach results in severely undersmoothed estimates of the time-varying mean in the speech processing setting. Indeed, as discussed earlier in Section 2.2, the physical process of modulation of the airflow by the vocal folds—resulting in the deterministic component of the source waveform that we are attempting to model with the time-varying mean $\mu[n]$—produces a relatively smooth signal. As a result, we have observed its wavelet decomposition to be relatively sparse at high-resolution levels. Accordingly, we combine (7.32) and (7.33) together with a nonlinear thresholding step as we now describe.

Let L be an integer such that 2^L denotes the number of scaling functions in the wavelet decomposition. Thus there are 2^L "coarse" coefficients associated to translates and dilations of the scaling function, and $N-p-2^L$ "detail" coefficients associated to translates and dilations of the mother wavelet. Accordingly, we partition the wavelet coefficient vector $\boldsymbol{\beta}$ via:

$$\boldsymbol{\beta} = \begin{pmatrix}\boldsymbol{\beta}_c^T & \boldsymbol{\beta}_d^T\end{pmatrix}^T,$$

where $\boldsymbol{\beta}_c \in \mathbb{R}^{2^L \times 1}$ and $\boldsymbol{\beta}_d \in \mathbb{R}^{(N-p-2^L)\times 1}$. Then the subspace selection algorithm based on wavelet shrinkage proceeds as described in Algorithm 7.1.

It is illustrative to describe the steps of Algorithm 7.1 in the time domain. At the ith iteration of Algorithm 7.1, the estimate of $\mu[n]$ based on $\widehat{\boldsymbol{\beta}}^{(i-1)}$ is subtracted from the waveform, and the covariance method of linear prediction is used to estimate $\boldsymbol{a}$ from the residual[1]. Next the waveform is inverse filtered, using a moving average filter with coefficients $\widehat{\boldsymbol{a}}^{(i)}$, and the discrete wavelet transform of the residual is obtained. A nonlinear thresholding step is then applied to the resultant detail coefficients $\boldsymbol{\beta}_d$ in order to regularize the overall solution and, in the case of hard thresholding, to promote sparsity. Thresholding is achieved through the hard or soft thresholding rule with the universal threshold of [235]. The required estimate of σ is obtained as the median absolute deviation of the finest-resolution wavelet coefficients divided by 0.6745. We note that alternative thresholding rules and stopping criteria are possible.

Using hard thresholding in Algorithm 7.1 explicitly promotes sparsity in the estimate of $\boldsymbol{\beta}$. An alternative means to the same end is to penalize the detail coefficients in the

[1]Thus the estimate of $\boldsymbol{a}$ during the first iteration is equivalent to that obtained by the covariance method.

Algorithm 7.1 Subspace Selection via Iterative Shrinkage

- Initialization: Set tolerance level ϵ_0, iteration counter $i = 0$, number of coarse levels L, and initial estimate of the coefficients $\widehat{\boldsymbol{\beta}}^{(0)} = \mathbf{0} \in \mathbb{R}^{(N-p)\times 1}$
- While $\epsilon > \epsilon_0$
  - Update estimates of $\boldsymbol{a}$ and $\boldsymbol{\beta}$ via:

$$\widehat{\boldsymbol{a}}^{(i)} = \left(\boldsymbol{X}_w^T \boldsymbol{X}_w\right)^{-1} \boldsymbol{X}_w^T \left(\boldsymbol{x}_w - \widehat{\boldsymbol{\beta}}^{(i-1)}\right)$$

$$\widehat{\boldsymbol{\beta}}^{(i)} \triangleq \left(\widehat{\boldsymbol{\beta}}_c^{(i)^T} \quad \widehat{\boldsymbol{\beta}}_d^{(i)^T}\right)^T = \boldsymbol{x}_w - \boldsymbol{X}_w \widehat{\boldsymbol{a}}^{(i)}$$

  - Calculate universal threshold:
    - Calculate median absolute deviation estimate of σ^2 as detailed in [235]
    - Set λ to $\sqrt{2\widehat{\sigma^2}\log(N-p-2^L)}$
  - Thresholding: for all $1 \le j \le N-p-2^L$

$$\text{Hard:} \quad \widehat{\boldsymbol{\beta}_d^{(i)}}(j) = \begin{cases} \widehat{\boldsymbol{\beta}_d^{(i)}}(j) & \text{if} \quad |\widehat{\boldsymbol{\beta}_d^{(i)}}(j)| > \lambda \\ 0 & \text{if} \quad |\widehat{\boldsymbol{\beta}_d^{(i)}}(j)| \le \lambda \end{cases}$$

$$\text{Soft:} \quad \widehat{\boldsymbol{\beta}_d^{(i)}}(j) = \text{sign}\left(\widehat{\boldsymbol{\beta}_d^{(i)}}(j)\right) \max\left(\left|\widehat{\boldsymbol{\beta}_d^{(i)}}(j)\right| - \lambda, 0\right)$$

  - Compute change in AR coefficient vector and increment: $i = i + 1$

$$\epsilon = \left\|\widehat{\boldsymbol{a}}^{(i)} - \widehat{\boldsymbol{a}}^{(i-1)}\right\|^2$$

- Return $\boldsymbol{a}^{(i)}$ and $\boldsymbol{\beta}^{(i)}$

least-squares problem of (7.31) directly using an ℓ_1 norm, which is also known to promote sparsity in the estimates [231]. Here, we estimate $\boldsymbol{a}$ and $\boldsymbol{\beta}$ as the solution to the following (convex) optimization problem:

$$\left(\widehat{\boldsymbol{a}}, \widehat{\boldsymbol{\beta}}\right) = \underset{(\boldsymbol{a},\boldsymbol{\beta})}{\text{argmin}} \ \|\boldsymbol{x}_w - \boldsymbol{X}_w\boldsymbol{a} - \boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}_d\|_1, \tag{7.34}$$

where λ is an appropriately-chosen threshold. The solution to (7.34) may be efficiently found using freely-available software packages such as CVX [131].

7.6 Hypothesis Testing

The ARX model specified by (7.2) and (7.3) naturally lends itself to testing whether or not the mean $\mu[n]$ is equal to 0. In the context of speech analysis, this amounts to testing whether or not a specific sound is voiced since the basis functions $\{g_1[n], \ldots, g_r[n]\}$ of (7.3) may be chosen to capture the deterministic component of the source waveform during voicing.

Recalling that the ARX(p) model reduces to the classical AR(p) model when $\beta_k = 0$ for all $1 \leq k \leq r$, we formulate the following hypothesis test for the presence of the mean $\mu[n]$:

$$\begin{aligned} \text{Model}: &\quad \text{ARX}(p) \text{ with parameters } \boldsymbol{a}, \boldsymbol{\beta}, \sigma^2; \\ \text{Hypotheses}: &\quad \begin{cases} \mathcal{H}_0 : \boldsymbol{\theta} &= \begin{pmatrix} \boldsymbol{a} & \boldsymbol{\beta} = \mathbf{0}_{r\times 1} & \sigma^2 \end{pmatrix} \\ \mathcal{H}_1 : \boldsymbol{\theta} &= \begin{pmatrix} \boldsymbol{a} & \boldsymbol{\beta} \neq \mathbf{0}_{r\times 1} & \sigma^2 \end{pmatrix} \end{cases} \end{aligned} \tag{7.35}$$

Each of these two hypotheses in turn induces a data likelihood in the observed signal $\boldsymbol{x} \in \mathbb{R}^{N\times 1}$, which we denote by $p_{\mathcal{H}_i}(\cdot)$ for $i = 0, 1$. The hypothesis test of (7.35) can then be realized using a number of test statistics; we discuss three choices corresponding to the generalized likelihood ratio test (GLRT), the Rao test, and the Wald test. All three involve computing a test statistic ($T_G(\boldsymbol{x})$, $T_W(\boldsymbol{x})$, and $T_R(\boldsymbol{x})$, respectively) and rejecting $\mathcal{H}_0$ in favor of $\mathcal{H}_1$ if the test statistic exceeds a given threshold γ.

7.6.1 Generalized Likelihood Ratio Test (GLRT)

The generalized ratio test (GLRT) statistic is defined by:

$$T_G(\boldsymbol{x}) \triangleq 2 \ln \frac{\sup_{\boldsymbol{a},\boldsymbol{\beta},\sigma^2} p_{\mathcal{H}_1}(\boldsymbol{x}\,;\, \boldsymbol{a}, \boldsymbol{\beta}, \sigma^2)}{\sup_{\boldsymbol{a},\sigma^2} p_{\mathcal{H}_0}(\boldsymbol{x}\,;\, \boldsymbol{a}, \sigma^2)}. \tag{7.36}$$

In order to compute (7.36), the (conditional) ML estimators of $\boldsymbol{a}, \boldsymbol{\beta}$ and σ^2 under $\mathcal{H}_1$ are obtained using (7.8) and (7.13), respectively. Estimates of $\boldsymbol{a}$ and σ^2 under $\mathcal{H}_0$ are obtained via the covariance method of linear prediction, which is equivalent to setting $q = 0$ in (7.8) and (7.13). Substituting these into (7.36) leads to the following form for $T_G(\boldsymbol{x})$:

$$T_G(\boldsymbol{x}) = (N - p) \ln \left(\widehat{\sigma^2}_{\mathcal{H}_0} / \widehat{\sigma^2}_{\mathcal{H}_1} \right). \tag{7.37}$$

7.6.2 Wald Test

The GLRT statistic of (7.36) requires fitting both under the null and the alternate hypotheses. On the other hand, the Wald test statistic requires fitting only under $\mathcal{H}_1$ and is defined by [152]:

$$T_W(\boldsymbol{x}) \triangleq \widehat{\boldsymbol{\beta}}^T \left[\boldsymbol{I}^{-1}(\widehat{\boldsymbol{\theta}}) \right]_{\boldsymbol{\beta}\boldsymbol{\beta}} \widehat{\boldsymbol{\beta}}, \tag{7.38}$$

where $\widehat{\boldsymbol{\theta}} \triangleq (\widehat{\boldsymbol{a}}, \widehat{\boldsymbol{\beta}}, \widehat{\sigma^2})$ is the (conditional) ML estimator obtained using (7.8) and (7.13) under $\mathcal{H}_1$, and

$$\left[\boldsymbol{I}^{-1}(\boldsymbol{\theta}) \right]_{\boldsymbol{\beta}\boldsymbol{\beta}} \triangleq \left(\boldsymbol{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\theta}) - \boldsymbol{I}_{\boldsymbol{\beta}\boldsymbol{a}}(\boldsymbol{\theta}) \boldsymbol{I}_{\boldsymbol{a}\boldsymbol{a}}^{-1}(\boldsymbol{\theta}) \boldsymbol{I}_{\boldsymbol{a}\boldsymbol{\beta}}(\boldsymbol{\theta}) \right), \tag{7.39}$$

is the inverse of the Schur complement of $\boldsymbol{I}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}(\boldsymbol{\theta})$ in $\boldsymbol{I}(\boldsymbol{\theta})$. Evaluating (7.39) using (7.26) yields:

$$\left[\boldsymbol{I}^{-1}(\boldsymbol{\theta}) \right]_{\boldsymbol{\beta}\boldsymbol{\beta}} = \left(\boldsymbol{G}^T \boldsymbol{G} - \boldsymbol{M} \left(\widetilde{\boldsymbol{R}} + \boldsymbol{M}^T \boldsymbol{M} \right)^{-1} \boldsymbol{M}^T \right) \sigma^{-2},$$

which is identical to the CRLB for $\boldsymbol{\beta}$ given by (7.28). Substituting the above into (7.38) shows that Wald test statistic takes the following form:

$$T_W(\boldsymbol{x}) = \frac{\widehat{\boldsymbol{\beta}}^T \boldsymbol{G}^T \left(\boldsymbol{I}_{N-p} - \widehat{\boldsymbol{M}} \left(\widehat{\widehat{\boldsymbol{R}}} + \widehat{\boldsymbol{M}}^T \widehat{\boldsymbol{M}} \right)^{-1} \widehat{\boldsymbol{M}}^T \right) \boldsymbol{G} \widehat{\boldsymbol{\beta}}}{\widehat{\sigma^2}}, \tag{7.40}$$

where $\boldsymbol{I}_N$ is an $N \times N$ identity matrix. In practice, in order to evaluate $T_W(\boldsymbol{x})$, we must first estimate $\widehat{\boldsymbol{a}}, \widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma^2}$ using (7.8), and (7.13), and then obtain $\widehat{\boldsymbol{R}}$ from $\widehat{\boldsymbol{a}}$ using the step-down procedure, and $\widehat{\boldsymbol{M}}$ from $\widehat{\boldsymbol{a}}$ and $\widehat{\boldsymbol{\beta}}$ using (7.25).

Next, observe that by the same manipulations as in (7.28) that the Wald test statistic may be rewritten according to:

$$\begin{aligned} \widehat{\sigma^2} T_W(\boldsymbol{x}) &= \widehat{\boldsymbol{\mu}}^T \left(\boldsymbol{I}_{N-p} - \widehat{\boldsymbol{M}} \left(\widehat{\widehat{\boldsymbol{R}}} + \widehat{\boldsymbol{M}}^T \widehat{\boldsymbol{M}} \right)^{-1} \widehat{\boldsymbol{M}}^T \right) \widehat{\boldsymbol{\mu}} \\ &= \widehat{\boldsymbol{\mu}}^T \left(\boldsymbol{I}_{N-p} + \frac{\widehat{\boldsymbol{M}} \widehat{\boldsymbol{R}}^{-1} \widehat{\boldsymbol{M}}^T}{N-p} \right)^{-1} \widehat{\boldsymbol{\mu}}. \end{aligned}$$

If the underlying process satisfies the regularity conditions of Theorem 8, the above manipulations imply that the Wald test statistic converges to the ratio of the energy of $\mu[n]$ and σ^2 in the model to:

$$\lim_{N \to \infty} T_W(\boldsymbol{x}) = \frac{\widehat{\boldsymbol{\mu}}^T \widehat{\boldsymbol{\mu}}}{\widehat{\sigma^2}} = \frac{\|\widehat{\boldsymbol{\mu}}\|^2}{\widehat{\sigma^2}}.$$

The asymptotic form of (7.6.2) is very intuitive—in the context of speech it simply measures the ratio of the voicing to the aspiration energy.

7.6.3 Rao Test

In contrast to the Wald and GLRT statistics of (7.36) and (7.40), the Rao test statistic only requires fitting under the null hypothesis and is defined by [152]:

$$T_R(\boldsymbol{x}) \triangleq \left. \frac{\partial \log p(\boldsymbol{x} \,|\, \boldsymbol{x}_p \,;\, \boldsymbol{\theta})}{\partial \boldsymbol{\beta}} \right|^T_{\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}} \left[\boldsymbol{I}^{-1}(\widetilde{\boldsymbol{\theta}}) \right]^{-1}_{\boldsymbol{\beta}\boldsymbol{\beta}} \left. \frac{\partial \log p(\boldsymbol{x} \,|\, \boldsymbol{x}_p \,;\, \boldsymbol{\theta})}{\partial \boldsymbol{\beta}} \right|_{\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}}, \tag{7.41}$$

where $\widetilde{\boldsymbol{\theta}} \triangleq (\widetilde{\boldsymbol{a}}, \widetilde{\sigma^2})$ is obtained using (7.8) and (7.13) under $\mathcal{H}_0$, and $\left[\boldsymbol{I}^{-1}(\boldsymbol{\theta}) \right]_{\boldsymbol{\beta}\boldsymbol{\beta}}$ is defined in (7.39). Note the differences between (7.38) and (7.41): In the Wald test all estimates are obtained under $\mathcal{H}_1$ and the Schur complement of $\boldsymbol{I}_{\boldsymbol{aa}}(\boldsymbol{\theta})$ in $\boldsymbol{I}(\boldsymbol{\theta})$ is used, whereas in the Rao test all estimates are obtained under $\mathcal{H}_0$ and the *inverse* of the Schur complement of $\boldsymbol{I}_{\boldsymbol{aa}}(\boldsymbol{\theta})$ in $\boldsymbol{I}(\boldsymbol{\theta})$ is used.

Evaluating the terms in (7.41) under $\mathcal{H}_0$ yields

$$\left. \frac{\partial \log p(\boldsymbol{x} \,|\, \boldsymbol{x}_p; \boldsymbol{\theta})}{\partial \boldsymbol{\beta}} \right|_{\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}} = -\frac{\boldsymbol{G}^T \widehat{\boldsymbol{e}}}{\widehat{\sigma}^2} \left[\boldsymbol{I}^{-1}(\widetilde{\boldsymbol{\theta}}) \right]^{-1}_{\boldsymbol{\beta}\boldsymbol{\beta}} = \left(\boldsymbol{G}^T \boldsymbol{G} \right)^{-1} \widehat{\sigma^2},$$

where $\widehat{\boldsymbol{e}} = \boldsymbol{x} - \boldsymbol{X}\widehat{\boldsymbol{a}}$ is the estimate of the residual (under $\mathcal{H}_0$). Substituting the above into (7.38) shows that the Rao test statistic takes the following form:

$$T_R(\boldsymbol{x}) = \frac{\widehat{\boldsymbol{e}}^T \boldsymbol{G} \left(\boldsymbol{G}^T \boldsymbol{G}\right)^{-1} \boldsymbol{G}^T \widehat{\boldsymbol{e}}}{\widehat{\sigma^2}} = \frac{\widehat{\boldsymbol{e}}^T \boldsymbol{P}_G \widehat{\boldsymbol{e}}}{\widehat{\sigma^2}} = \frac{\widehat{\boldsymbol{e}}^T \boldsymbol{P}_G \boldsymbol{P}_G \widehat{\boldsymbol{e}}}{\widehat{\sigma^2}} = \frac{\|\boldsymbol{P}_G \widehat{\boldsymbol{e}}\|^2}{\widehat{\sigma^2}} \tag{7.42}$$

7.6.4 Detection Performance

The GLR, Wald, and Rao test statistics all boil down to a ratio of estimated energies of the exogenous mean $\mu[n]$ and the variance of the Gaussian innovations sequence $w[n]$. In the finite-sample-size setting, the differences manifest themselves primarily in the quality of the underlying estimators—we study these tradeoffs empirically below. On the other hand, these tests are known to be asymptotically equivalent. The GLRT statistic has the following asymptotic behavior [152]:

$$\lim_{N\to\infty} T_G(\boldsymbol{x}) \overset{\mathcal{D}}{\sim} \chi^2_d(\lambda), \quad \begin{cases} \lambda = 0 & \text{under } \mathcal{H}_0, \\ \lambda > 0 & \text{under } \mathcal{H}_1. \end{cases}, \tag{7.43}$$

where the degrees of freedom d is equal to q the number of unrestricted parameters under the null hypothesis, and the noncentrality parameter λ is given by:

$$\lambda = \sigma^{-2} \boldsymbol{\beta}^T \left(\boldsymbol{G}^T \boldsymbol{G}\right) \boldsymbol{\beta} = \sigma^{-2} \boldsymbol{\mu}^T \boldsymbol{\mu}. \tag{7.44}$$

Note that since the asymptotic distribution of the test statistics under $\mathcal{H}_0$ depends *only* on q, which is set in advance, we can determine a constant false alarm rate (CFAR) threshold γ by fixing a desired false alarm value (say, 5%), and evaluating the inverse cumulative distribution function of $\chi^2_q(0)$ to obtain the value of γ that guarantees the specified (asymptotic) constant false alarm rate. In addition note, that the asymptotic results of (7.43) hold for only small source-harmonics-to-noise ratios [152].

7.7 Experiments

In this section, we study the behavior of various methods derived from the ARX(p) model of (7.2) using synthetic data in Section 7.7.1 and natural data in Section 7.7.2. The experiments using synthetic speech data are especially important since, in practice, the true source waveform cannot be measured—at best only correlates such as the electroglottograph signal (EGG) are available. Additional synthetic experiments are reported on in Appendix 7.A.

Some of the reported results in the sections below rely on the *source* harmonics-to-noise ratio (SHNR) ρ defined (in dB) by:

$$\rho \triangleq 10 \log_{10} \left(\|\mu[n]\|^2 / (N\sigma^2) \right), \tag{7.45}$$

where N is the length of the waveform; the importance of this measure in clinical applications and its properties are detailed in [236]. Note that relationship between (7.45) and the noncentrality parameter (7.44)—they differ only in that the average energy of $\mu[n]$ is

used in (7.45) rather than the total energy in (7.44)[2]. We will also consider the problem of estimating SHNR in the experiments below.

In addition, we measure the distance between two power-spectra $S_1(\omega)$ and $S_2(\omega)$ via the log-spectral (LS) distance defined (in dB) over L frequency bins $\{\omega_1, \ldots, \omega_L\}$ by:

$$d_{\text{LS}}(S_1, S_2) \triangleq \sqrt{\frac{1}{L}\sum_{k=1}^{L}\left[10\log_{10} S_1(\omega_k) - 10\log_{10} S_2(\omega_k)\right]^2},$$

though other choices, including the many metrics discussed in our treatment of regularized TVAR models in Section 4.6, are possible.

7.7.1 Synthetic Speech Experiments

In all the experiments in this section, speech was synthesized by a formant synthesizer in which the vocal tract is represented by a cascade of three second-order digital resonators. This filter is subsequently excited by a mixture of a deterministic voicing component $\mu[n]$ generated using the Rosenberg-Klatt model [18] and white Gaussian noise $\sigma w[n]$ modeling the aspiration stochastically. The relative power of the deterministic and stochastic components are determined by selecting an SHNR value prior to synthesis. Note that this approach allows us to synthesize data according to the ARX model when the number of AR coefficients is set as $p = 6$. Though illustrated only for a range of phonemes below and in Appendix 7.A, the results below hold across a variety of phonemes, pitch frequencies, wavelet choices, and SHNR values.

In Section 7.7.1.1, we show that, in the presence of voicing, estimating the vocal tract using ARX modeling results in *more accurate* estimates of the vocal tract transfer function than via classical linear prediction. In addition, we examine the effect of time-varying pitch on the spectral estimates. Next, in Section 7.7.1.2, we evaluate the model selection approaches of Section 7.5. In Section 7.7.1.3, we empirically study the detection performance of the GLRT, Rao, and Wald tests of Section 7.6, and consider the closely-related problem of SHNR estimation.

7.7.1.1 Spectral Analysis

In the first experiment, we fit the ARX(6) model to a 32 ms, synthetic phoneme [ɪ] (as in bit), generated at a sampling rate of 16 kHz, a fundamental frequency of 200 Hz, and an SHNR of $\rho = 10$ dB. The waveform and the deterministic voicing component $\mu[n]$ used to synthesize it are shown in the top-left and top-right panels of Figure 7.2, respectively. An ARX(6) model of (7.2) was fit to the data via (7.8) using the first $r = 64$ Daubechies 6 wavelets—we provided some justification for this choice of basis functions in Section 7.5.1; an online subspace selection approach is considered in Section 7.7.1.2 below. For comparison, we fit an AR(8) model to the data after a pre-emphasis step to remove the spectral tilt induced by the shape of the glottal flow pulses. We compare an ARX(6) to an AR(8),

[2]This difference is quite sensible since as we get more data from a signal with fixed SHNR ρ, the detection performance of hypothesis tests should improve as a function of an increase in the value λ.

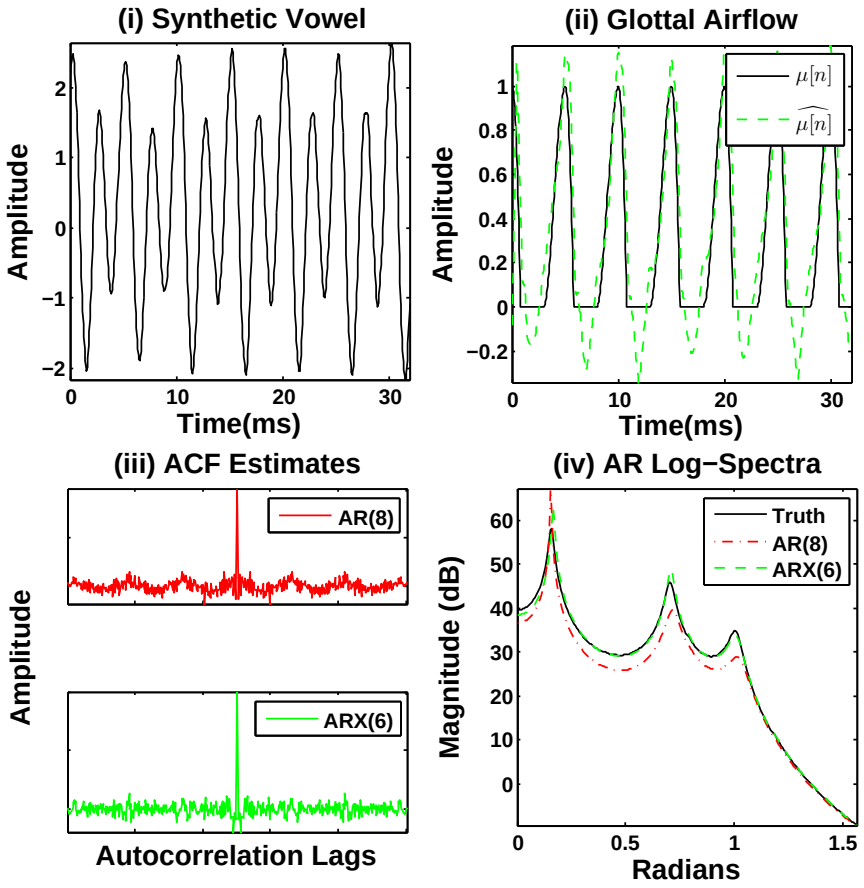


Figure 7.2: Comparison of ARX(6) and AR(8) models on a synthetic vowel with a constant pitch contour: (i) synthesized phoneme [ɪ]; (ii) voiced component of glottal flow (solid, black) and its estimate (dashed, green); (iii) autocorrelations of the AR(8) (red) and ARX(6) (green) residuals; (iv) log-spectra for the true AR(6) (solid, black) and fitted AR(8) (dashed-dot, red) and ARX(6) (dashed, green) models.

instead of an AR(6), model because the two additional poles help to roughly capture the glottal pulse shape as described in Section 2.4.6; adding more than two poles did not appreciably alter the results, but without adding at least two poles, the performance of the classical linear prediction approach degrades *significantly*.

The resultant estimate of $\mu[n]$, computed according to $\widehat{\mu[n]} = \boldsymbol{G}\widehat{\beta}$ and shown in the top-right panel of Figure 7.2, clearly captures the general quality and periodicity of the true glottal flow. Indeed, comparing the autocorrelation of the AR(8) and ARX(6) residuals, as shown in the bottom-left panel of Figure 7.2, reveals that, unlike the AR(8) residual, the ARX(6) residual has virtually no periodic structure; the model was able to explain the variation due to the deterministic voicing component. Finally, both estimates of the all-pole power spectra are shown in bottom-right panel of Figure 7.2 alongside the *true* AR(6) power spectrum. It is evident that the all-pole spectrum estimated via the ARX(6) model is significantly closer to the true spectrum than the estimate obtained via

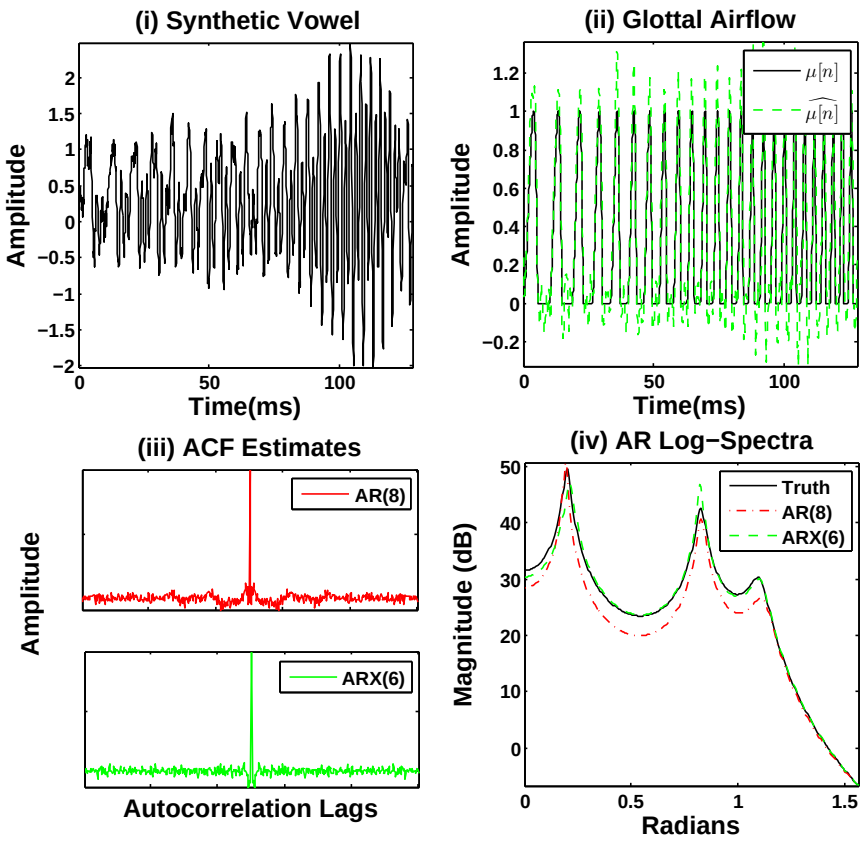


Figure 7.3: Comparison of ARX(6) and AR(8) models on a synthetic vowel with a time-varying pitch contour: (i) synthesized phoneme [ɛ]; (ii) voiced component of glottal flow (solid, black) and its estimate (dashed, green); (iii) autocorrelations of the AR(8) (red) and ARX(6) (green) residuals; (iv) log-spectra for the true AR(6) (solid, black) and fitted AR(8) (dashed-dot, red) and ARX(6) (dashed, green) models.

the covariance method (1.08 dB and 3.03 dB LS distance, respectively). Indeed, the ARX(6) model captures the *valleys* more accurately; as is further demonstrated in Appendix (7.A), these results are typical over a broad range of synthetic phonemes. Note that we do not pre-emphasize the data prior to fitting the ARX(6) model since $\mu[n]$ also captures the spectral tilt due to the glottal pulse shape.

In the second experiment, we study the performance of the ARX model in the presence of pitch variation by repeating the first experiment on a 128 ms vowel [ɛ] (as in bait) synthesized with linearly-increasing pitch from 100 to 300 Hz over its duration, and shown in the top-left panel of Figure 7.3. An ARX(6) model was fitted to the data via (7.8) using the first 128 Daubechies 6 wavelets (since the waveform is four times longer than that of the first experiment). Clearly, despite the extreme pitch variation, the ARX(6) model yields an accurate fit of $\mu[n]$ and a more accurate estimate of the true spectrum than the covariance method (1.10 dB and 2.86 dB LS distance, respectively). Moreover, the AR(8) residual shows periodicity whereas the ARX(6) residual is noiselike. The last experiment highlighted

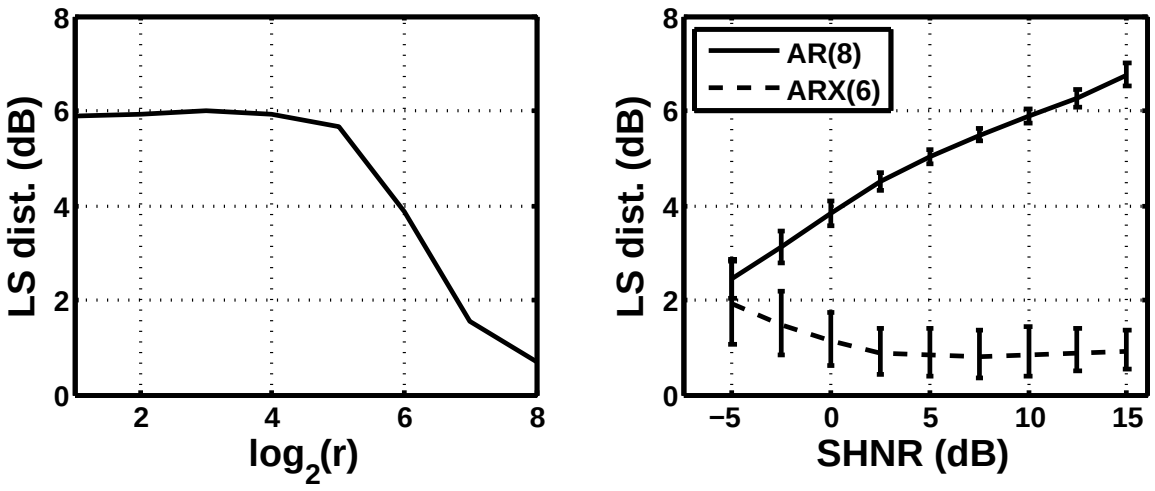


Figure 7.4: Accuracy of the ARX(6) spectral estimator as a function of the number of wavelets used (left) and the ARX(6) and AR(8) spectral estimators as a function of the SHNR ρ (right).

the advantage of using a *local* wavelet basis; using a *global* Fourier series expansion instead leads to poor results as the underlying source waveform is not periodic. Here, we did not attempt to optimize performance among different wavelet families; indeed, many (e.g., Haar) produce similar results.

Next we consider the accuracy of spectral estimation as a function of r—the number of wavelets used—for a 64 ms vowel [ɪ] synthesized exactly as in the first experiment. As shown in the left panel of Figure 7.4, the estimation accuracy improves with increasing r, as expected. We also studied the relative performance of the ARX- and AR-based spectral estimators as a function of ρ, the source harmonics-to-noise ratio. Both methods were applied to a 64 ms segment of the phoneme [ɪ] synthesized for values of ρ ranging between -5 and 15 dB; the average LS distances over 500 Monte Carlo trials are shown in the right-panel of Figure 7.4. Clearly, as ρ increases then so does the error of the AR spectral estimator (similar results are obtained when comparing against an AR(p) model for all $p \geq 8$), whereas the performance of the ARX spectral estimator is relatively unaffected. At low SHNR values, the behavior of the AR and ARX models is similar since, in this case, the glottal flow is mostly aspiration. The experiment shows that the accuracy of the ARX spectral estimator does not significantly depend on ρ—a desirable feature for speech analysis.

7.7.1.2 Signal-Adaptive Subspace Selection

Next, we illustrate the online signal-adaptive subspace selection methods of 7.5.2. First, we apply Algorithm 7.1 to the signal shown in Figure 7.2, but do not employ any thresholding rule. The resultant procedure reduces to iteratively computing (7.32) and (7.33), and leads to the results shown in Figure 7.5. It can be seen that the estimate of the source waveform is significantly undersmoothed, and consequently the estimated spectrum is biased.

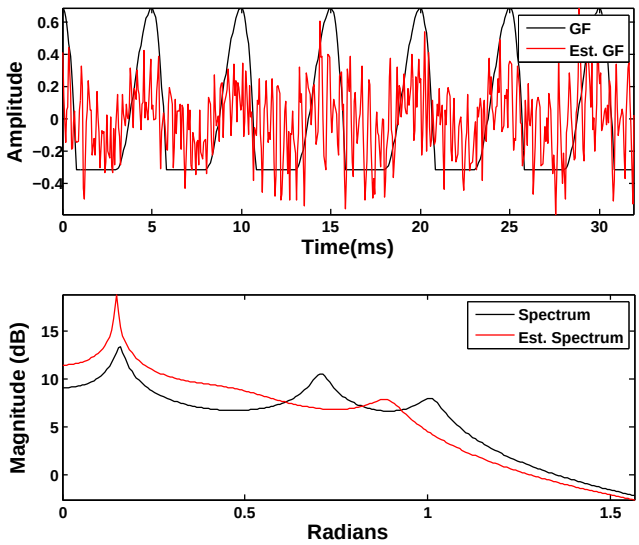


Figure 7.5: Subspace selection via iterative estimation without shrinkage. Top panel: the time-varying mean $\mu[n]$ (solid, black) is overlaid with an estimate of $\mu[n]$ obtained by Algorithm 7.1 (dashed, red). Bottom panel: the log-spectrum for the true AR(6) process is overlaid with a spectral estimate obtained by Algorithm 7.1 (dashed, red).

To address this shortcoming, we apply Algorithm 7.1 using a hard thresholding step to the same waveform. We employed Daubechies 6 wavelets and set the number of coarse levels L was set to 0 so that only the first coefficient was not subject to shrinking. Afterward, the wavelets associated to the 64 largest-magnitude inferred coefficients were taken as columns of a matrix $\boldsymbol{G}$ and the conditional maximum likelihood estimator of (7.8) was used to obtain a joint fit of the AR coefficients and the glottal flow.

The results are shown in the top- and bottom-left panels of Figure 7.6 and it is clear that Algorithm 7.1 is able to accurately estimate both $\mu[n]$ and the AR spectrum (1.65 dB LS distance). Subsequent refitting using the largest-magnitude wavelets via (7.8) further improved the spectral estimate (0.55 dB LS distance). As shown in the top- and bottom-panels of Figure 7.6, similar performance is observed in our second example using the waveform shown in Figure 7.3 with Algorithm 7.1 yielding a comparably accurate estimate of the spectrum (1.19 dB LS distance). No appreciable change is observed after subsequent refitting using (7.8). In addition, we applied Algorithm 7.1 using soft thresholding and the ℓ_1-regularization approach of (7.34) to the constant-pitch example delineated above. In the case of soft thresholding, shown in the top- and bottom-left panels of Figure 7.7, we can see that the estimate of $\mu[n]$ only partially reflects its periodicity and general quality, and there is a strong bias in the spectral estimate (7.62 dB LS distance). However, a subsequent refitting using the top 64 wavelets and the estimator (7.8) yields a significant improvement (0.82 dB LS distance). The estimates produced by the ℓ_1 approach are comparable (1.57 dB LS distance) in accuracy. Both of these estimators show some promise and need to be further investigated. More generally, level-dependent thresholding or penalization of wavelet coefficients may improve the performance of the presented algorithms.

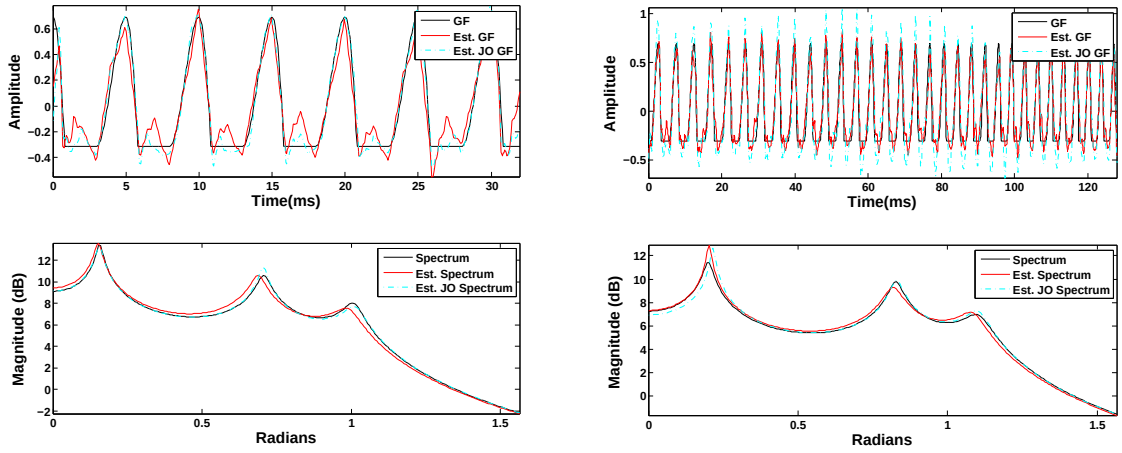


Figure 7.6: Subspace selection via iterative wavelet shrinkage and hard thresholding on the synthetic phoneme [ɪ] with constant (left) and time-varying (right) pitch contours. Top panels: the time-varying mean $\mu[n]$ (solid, black) is overlaid with estimates of $\mu[n]$ obtained by Algorithm 7.1 (dashed, red) and subsequent refitting via (7.8) (dashed, cyan). Bottom panels: the log-spectrum for the true AR(6) process is overlaid with spectral estimates obtained by Algorithm 7.1 (dashed, red) and subsequent refitting via (7.8) (dashed, cyan).

7.7.1.3 Hypothesis Testing

Here we study the detection performance of the hypothesis tests of Section 7.6 in a small-sample-size setting and when the number of observations N is large, so that we are essentially in the asymptotic regime. In addition, we consider the closely-related problem of estimating SHNR from data.

In the first experiment, we consider the convergence of the GLR, Wald, and Rao test statistics of (7.36), (7.40), and (7.42), respectively, to their (shared) asymptotic distribution. To this end, a 64 ms phoneme [ɛ] (as in bet) was synthesized with a constant fundamental frequency of 100 Hz and an SHNR of $\rho = -15$ dB. In addition, so as to avoid any model mismatch, a least-squares $r = 16$ Daubechies 4 wavelet approximation to the source waveform was used during synthesis. In turn, the test statistics were calculated under the assumption of $p = 6$ and using the first $r = 16$ Daubechies 4 wavelets.

The resultant sampling distributions of the GLRT statistic of (7.36) under $\mathcal{H}_0$ and $\mathcal{H}_1$ were obtained via 5000 Monte Carlo trials and are shown in the top-left panel of Figure 7.8. The sampling distributions are overlaid with the plots of the chi-squared densities with $d = pr$ degrees of freedom and with $\lambda = 0$ under $\mathcal{H}_0$, and calculated according to (7.44) under $\mathcal{H}_1$. The receiver operating characteristics associated to the GLRT as well as the Rao and Wald tests, calculated over the same set of trials, are shown in the bottom-left panel of Figure 7.8.

The same experiments were repeated for 5000 instantiations of the same phoneme, but now using 16 ms waveforms with all other parameters left unchanged. The resultant sampling distributions and associated ROC curves are shown in the top- and bottom-right panels of Figure 7.8, respectively. These results confirm that in the near-asymptotic regime—a 64 ms waveform has $N = 1024$ samples at 16 kHz relative to $p \cdot r = 6 \cdot 16 = 96$

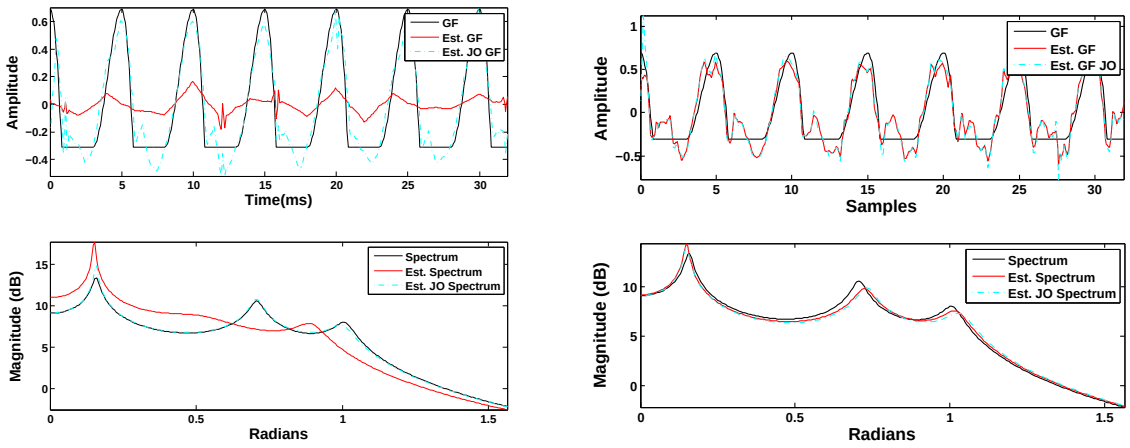


Figure 7.7: Subspace selection via iterative wavelet shrinkage with soft thresholding (left) and ℓ_1 regularization (right) on the synthetic phoneme [ɪ] with a constant pitch contour. Top panels: the time-varying mean $\mu[n]$ (solid, black) is overlaid with estimates of $\mu[n]$ (dashed, red) and subsequent refitting via (7.8) (dashed, cyan). Bottom panels: the log-spectrum for the true AR(6) process is overlaid with spectral estimates (dashed, red) and subsequent refitting via (7.8) (dashed, cyan).

parameters—all three test statistics (nearly) converge to their shared asymptotic distribution. On the other hand, the sampling distributions of the test statistics differ from one another, in the non-asymptotic regime, when $N = 256$ relative to $p \cdot r = 6 \cdot 16 = 96$ parameters. As shown in the bottom-right panel of Figure 7.8, the Rao test yields the more powerful test in this case. We have, in fact, observed the Rao test to outperform the GLR and Wald tests for a broad range of synthetic speech experiments in the small-sample-size setting. The Rao test performs especially well relative to the other tests when the source waveform used for synthesis is not in the column space of $\boldsymbol{G}$, even if the energy of the projection of $\mu[n]$ onto the subspace spanned by the basis functions is large.

In the second experiment we study the detection performance of the GLRT statistic of (7.36) as a function of signal length and SHNR. To this end, we synthesize the phoneme [ɛ] for the durations of $8, 16, 32$ and 64 ms and at SHNRs of $-25, -20, -15$ and -10 dB. The number of basis functions r was doubled for each doubling of signal length during fitting under $\mathcal{H}_1$. The resultant ROC curves are shown in Figure 7.9; the detection performance improves with increasing signal length and SHNR, in agreement with our intuition.

Finally, we turn to the question of estimating SHNR from data. The straightforward plug-in estimator of ρ is obtained by first estimating $\mu[n]$ and σ^2 via (7.8), and substituting these estimates into (7.45) to yield:

$$\widehat{\rho} \triangleq 10 \log_{10} \left(\|\widehat{\mu}[n]\|^2 / (N\widehat{\sigma^2}) \right). \tag{7.46}$$

To evaluate the performance of (7.46), we synthesized 32 ms waveforms of the phoneme [ɪ] for different values of SHNR ranging from -20 dB to 20 dB in increments of 5 dB. The value of $\mu[n]$ used during synthesis was obtained by projecting the Rosenberg flow onto the span of the first $r \in \{8, 16, 32, 64, 128\}$ Daubechies wavelets. For each SHNR ρ and wavelet number r an estimate of SHNR was obtained via (7.46); the corresponding results

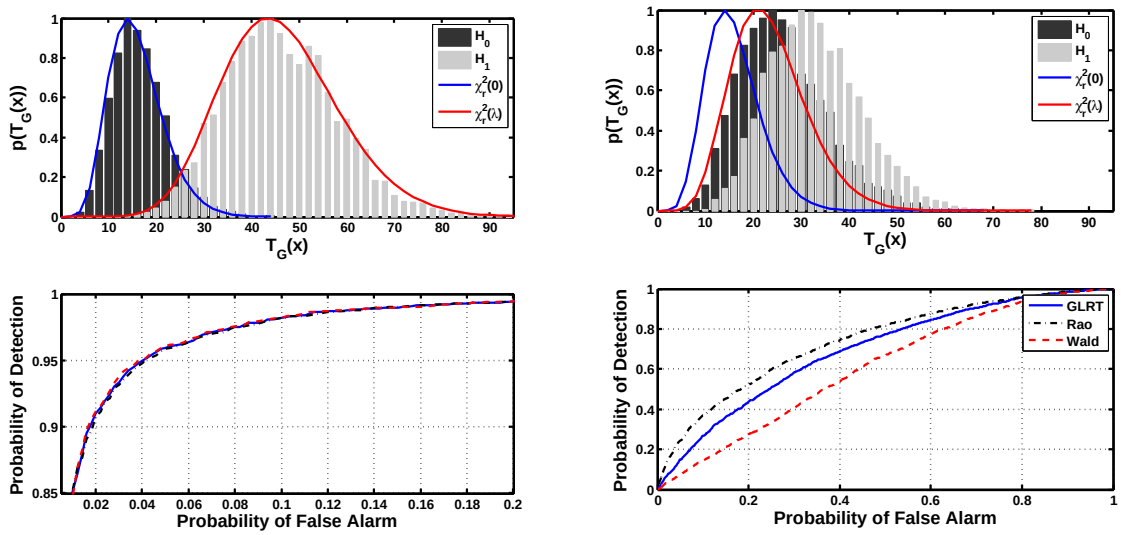


Figure 7.8: Detection performance of the GLRT (7.36), Wald (7.40), and Rao (7.42) tests for small and large sample size. The sampling distributions of the GLRT statistic under $\mathcal{H}_0$ and $\mathcal{H}_1$, computed for 5000 instantiations of the phoneme [ɛ], are shown overlaid with their associated asymptotic forms for large sample size (top-left, 64 ms) and sample size (top-right, 16 ms), respectively. The ROC curves corresponding to the GLR (solid, blue), Rao (dot-dashed, black), and Wald (dashed, red) tests are shown in both scenarios in the bottom two panels.

of 100 Monte Carlo trials are shown in Figure 7.10. Clearly, it is easier to estimate SHNR when there are less parameters to estimate in order to accurately capture the glottal flow (smaller number of wavelets) and when the underlying SHNR is large. Consequently, the most difficult cases correspond to estimating a complex model for $\mu[n]$ (128 parameters) in the greatest amount of noise.

7.7.2 Natural Speech Experiments

In the first experiment aimed at studying the behavior of the ARX model on recorded speech data, we estimated the vocal tract envelope and deterministic component of the glottal flow from 100 ms segments of the vowel [a] as in "father" recorded at 16 kHz and downsampled to 10 kHz prior to subsequent analysis. An ARX(10) model with $r = 128$ Daubechies 8 wavelets was fitted to the waveform via (7.8) and an AR(12) model was fitted to the waveform after pre-emphasis using the covariance method; the associated estimates are shown in Figure 7.11. The coherent and periodic structure of the estimated glottal flow $\widehat{\mu}[n]$ is transparent as shown in the top-right panel of Figure 7.11. It is important to note that the ARX(10) spectral estimate captures the shape of the power spectrum corresponding to the *pre-emphasized* waveform, but was obtained from the *original* waveform that was not pre-emphasized. Consequently, the spectral tilt due to radiation impedance at the lips has been adequately captured by $\widehat{\mu}[n]$, and no pre-emphasis is necessary as was the case with synthetic data. Although the residual of the ARX(10) fit is not white—the 128 wavelets employed did not capture the extremely sharp peaks at the glottal closures—its norm is

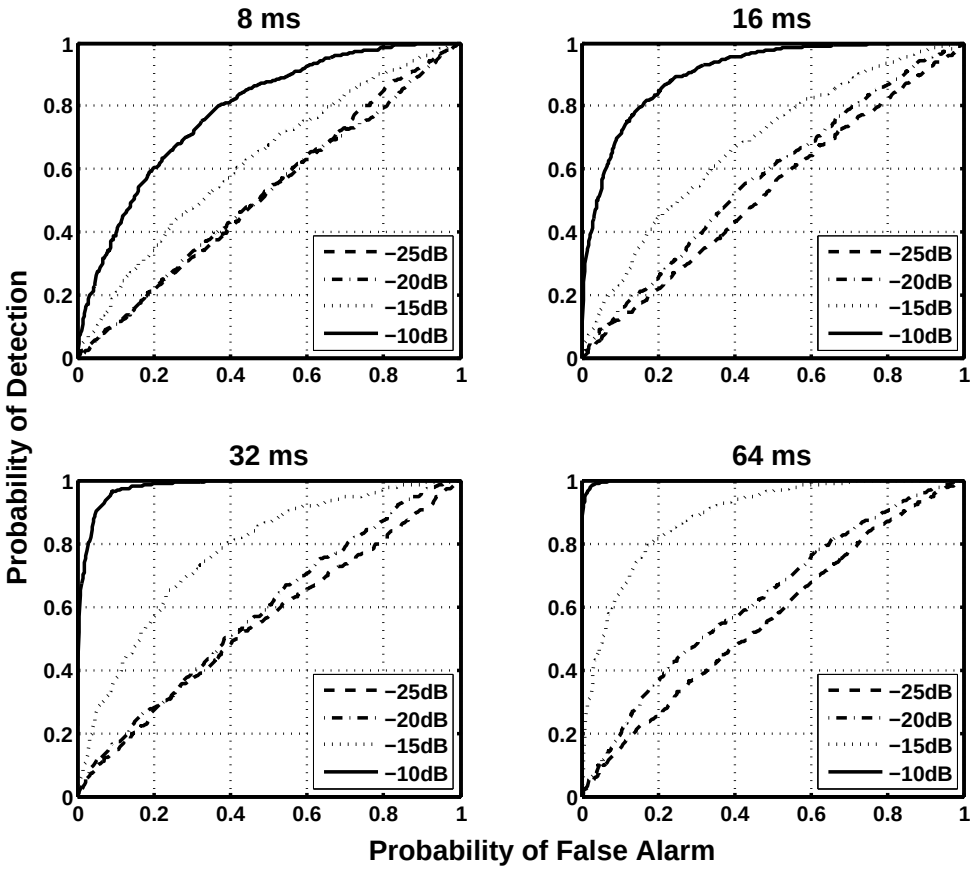


Figure 7.9: Detection performance of the GLRT. Receiver operating characteristic curves are shown for various signal lengths and SHNR values.

less than that of the corresponding AR(12) residual.

Next, we estimated the deterministic component of the glottal flow from two 75 ms segments of the vowel [ɪ] and fricative [f] recorded at 44.1 kHz and subsequently downsampled to 16 kHz both shown in Figure 7.12. The estimators of (7.8) and (7.13) used $p = 16$ AR coefficients and $q = 128$ Daubechies 4 wavelets—the glottal flow estimates are shown in the panels (ii) and (iv) of Figure 7.12. The estimates of $\mu[n]$ reveal its coherent and periodic structure in the case of the vowel, whereas the estimated source waveform is incoherent and noise-like in the case of the fricative. The right-hand side panels of Figure 7.12 show that the relative power of the voicing to aspiration is an order of magnitude higher, as expected, for the vowel. Further, its estimated $\mu[n]$ strongly resembles the glottal flow derivative and highlights the potential of using the ARX method for inverse filtering *without* needing to first identify the closed-phase.

As detailed in [236], estimating the SHNR ρ is a valuable non-invasive diagnostic tool in the clinical setting. To complement our experiments with synthetic speech, we estimated the SHNR on recordings of three vowels [ɑ], [ɪ] and [u] (as in father, bit and boot) produced by a female speaker using (a) a normal conversational voice and (b) a breathy

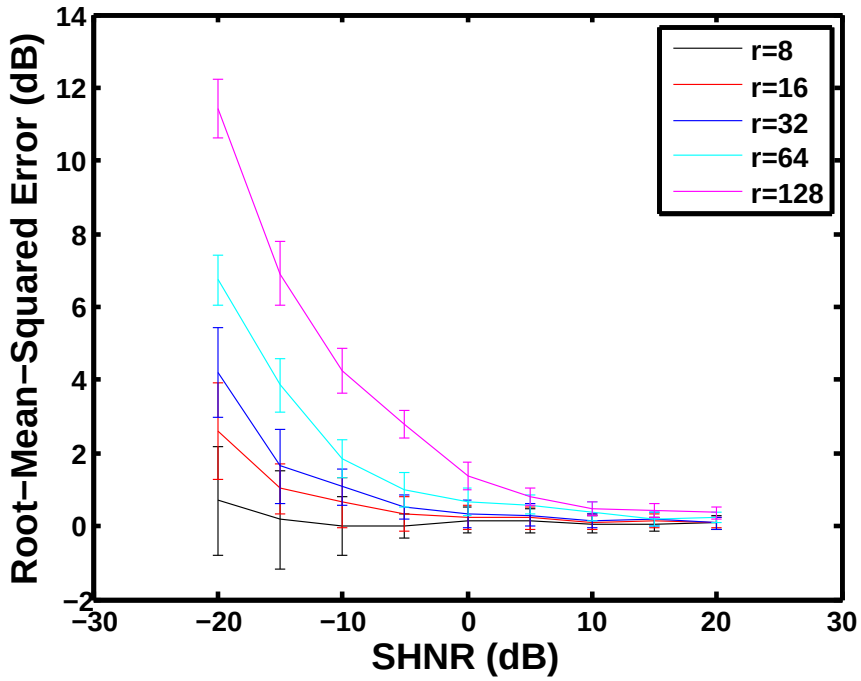


Figure 7.10: Performance of the plug-in SHNR estimator as a function of SHNR and number of wavelets r as evaluated via root mean-squared error.

voice with exaggerated aspiration, recorded at 44.1 kHz and downsampled to 16 kHz. The estimators of (7.8) and (7.13) used $p = 12$ AR coefficients and $q = 128$ Daubechies 4 wavelets; the resultant estimates were plugged into (7.46)—the results are summarized in Table 7.1. It can be seen that the SHNR values obtained via the ARX method are

	[ɑ]/	[ɪ]	[u]
Normal (dB)	27.3	22.1	36.5
Breathy (dB)	19.9	13.9	27.6

Table 7.1: SHNR estimates for normal and aspirated vowels

sensitive to the relative amount of aspiration and voicing present in these speech data since the SHNR values are consistently lower in the breathy vowel recordings. In addition, the reported SHNR values lie in the range for normal speakers [236], underscoring the promise of our approach—a comparative study of different SHNR estimation methods is beyond our current scope, but is planned for the future.

7.8 Extension to Time-Varying Autoregressions

Much of the material in this chapter can be naturally extended to allow for time-varying autoregressive coefficients. This could prove to be useful when working on time scales over which both the source waveform and the vocal tract configuration may

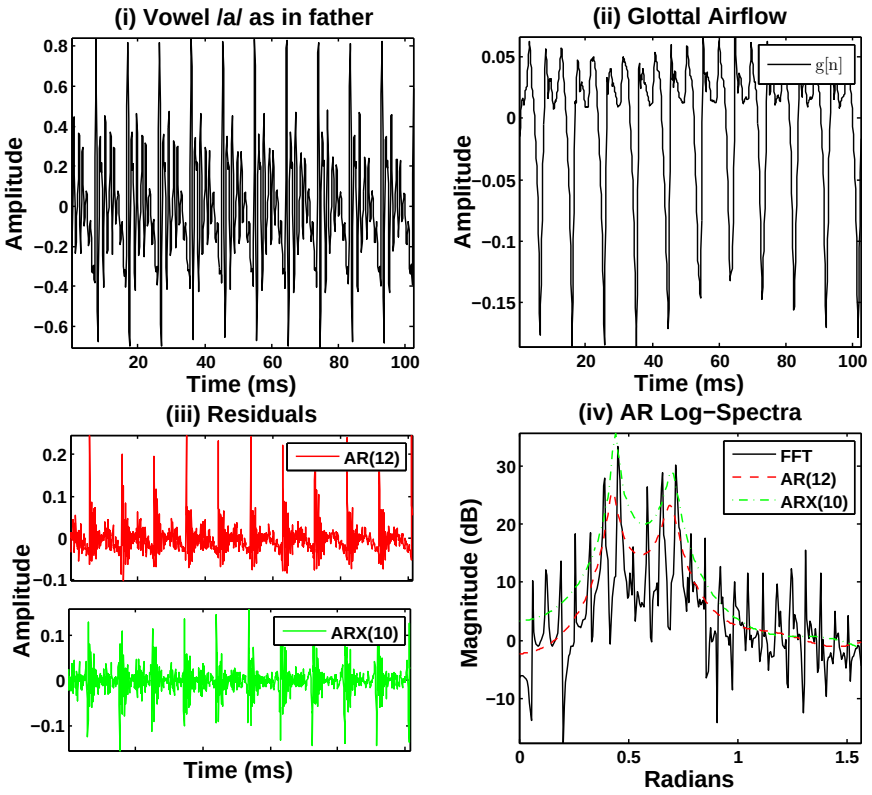


Figure 7.11: Analysis of a recorded vowel [a] using the ARX model: (i) recorded waveform [a]; (ii) estimated voiced component of glottal flow (solid, black); (iii) AR(12) (red) and ARX(10) (green) residuals; (iv) log-spectra for the fitted AR(12) (dashed-dot, red) and ARX(10) (dashed, green) models overlaid on the periodogram estimate of the power spectrum of the pre-emphasized waveform.

change. In this section, we precisely define this model and detail the associated (conditional) maximum-likelihood estimator. Then we discuss a number of interesting hypothesis testing and inference problems, which can form the basis for future work.

We define time-varying autoregressive model with exogenous input (TVARX) model as follows:

$$x[n] = \sum_{i=1}^{p} a_i[n]x[n-i] + \mu[n] + \sigma w[n], \tag{7.47}$$

where $w[n]$ is a white Gaussian sequence scaled by a gain $\sigma > 0$. Examination of (7.47) reveals that $x[n]$ is a Gaussian process with time-varying first and second moments. The time-varying mean is induced by the exogenous variable $\mu[n]$, whereas the time-varying correlation structure is induced by the temporal variation in the AR coefficients $a_i[n]$.

We consider a parametric version of (7.47) by specifying the evolution of each

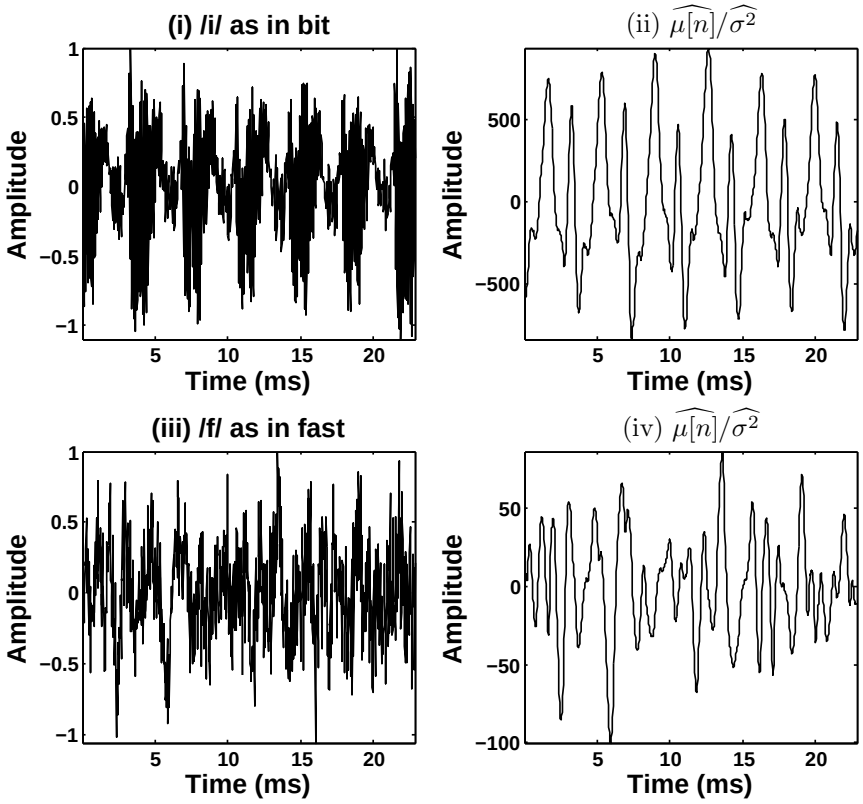


Figure 7.12: Analysis of a vowel [ɪ] and fricative [f] using the ARX model: waveforms of the vowel (i) and fricative (iii) with associated estimates of the voicing components in panels (ii) and (iv), respectively.

TVAR coefficient and the time-varying mean $\mu[n]$ via a functional expansion as follows:

$$a_i[n] = \sum_{j=0}^{q} \alpha_{ij} f_j[n] \qquad \text{and} \qquad \mu[n] = \sum_{k=1}^{r} \beta_k g_k[n], \tag{7.48}$$

where the $q+1$ functions $\{f_j[n] \,|\, 0 \le j \le q\}$ and r functions $\{g_k[n] \,|\, 1 \le k \le r\}$ are known a priori. In addition, we let the "constant" function $f_0[n] = 1$ be among the chosen set, so that the classical AR(p) model of (7.47) is recovered as $a_i \equiv \alpha_{i0} \cdot 1$ whenever $\alpha_{ij} = 0$ for all $j > 0$ and $\beta_k = 0$ for all $1 \le k \le r$. The model orders p, q and r are assumed known.

Consequently the model implied by (7.47) and (7.48) is parameterized by $p(q+1)+r+1$ parameters, which we gather in a vector $\boldsymbol{\theta}$ defined by:

$$\boldsymbol{\theta} \triangleq \begin{pmatrix} \boldsymbol{\alpha}_{\text{AR}} & \boldsymbol{\alpha}_{\text{TV}} & \boldsymbol{\beta}^T & \sigma^2 \end{pmatrix}^T, \tag{7.49}$$

where $\boldsymbol{\alpha}_{\text{AR}} \triangleq \boldsymbol{\alpha}_0^T$, $\boldsymbol{\alpha}_{\text{TV}} \triangleq \begin{pmatrix} \boldsymbol{\alpha}_1^T & \boldsymbol{\alpha}_2^T & \cdots & \boldsymbol{\alpha}_q^T \end{pmatrix}$, and the vectors $\boldsymbol{\alpha}_j$ are defined according to $\boldsymbol{\alpha}_j \triangleq \begin{pmatrix} \alpha_{1j} & \alpha_{2j} & \cdots & \alpha_{pj} \end{pmatrix}^T$ and $\boldsymbol{\beta} \triangleq \begin{pmatrix} \beta_1 & \beta_2 & \cdots & \beta_r \end{pmatrix}^T$.

The model specified by (7.47) and (7.48) is consistent with our earlier approach and retains the linear relationship between the parameters and observations. This allows for the specification of efficient maximum likelihood estimators as we show next.

7.8.1 Maximum Likelihood Estimation

Given N observations of a TVARX process of (7.47), partitioned according to

$$\boldsymbol{x} = (\boldsymbol{x}_p \,|\, \boldsymbol{x}_{N-p})^T \triangleq \big(x[0] \cdots x[p-1] \;\mid\; x[p] \cdots x[N-1]\big)^T,$$

the joint probability density function of $\boldsymbol{\theta}$ is given by:

$$p(\boldsymbol{x}\,;\boldsymbol{\theta}) = p(\boldsymbol{x}_{N-p} \,|\, \boldsymbol{x}_p\,;\boldsymbol{\theta}) p(\boldsymbol{x}_p\,;\boldsymbol{\theta}). \tag{7.50}$$

As before, we approximate (7.50) by the conditional likelihood $p(\boldsymbol{x}_{N-p} \,|\, \boldsymbol{x}_p\,;\boldsymbol{\theta})$. Gaussianity of $w[n]$ implies the conditional likelihood

$$p(\boldsymbol{x}_{N-p} \,|\, \boldsymbol{x}_p\,;\boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^{(N-p)/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{n=p}^{N-1}\left(x[n] - \sum_{i=1}^{p} a_i[n]x[n-i] - \mu[n]\right)^2\right),$$

Maximizing the associated log-likelihood is therefore equivalent to solving the following least-squares problem:

$$\boldsymbol{x}_{N-p} = \boldsymbol{H}_{\boldsymbol{x}}\boldsymbol{\alpha} + \boldsymbol{G}\boldsymbol{\beta} + \sigma\boldsymbol{w} = \big(\boldsymbol{H}_{\boldsymbol{x}} \,|\, \boldsymbol{G}\big)\begin{pmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{pmatrix} + \boldsymbol{w}, \tag{7.51}$$

where the $(n-p+1)$th row of the matrix $\boldsymbol{H}_{\boldsymbol{x}} \in \mathbb{R}^{(N-p)\times p(q+1)}$ is given by the Kronecker product $(x[n-1] \;\cdots\; x[n-p]) \otimes (f_0[n] \; f_1[n] \;\cdots\; f_q[n])$ for any $p \le n \le N-1$, the matrix $\boldsymbol{G}$ is defined as in (7.7), and $\boldsymbol{w}_{N-p} \triangleq \big(w[p] \quad w[p+1] \quad \cdots \quad w[N-1]\big)^T$.

It is easy to see from (7.51) that the (conditional) ML estimate of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is given by the pseudo-inverse of $(\boldsymbol{H}_{\boldsymbol{x}} \,|\, \boldsymbol{G})$ as in:

$$\begin{pmatrix}\widehat{\boldsymbol{\alpha}}\\ \widehat{\boldsymbol{\beta}}\end{pmatrix} = \big(\boldsymbol{H}_{\boldsymbol{x}} \,|\, \boldsymbol{G}\big)^{\#}\boldsymbol{x}_{N-p} = \begin{pmatrix}\boldsymbol{H}_{\boldsymbol{x}}^T\boldsymbol{H}_{\boldsymbol{x}} & \boldsymbol{H}_{\boldsymbol{x}}^T\boldsymbol{G}\\ \boldsymbol{G}^T\boldsymbol{H}_{\boldsymbol{x}} & \boldsymbol{G}^T\boldsymbol{G}\end{pmatrix}^{-1}\begin{pmatrix}\boldsymbol{H}_{\boldsymbol{x}}^T\\ \boldsymbol{G}^T\end{pmatrix}\boldsymbol{x}_{N-p}, \tag{7.52}$$

and that the associated variance estimator is given by:

$$\widehat{\sigma^2} = \frac{1}{N-p}\left(\boldsymbol{x}_{N-p} - \boldsymbol{H}_{\boldsymbol{x}}\widehat{\boldsymbol{\alpha}} - \boldsymbol{G}\widehat{\boldsymbol{\beta}}\right). \tag{7.53}$$

An example of fitting the TVARX model to synthetic data is shown in Figure 7.13. The sonorant [a ʊ] as in "how" was synthesized by shaping the glottal flow *derivative* using an all-pole filter with time-varying coefficients, and both an ARX(6) and a TVARX(6) model were fitted to the resultant waveform. Clearly, the TVARX(6) model is able to accurately capture the temporal evolution of the time-varying AR coefficients and the glottal flow derivative. On the other hand, the ARX(6) approach yields estimates that represent the average coefficient trajectories. As a result, the estimated source waveform attempts to capture the formant variation as can be seen in the top-right panel of Figure 7.13.

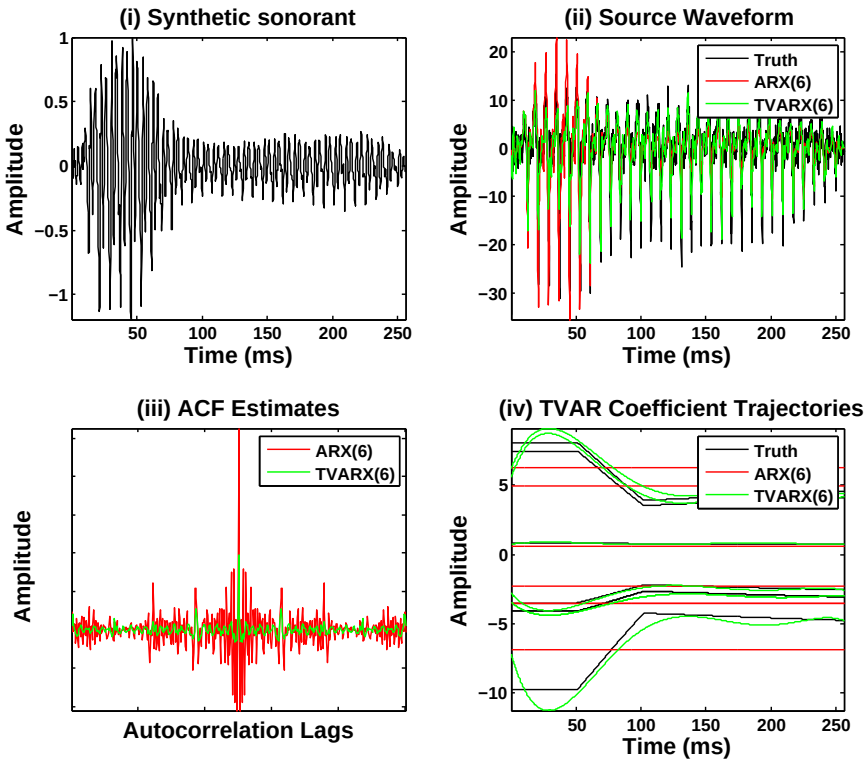


Figure 7.13: Comparison of ARX(6) and TVARX(6) on a synthetic sonorant [a ʊ] as in "how" with time-varying formants: (i) synthesized sonorant; (ii) voiced component of glottal flow (black) and its estimate obtained by ARX(6) (red) and TVARX(6) (green) models; (iii) autocorrelations of the ARX(6) (red) and TVARX(6) (green) residuals; (iv) true TVAR coefficient trajectories for the true TVARX(6) (black) and fitted ARX(6) (red) and TVARX(6) (green) models.

7.8.2 Hypothesis Testing

The hypothesis testing approaches of Chapter 5 to testing for stationarity using TVAR modeling and of Section 7.6 for voicing detection can be naturally combined in the context of a TVARX model with parameter vector $\boldsymbol{\theta} = \{\boldsymbol{\alpha}_{\text{AR}}, \boldsymbol{\alpha}_{\text{TV}}, \boldsymbol{\beta}, \sigma^2\}$. Indeed four natural hypothesis testing problems suggest themselves, two of which we have addressed previously and two of which are new:

- Time-varying vocal tract, no voicing model (Chapter 5)

$$\begin{aligned} \mathcal{H}_0 &: \boldsymbol{\theta} = \left(\boldsymbol{\alpha}_{\text{AR}} \;\; \boldsymbol{\alpha}_{\text{TV}} = \mathbf{0}_{pq\times 1} \;\; \boldsymbol{\beta} = \mathbf{0}_{r\times 1} \;\; \sigma^2\right) \\ \mathcal{H}_1 &: \boldsymbol{\theta} = \left(\boldsymbol{\alpha}_{\text{AR}} \;\; \boldsymbol{\alpha}_{\text{TV}} \neq \mathbf{0}_{pq\times 1} \;\; \boldsymbol{\beta} = \mathbf{0}_{r\times 1} \;\; \sigma^2\right) \end{aligned} \tag{7.54}$$

- Time-varying vocal tract with voicing model

$$\begin{aligned} \mathcal{H}_0 &: \boldsymbol{\theta} = \left(\boldsymbol{\alpha}_{\text{AR}} \;\; \boldsymbol{\alpha}_{\text{TV}} = \mathbf{0}_{pq\times 1} \;\; \boldsymbol{\beta} \;\; \sigma^2\right) \\ \mathcal{H}_1 &: \boldsymbol{\theta} = \left(\boldsymbol{\alpha}_{\text{AR}} \;\; \boldsymbol{\alpha}_{\text{TV}} \neq \mathbf{0}_{pq\times 1} \;\; \boldsymbol{\beta} \;\; \sigma^2\right) \end{aligned} \tag{7.55}$$

- Voicing test, time-invariant vocal tract (Section 7.6)

$$\begin{aligned}\mathcal{H}_0 &: \boldsymbol{\theta} = \left(\boldsymbol{\alpha}_{\mathrm{AR}} \;\; \boldsymbol{\alpha}_{\mathrm{TV}} = \mathbf{0}_{pq\times 1} \;\; \boldsymbol{\beta} = \mathbf{0}_{r\times 1} \;\; \sigma^2\right) \\ \mathcal{H}_1 &: \boldsymbol{\theta} = \left(\boldsymbol{\alpha}_{\mathrm{AR}} \;\; \boldsymbol{\alpha}_{\mathrm{TV}} = \mathbf{0}_{pq\times 1} \;\; \boldsymbol{\beta} \neq \mathbf{0}_{r\times 1} \;\; \sigma^2\right)\end{aligned} \tag{7.56}$$

- Voicing test, time-varying vocal tract

$$\begin{aligned}\mathcal{H}_0 &: \boldsymbol{\theta} = \left(\boldsymbol{\alpha}_{\mathrm{AR}} \;\; \boldsymbol{\alpha}_{\mathrm{TV}} \;\; \boldsymbol{\beta} = \mathbf{0}_{r\times 1} \;\; \sigma^2\right) \\ \mathcal{H}_1 &: \boldsymbol{\theta} = \left(\boldsymbol{\alpha}_{\mathrm{AR}} \;\; \boldsymbol{\alpha}_{\mathrm{TV}} \;\; \boldsymbol{\beta} \neq \mathbf{0}_{r\times 1} \;\; \sigma^2\right)\end{aligned} \tag{7.57}$$

The estimators of (7.52) and (7.53) could be used to realize all the above hypothesis tests via the generalized likelihood ratio or Wald test statistics. The Rao test may also be implemented. Applying these hypothesis tests to speech analysis is an interesting direction for further research.

7.9 Summary

In this chapter we extended the classical linear prediction framework by incorporating model of source waveform via nonparametric wavelet regression in order to take into account its quasi-periodic nature during voicing. The resultant model admits efficient linear estimators for the vocal tract transfer function, glottal flow and aspiration noise power, and exhibits robustness to pitch variation. In addition to summarizing the asymptotic properties of the ML estimator and studying the associated Cramer-Rao bound for the case of a fixed subspace for modeling $\mu[n]$, we introduced a number of online, data-dependent subspace selection algorithms. We also considered the problem of detecting the presence of voicing via hypothesis tests realized via the GLR, Wald and Rao test statistics and explored their detection performance. Evaluation using synthetic and recorded speech highlights the advantages afforded by our method and points the way toward using it for inverse filtering as well as for time-domain SHNR estimation—both of value in clinical applications.

7.A Appendix: Additional Synthetic Examples

In this appendix, we gather a number of additional experiments illustrating the application of the ARX modeling framework to a broad range of synthetic speech waveforms. All experiments are performed in the same manner as in Section 7.7.1.1, however, in order to demonstrate the versatility of the approach a different glottal airflow synthesizer was used. In particular, the implementation of the Liljencrantz-Fant (LF) model from the Voicebox speech processing toolbox for MATLAB [172] was employed in generating the glottal airflow waveforms and the associated derivatives. We describe all the experiments below, the results consistently show the robustness of the ARX framework.

The first two examples are included in order to illustrate the application of the ARX framework to short data records. In both cases approximately one pitch period (16 ms) of the vowel [a] as in "father" was synthesized at 16 kHz using the LF glottal airflow and its derivative as shown in the top-left panels of Figures 7.14 and 7.15, respectively. An ARX(6)

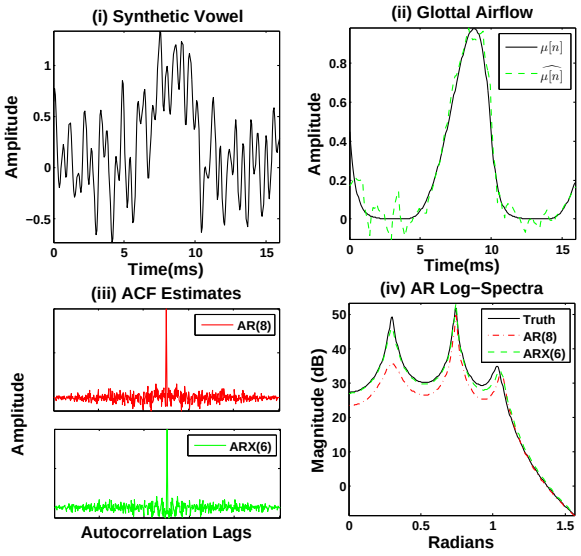


Figure 7.14: (i) synthesized waveform; (ii) voiced component of glottal flow (solid, black) and its estimate (dashed, green); (iii) autocorrelations of the AR(8) (red) and ARX(6) (green) residuals; (iv) log-spectra for the true AR(6) (solid, black) and fitted AR(8) (dashed-dot, red) and ARX(6) (dashed, green) models.

model with 32 Daubechies 6 wavelets was fit to the data using the least-squares estimator of (7.8); an AR(8) model was fitted using the covariance method after pre-emphasis. In first case, the ARX and AR spectral estimates incurred 1.09 dB and 4.07 dB error in LS distance, respectively. In the second case, the errors were 1.15 dB and 3.49 dB LS distance, respectively.

The next four examples, shown in Figures 7.16, 7.17, 7.18, and 7.19, illustrate the performance of the ARX framework on the vowel ɪ synthesized at a fundamental frequency of 100 Hz, 150 Hz, 200 Hz, and 250 Hz, respectively. The associated LS distances obtained by ARX(6)/AR(8) fits are (1.21/3.30) Hz, (1.73/2.25) Hz, (1.25/2.14) Hz, and (1.26/2.61) Hz, respectively.

The last three examples, shown in Figures 7.20, 7.21, and 7.22, illustrate the performance of the ARX framework on the vowels [ɪ] as in "bit", [o] as in "bot", and [u] as in "but", respectively. All vowels were synthesized at a fundamental frequency of 200 Hz. The associated LS distances obtained by ARX(6)/AR(8) fits are (1.25/2.14) Hz, (1.35/2.44) Hz, and (1.98/2.54) Hz, respectively.

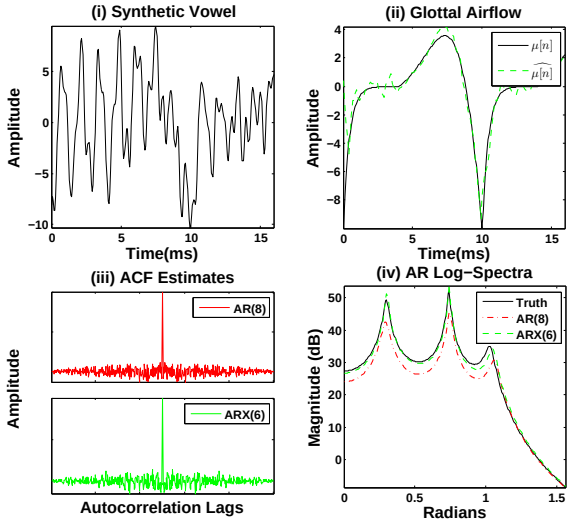


Figure 7.15: Comparison of ARX(6) and AR(8) models on one pitch period of a synthetic vowel æ as in "bat": (i) synthesized waveform; (ii) voiced component of glottal flow derivative (solid, black) and its estimate (dashed, green); (iii) autocorrelations of the AR(8) (red) and ARX(6) (green) residuals; (iv) log-spectra for the true AR(6) (solid, black) and fitted AR(8) (dashed-dot, red) and ARX(6) (dashed, green) models.

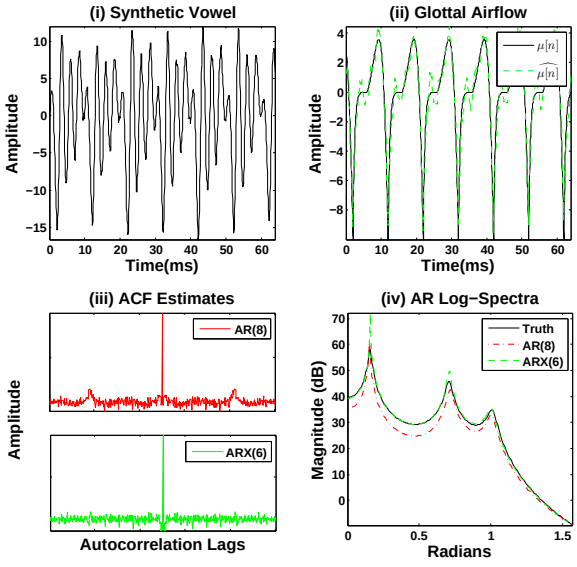


Figure 7.16: Comparison of ARX(6) and AR(8) models on a synthetic vowel with a constant pitch contour 100 Hz: (i) synthesized phoneme [ɪ]; (ii) voiced component of glottal flow (solid, black) and its estimate (dashed, green); (iii) autocorrelations of the AR(8) (red) and ARX(6) (green) residuals; (iv) log-spectra for the true AR(6) (solid, black) and fitted AR(8) (dashed-dot, red) and ARX(6) (dashed, green) models.

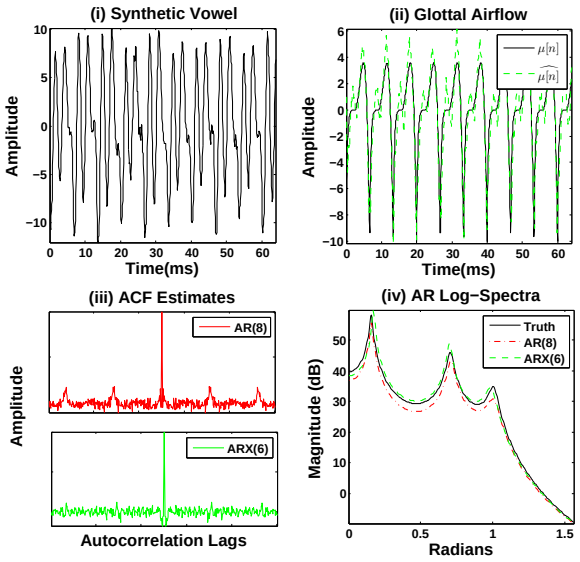


Figure 7.17: Comparison of ARX(6) and AR(8) models on a synthetic vowel with a constant pitch contour 150 Hz: (i) synthesized phoneme [ɪ]; (ii) voiced component of glottal flow (solid, black) and its estimate (dashed, green); (iii) autocorrelations of the AR(8) (red) and ARX(6) (green) residuals; (iv) log-spectra for the true AR(6) (solid, black) and fitted AR(8) (dashed-dot, red) and ARX(6) (dashed, green) models.

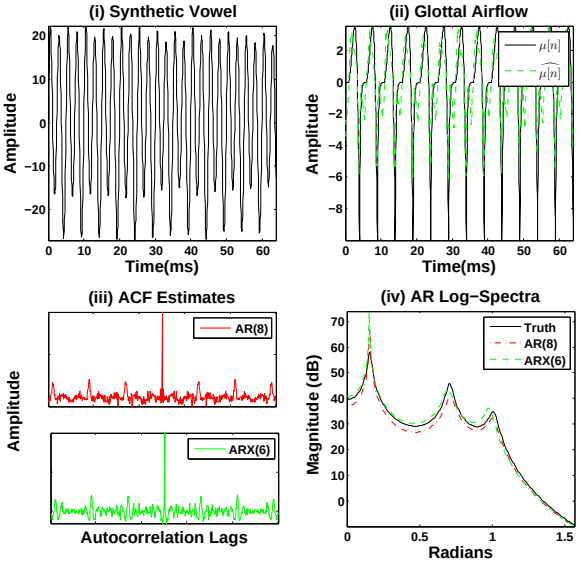


Figure 7.18: Comparison of ARX(6) and AR(8) models on a synthetic vowel with a constant pitch contour 200 Hz: (i) synthesized phoneme [ɪ]; (ii) voiced component of glottal flow (solid, black) and its estimate (dashed, green); (iii) autocorrelations of the AR(8) (red) and ARX(6) (green) residuals; (iv) log-spectra for the true AR(6) (solid, black) and fitted AR(8) (dashed-dot, red) and ARX(6) (dashed, green) models.

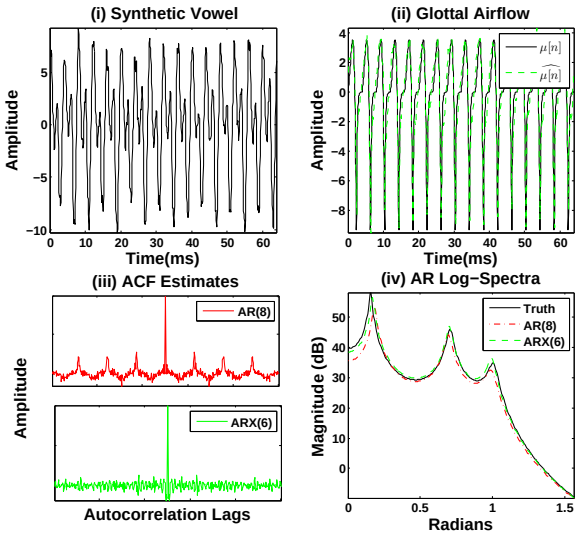


Figure 7.19: Comparison of ARX(6) and AR(8) models on a synthetic vowel with a constant pitch contour 250 Hz: (i) synthesized phoneme [ɪ]; (ii) voiced component of glottal flow (solid, black) and its estimate (dashed, green); (iii) autocorrelations of the AR(8) (red) and ARX(6) (green) residuals; (iv) log-spectra for the true AR(6) (solid, black) and fitted AR(8) (dashed-dot, red) and ARX(6) (dashed, green) models.

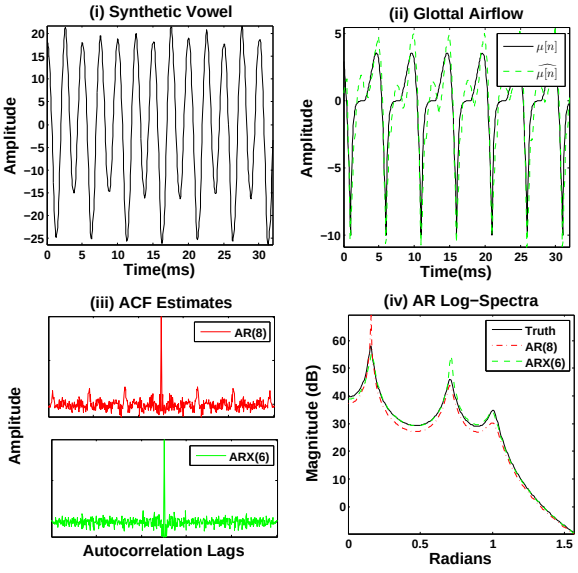


Figure 7.20: Comparison of ARX(6) and AR(8) models on a synthetic vowel ɪ as in "bit": (i) synthesized phoneme; (ii) voiced component of glottal flow (solid, black) and its estimate (dashed, green); (iii) autocorrelations of the AR(8) (red) and ARX(6) (green) residuals; (iv) log-spectra for the true AR(6) (solid, black) and fitted AR(8) (dashed-dot, red) and ARX(6) (dashed, green) models.

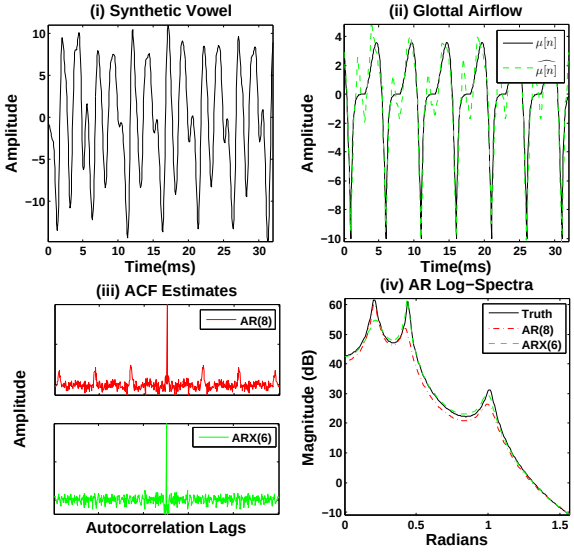


Figure 7.21: Comparison of ARX(6) and AR(8) models on a synthetic vowel o ʊ as in "boat": (i) synthesized phoneme; (ii) voiced component of glottal flow (solid, black) and its estimate (dashed, green); (iii) autocorrelations of the AR(8) (red) and ARX(6) (green) residuals; (iv) log-spectra for the true AR(6) (solid, black) and fitted AR(8) (dashed-dot, red) and ARX(6) (dashed, green) models.

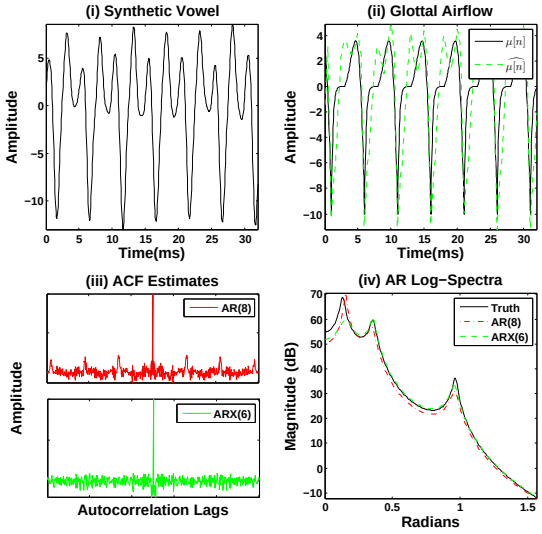


Figure 7.22: Comparison of ARX(6) and AR(8) models on a synthetic vowel ʊ as in "book": (i) synthesized phoneme; (ii) voiced component of glottal flow (solid, black) and its estimate (dashed, green); (iii) autocorrelations of the AR(8) (red) and ARX(6) (green) residuals; (iv) log-spectra for the true AR(6) (solid, black) and fitted AR(8) (dashed-dot, red) and ARX(6) (dashed, green) models.

Chapter 8

Conclusions and Future Directions

In this chapter we first review the major themes and primary contributions of this thesis. In addition, we outline a number of open and interesting questions that remain for future investigation.

8.1 Summary of Contributions

Many real-world time-series such as speech and music waveforms exhibit a degree of controlled nonstationarity, and accurate models of the associated temporal variations are extremely valuable for a wide range of applications. Motivated by the physiology and acoustics of speech production, we developed statistical methods for the analysis of nonstationary time series and applied them to a variety of problems arising in speech signal processing. Throughout, we have aimed to design methods that are simultaneously computationally-efficient, can be applied to many different problems in speech analysis, and whose properties can be formally analyzed.

The exploration of these themes has led to several major contributions:

- To improve and extend existing formant tracking methods;
- To develop the theory of time-varying linear prediction and apply time-varying autoregressive models to a number of problems in speech analysis;
- To propose parametric and nonparametric hypothesis tests for stationarity;
- To extend standard short-time Fourier methods by developing signal-adaptive, linear time-frequency representations that admit fast reconstruction;
- To describe a semiparametric approach for jointly modeling the source waveform and vocal tract.

Formant tracking was addressed in Chapter 3, where we developed a generative probabilistic model for formant evolution that enabled improved modeling of spectral valleys, the incorporation of a speech activity detector, and the application of proper system identification methods for parameter estimation. The analysis of time-varying autoregressive models formed the basis of Chapter 4 and the first half of Chapter 5. The contributions

contained therein advance our understanding of these models with respect to statistical questions such as covariance estimation and parameter estimation from noisy data, system-theoretic issues related to time-varying lattice filters and stability, and a broader exploration of different classes of TVAR models. In the first part of Chapter 5, we sought to determine through a hypothesis testing framework, whether TVAR models can accurately capture the time-varying nature of vocal tract.

In the second half of Chapter 5, we consider alternatives to purely parametric approaches based on autoregressive modeling. After developing a nonparametric counterpart to the TVAR-based GLRT based on empirical short-time Fourier coefficients, we generalize the standard fixed-resolution time-frequency analysis and propose a family of novel signal-adaptive representations, termed superposition frames, whose properties allow their successful application in practice. Finally in Chapter 7, we augment the parametric AR model for the vocal tract with a nonparametric regression model for the source waveform resulting in a semiparametric realization of the source-filter model.

Throughout, we apply the developed methodology to various problems in speech signal processing: formant tracking in Chapter 3; speech modeling and synthesis in Chapter 3; formant change detection, glottal opening instant estimation, glottal closing instant estimation, and audio enhancement in Chapter 5; speech enhancement in Chapter 6; and vocal tract and source-harmonics-to-noise ratio estimation, inverse filtering, and voicing detection in Chapter 7. However, we feel that many of the proposed methods can be be applied beyond speech processing to other domains where analysis of nonstationary time series is required.

8.2 Suggestions for Future Research

We conclude by discussing a number of interesting, open questions raised in this thesis, and suggesting ways in which they may be addressed. In addition, we highlight further promising applications of the developed methodologies.

8.2.1 Time-Varying Autoregressive Models and Properties

As discussed in Chapters 4 and 5, time-varying autoregressive models have been widely used beyond speech processing, in a wide variety of applications ranging from blind source separation [101,103–106] and radar signal processing [107] to biomedical signal analysis [109–111], and instantaneous frequency estimation [112–114]. Consequently, it is important to further develop our understanding of their properties. We have raised two specific open problems described below.

Exact ML Estimation Many of the algorithms for estimating AR coefficients in the time-invariant case have been generalized to the time-varying setting. Generalizations of the covariance (Section 2.4.1) and autocorrelation (Section 2.4.3) methods of linear prediction were described in Sections 4.3.1 and 4.3.2, respectively. Moreover, in Section 4.4 we described how lattice-based estimators can be generalized to the time-varying setting with and without frozen-time stability constraints, and, in Section 4.3.3, we proposed a

TVAR coefficient estimator from noise-corrupted observations by generalizing the Kalman-EM approach of [128, 129]. However, all these estimators are based on approximating the unconditional likelihood of TVAR(p) parameters $p(\boldsymbol{x}\,;\,\boldsymbol{\alpha},\sigma^2)$ by the conditional likelihood $p(\boldsymbol{x}_{N-p}\,|\,\boldsymbol{x}_p\,;\,\boldsymbol{\alpha},\sigma^2)$, thereby maximizing the likelihood with respect to only $N-p$ out of N observations.

This approximation is reasonable when $N >> p$, but in situations when little data is available and a high-order AR model is used, the exact ML estimator is preferable. In the time-invariant case, the exact ML estimate may be obtained by the methods described in Section 2.4.2, but none have been generalized to the time-varying setting—doing so is an open, but somewhat intricate algebraic question. The most promising way forward seems to be to generalize the approach of Scharf [33] by relying on the Dym-Gohberg formulas for the inverse of a block-Toeplitz matrix, as described in [123], in lieu of the Gohberg-Semencul formulas for the inverse of a Toeplitz covariance matrix that Scharf uses. An alternative approach based on generalizing the order-recursive procedure of [32] may also be possible. The development of an exact ML estimator for TVAR models is important for applications such as detecting glottal opening and closing instants, as described in Chapter 5, since little data is available in a single pitch period of a typical speech waveform.

TVAR Models and Differential Geometry In Section 4.6, we considered TVAR models from a geometric perspective, by viewing the TVAR coefficient trajectories as defining a path on the space of AR processes. Loosely speaking, this gives us a way of measuring the "degree" of nonstationarity—the longer the path "length" of a process, the more nonstationary the associated process. This idea formed the basis of our approach to modeling the temporal evolution of TVAR coefficients—whereby the overall coefficient variation was regularized by bounding a measure of distance between induced AR models along the TVAR path. Consequently, the inverse problem of estimating TVAR coefficients from data could be approached by minimizing a least-squares criterion subject to an overall path-length constraint. However, the solubility of the resultant problem strongly depends on the form of the relationship between the path-length and the AR coefficients.

When path-length is measured via an ℓ_q-norm-derived distance between adjacent AR coefficient vectors as in (4.70), for instance, then estimating the TVAR coefficients from a time series of observations results in a convex optimization problem, since the appropriate objective function is quadratic in the AR coefficients and the constraints are convex. However, as discussed in Section 4.6.4, it is more natural to use a measure of distance intrinsic to the space of AR processes—we discussed a number of Riemannian metrics, other choices are detailed in [146]. Unfortunately, the relationship between path length, induced by such metrics (see, e.g., (4.83)), and the associated AR coefficients is highly nonlinear—the metric depends critically on the covariance structure of the AR process, which is a nonlinear function of the AR coefficients. This, in turn, causes difficulty in solving for the associated TVAR process. The question of how to employ any Riemannian metric in lieu of (4.70) in order to define classes of nonstationary TVAR processes that are amenable to efficient estimation procedures remains open. Any solution, if it exists, would likely require a combination of a judiciously-chosen metric and clever optimization techniques.

8.2.2 Speech Enhancement

A number of techniques presented in this thesis can be applied to the problem of single-channel speech enhancement. Speech enhancement algorithms are often judged on the basis of their ability to improve both speech quality and speech intelligibility. Most of the well-known algorithms have been shown to improve waveform quality, but have not been as successful in improving its intelligibility [237, 238].

Although this latter goal remains an open problem, recent studies [238, 239] have suggested that improvements in intelligibility may be obtained by reliable estimation of SNR in each time-frequency (TF) unit of a joint TF signal representation. Accurate modeling of the time-varying covariance structure of the speech waveform, in the manner of this thesis, is certainly consistent with this goal. Thus, we describe two potential approaches to speech enhancement based on the contributions of this thesis.

Time-Domain Enhancement using TVAR modeling The first approach is a time-domain parametric method based on modeling speech using time varying autoregressions, as described in Chapters (4) and (5). Since only noisy observations are available in the enhancement setting, the Kalman-EM parameter estimation algorithm of Section 4.3.3 should be used for parameter estimation.

If only a constant function were used to model the temporal trajectories of the AR coefficients (i.e., $q = 0$ in (4.25)), then the estimation framework reduces to the enhancement algorithms described in [128] and [129]. However, TVAR(p) modeling offers the potential of more accurately preserving transitional information between phonemes, and may allow for the use of of longer segments than in the traditional enhancement setting. Using an adaptive segmentation algorithm (e.g., based on hypothesis testing as in Chapters 5 or dynamic programming as in 6) to identify variable-length segments for subsequent TVAR modeling and enhancement seems appropriate. Finally, a clever initialization procedure for the Kalman-EM algorithm will be required in order to obtain good performance results in low-SNR (< 5 dB) regimes.

Adaptive Nonparametric Enhancement using Superposition Frames Another family of approaches was suggested in Chapter 6, whereby a signal-adaptive variable-length time-frequency representation (i.e., a superposition frame) can be used together with a statistical shrinkage rule (e.g., spectral subtraction [237]) to efficiently enhance noisy signals in the transform domain. Indeed, our initial experiments and listening tests yielded promising results as described in Chapter 6 and support further investigation. The superposition frames framework offers much flexibility with respect to the signal-adaptation rule and choice of suppression or shrinkage rule; it would be valuable to investigate a variety of possible choices, and to evaluate them specifically with respect to improving speech quality. Furthermore, since adapting the time-frequency plane to the time-varying content of the underlying waveform should serve to improve SNR in each time-frequency bin, the resultant enhancement algorithms may also lead to improve the intelligibility of the enhanced waveform. Finally, it would be interesting to compare enhancement based on superposition frames with the recent approach of [155], since these authors estimate the clean speech

periodogram—required for the shrinkage rule—based on an adaptive segmentation of the signal, but apply the scheme using a fixed-resolution short-time analysis.

8.2.3 Semiparametric Source-Filter Modeling

In Chapter 7, we introduced a novel semiparametric model of the speech waveform, whereby the vocal tract (filter) was modeled parametrically using an AR(p) process, and the source waveform (glottal airflow volume velocity or its derivative) was modeled nonparametrically using wavelet regression. A number of important questions remain for future investigation.

Subspace Selection On the methodological side, the question of subspace selection, as outlined in Section 7.5, needs to be further addressed. Recall that the ARX(p) model posits that an N-sample waveform is modeled as a convolution of a source waveform, modeled via an order-r wavelet regression, and an order-p autoregressive system. If the number of wavelets r were set to the number of samples N, then source waveforms of all shapes could be modeled perfectly, but the inverse problem would be ill-posed since more variables than data points would need to be estimated. Indeed, the maximum likelihood estimator of (7.9) cannot be used in this case, since $P_{\boldsymbol{G}}^{\perp}\boldsymbol{X} = \boldsymbol{0}$. Consequently, it is necessary to select judiciously $r << N$ functions from a basis or dictionary for modeling the source waveform.

To this end, in Section 7.5.2, we proposed a number of signal-adaptive algorithms for subspace selection based on hard shrinkage, soft shrinkage, and ℓ_1-optimization. The initial results are promising, as demonstrated by the examples in Section 7.7, but further study of the properties of these algorithms is warranted; proving the convergence of the iterative shrinkage algorithms to a fixed point and establishing consistency properties in the asymptotic regime are two problems of obvious interest. The former question may be amenable to the approach of Fadili-Bullmore [233], though the presence of an autoregressive component significantly complicates the required analysis of both questions.

Applications to Speech Analysis Since the ARX(p) model of Chapter 7 is an encapsulation of the source-filter model of the speech production system, there is a wide range of specific problems to which the associated methodology can be applied. We have already demonstrated a number of potential applications such as improved vocal tract estimation, inverse filtering and source harmonics-to-noise ratio (SHNR) estimation, but much more can be done as we describe below.

In a clinical setting, a problem of great interest is to find acoustic correlates of features symptomatic of vocal disorders to aid their diagnosis and characterization. Breathiness—roughly, the perceived ratio of aspiration to periodicity in a voiced utterance—is an important example, and has been found to be a symptom of organic and functional voice disorders [240]. An interesting question is to determine whether the source harmonics to noise ratio, defined by (7.45) and estimated by fitting the ARX(p) model to data can yield a reliable acoustic correlate of breathiness. Conducting perceptual studies along the lines of [240–243], with the goal of finding statistically significant correlation between the estimated SHNR and the perceptual rating of breathiness on a numerical scale by trained

listeners, would be of great interest. In addition, estimated SHNR can be used to construct a voicing detector (it is essentially equivalent to the Wald test statistic described in Section 7.6.2, which is useful in applications such as speech coding and recognition.

For many applications, it is of great interest to estimate the source waveform itself rather than features derived from it such as SHNR. Accurately estimating the source waveform has been a long-standing open problem in speech processing relevant to multiple applications (see e.g., [244] for a recent survey). Recall that the ARX(p) model of Chapter 7 uses wavelet regression to model the source waveform, where *a priori* estimation of pitch and glottal closing and opening instants is not necessary. This makes the proposed methodology attractive for signals with time-varying pitch and, especially in the clinical setting, for waveforms with highly irregular timings of pitch periods. It would be interesting to compare the proposed approach to existing methods including closed-phase inverse filtering [244] and various model-based approaches [20–22, 43], though the evaluation of all these methods is confounded by the inability to non-invasively measure the glottal volume velocity waveform in practice. The most promising approaches may employ perceptual and multi-modal (i.e., obtained by multiple sensors such as electroglottagraphs or high-speed videoendoscopy systems) correlates to the source waveform to validate the results.

Finally, recall that the ARX(p) model allows the source waveform to be modeled over multiple pitch periods, which in turn suggests to fit the model to longer speech segments. However, at longer time scales, it is quite possible that the vocal tract is also time-varying, which suggests applying the TVARX(p) model described in Section 7.8. The discussion contained therein demonstrates that the resultant estimation methods are still linear and can be efficiently implemented, and that hypothesis tests to verify whether such a complex model is warranted are easily realized. This offers interesting possibilities with respect to precise analysis of the source waveform, and by proxy the behavior of the vocal folds, during sounds other than sustained vowels typically used in the clinical setting.

Bibliography

[1] L. Deng, A. Acero, and I. Bazzi, "Tracking vocal tract resonances using a quantized nonlinear function embedded in a temporal constraint," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, pp. 425–434, 2006.

[2] L. Deng, L. J. Lee, H. Attias, and A. Acero, "Adaptive Kalman filtering and smoothing for tracking vocal tract resonances using a continuous-valued hidden dynamic model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, pp. 13–23, 2007.

[3] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, Upper Saddle River, NJ: Prentice-Hall, 2002.

[4] S. Kay, *Modern Spectral Estimation: Theory and Application*, Upper Saddle River, NJ: Prentice-Hall, 1988.

[5] P. Flandrin, *Time-Frequency Time-Scale Analysis*, Academic, London, U.K., 1998.

[6] M. R. Portnoff, "A quasi-one-dimensional digital simulation for the time-varying vocal tract," M.S. thesis, Massachusetts Institute of Technology, 1973.

[7] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, 1978.

[8] J. D. Markel and Jr. A. H. Gray, *Linear Prediction of Speech*, Springer-Verlag: New York, 1976.

[9] H. M. Hanson, K. N. Stevens, H.-K. J. Kuo, M. Y. Chen, and J. Slifka, "Towards models of phonation," *J. Phonetics*, vol. 29, pp. 451–480, 2001.

[10] K. Ishizaka and M. Matsudaira, "What makes the vocal cords vibrate.," in *Proc. Sixth Intl. Cong. Acoust., Tokyo, Japan*, 1968, pp. B9–B12.

[11] K. Ishizaka and J. L. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords," *Bell System Technical Journal*, vol. 51, pp. 1233–1268, 1972.

[12] B. H. Story and I. R. Titze, "Voice simulation with a body-cover model of the vocal folds," *J. Acoust. Soc. Am.*, vol. 97, pp. 1249–1260, 1995.

[13] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *J. Acoust. Soc. Am.*, vol. 49, pp. 583–590, 1971.

[14] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *Speech Transmission Lab. Quart. Prog. Status Rep., Royal Institute of Technology, Stockholm, Sweden*, vol. 4, pp. 1–13, 1985.

[15] P. Hedelin, "A glottal LPC-vocoder," in *Proc. IEEE Intl. Conf. Acoust., Speech, Signal Process.*, 1984, pp. 21–24.

[16] H. Fujisaki and M. Ljungqvist, "Proposal and evaluation of models for the glottal source waveform," in *Proc. IEEE Intl. Conf. Acoust., Speech, Signal Process.*, 1986, pp. 1605–1608.

[17] H. Fujisaki and M. Ljungqvist, "Estimation of voice source and vocal tract parameters based on ARMA analysis and a model for the glottal source waveform," in *Proc. IEEE Intl. Conf. Acoust., Speech, Signal Process.*, 1987, pp. 637–640.

[18] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Am.*, vol. 87, pp. 820–857, 1990.

[19] R. Veldhuis, "A computationally efficient alternative for the Liljencrants-Fant model and its preceptual evaluation," *J. Acoust. Soc. Am.*, vol. 103, pp. 566–571, 1998.

[20] Y. Miyanaga, N. Miki, N. Nagai, and K. Hatori, "A speech analysis algorithm which elminates the influence of pitch using the model reference adaptive system," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 30, pp. 88–96, 1982.

[21] P. Milenkovic, "Glottal inverse filtering by joint estimation of an AR system with a linear input model," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 34, pp. 28–42, 1986.

[22] M. M. Thomson, "A new method for determining the vocal tract transfer function and its excitation from voiced speech," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.*, 1992, pp. 23–26.

[23] M. S. Howe, *Theory of Vortex Sound*, Cambridge University Press, Cambridge UK, 2003.

[24] D. Mehta and T. F. Quatieri, "Pitch-scale modification using the modulated aspiration noise source," in *Proc. 9th Ann. Conf. Intl. Speech Commun. Ass.*, 2006, pp. 2490–2493.

[25] T. W. Anderson, *The Statistical Analysis of Time Series*, John Wiley and Sons, 1971.

[26] J. D. Hamilton, *Time Series Analysis*, Princeton University, 1994.

[27] B. Friedlander, "Lattice filters for adaptive processing," *Proc. IEEE*, vol. 8, pp. 829–866, 1982.

[28] M. Kendall, A. Stuart, J. K. Ord, and S. Arnold, *Kendall's Advanced Theory of Statistics*, vol. 2a, Hodder Arnold, 1999.

[29] J. Makhoul, "Linear prediction: A tutoral review," *Proc. IEEE*, vol. 63, pp. 561–581, 1975.

[30] P. Kabal and R. P. Ramachandran, "The computation of line spectral frequencies using Chebyshev polynomials," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 34, pp. 1419–1427, 1986.

[31] S. L. Marple, *Digital Spectral Analysis*, Englewood Cliffs, NJ: Prentice-Hall, 1987.

[32] S. M. Kay, "Recursive maximum likelihood estimation of autoregressive processes," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 31, pp. 56–66, 1983.

[33] L. T. McWhorter and L. Scharf, "Nonlinear maximum likelihood estimation of autoregressive time series," *IEEE Trans. Signal Process.*, vol. 43, pp. 2909–2920, 1995.

[34] U. Grenander and G. Szego, *Toeplitz Forms and Their Applications*, Univ. of California, Berkeley, CA, 1958.

[35] B. Porat, *Digital Processing of Random Signals: Theory and Methods*, Upper Saddle River, NJ: Prentice-Hall, 1993.

[36] P. Stoica and T. Soderstrom, "Optimal instrumental variable estimation and approximate implementations," *IEEE Trans. Automat. Control*, vol. 28, pp. 757–772, 1983.

[37] J. Makhoul, "Stable and efficient lattice methods for linear prediction," *IEEE Trans. on Acoust. Speech and Signal Process.*, vol. 25, pp. 423–428, 1977.

[38] S. Kay and J. Makhoul, "On the statistical of the estimated reflection coefficients of an autoregressive process," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 31, pp. 1447–1456, 1983.

[39] C.-H. Lee, "On robust linear prediction of speech," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 36, pp. 642–650, 1988.

[40] R. P. Ramachandran, M. S. Zilovic, and R. J. Mammone, "A comparative study or robust linear predictive analysis methods with applications to speaker identification," *IEEE Trans. Speech Audio Process.*, vol. 3, pp. 117–126, 1995.

[41] P. Jinachitra and J. O. Smith III, "Joint estimation of glottal source and vocal tract for vocal synthesis using Kalman smoothing and EM algorithm," in *Proc. IEEE Worksh. Appl. Signal Process. Audio Acoust.*, 2005, pp. 327–330.

[42] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Trans. Speech Audio Process.*, vol. 7, pp. 569–586, 1999.

[43] Q. Fu and P. Murphy, "Robust glottal source estimation based on joint source-filter model optimization," *IEEE Trans. Audio, Speech Lang.*, vol. 14, pp. 492–501, 2006.

[44] H. Akaike, "A new look at statistical model identification," *IEEE Trans. Autom. Contr.*, vol. 19, pp. 716–723, 1974.

[45] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[46] H. Akaike, "A Bayesian extension of the minimum AIC procedure of autoregressive model fitting," *Biometrika*, vol. 19, pp. 237–242, 1979.

[47] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde, "Bayesian measures of model complexity and fit (with discussion)," *J. Roy. Stat. Soc. B*, vol. 64, pp. 583–639, 2002.

[48] J. S. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, Linguistic Data Consortium, Philadelphia, PA, 1993.

[49] K. Sjölander and J. Beskow, "WaveSurfer 1.8.5 for Windows," 2005.

[50] T. V. Ananthapadmanabha and G. Fant, "Calculation of true glottal flow and its components," *Speech Commun.*, vol. 1, pp. 167–184, 1982.

[51] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, 1984.

[52] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods*, Springer-Verlag, 1991.

[53] M. B. Priestley, "Evolutionary spectra and non-stationarity of time-series," *J. Roy. Stat. Soc. B*, vol. 27, pp. 204–229, 1965.

[54] D. Tjostheim, "Spectral generating operators for non-stationary processes," *Adv. Applied. Probab.*, vol. 8, pp. 821–846, 1976.

[55] H. Cramer, "On the structure of purely nondeterministic stochastic processes," *Ark. Mat.*, vol. 4, pp. 249–266, 1961.

[56] L. A. Zadeh, "Frequency analysis of variable networks," *Proc. I.R.E.*, vol. 28, pp. 291–299, 1950.

[57] Y. Grenier, *Parametric time-frequency representations, in: Traitement du Signal/Signal Processing*, pp. 339–397, Les Houches, Session XLV, North-Holland, Amsterdam, 1987.

[58] Y. Grenier, "Time-dependent ARMA modeling of nonstationary signals," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 31, pp. 899–911, 1983.

[59] J. A. Sills and E. W. Kamen, "On some classes of nonstationary parametric processes," *Proc. 29th Conf. Inf. Sci. Sys.*, pp. 136–191, 1995.

[60] J. A. Sills and E. W. Kamen, "On some classes of nonstationary parametric processes," *J. Frankl. Inst.*, vol. 337, pp. 217–249, 2000.

[61] N. Huang and J. K. Aggrawal, "On linear shift-variant digital filters," *IEEE Trans. Circuits Systems*, vol. 8, pp. 672–679, 1980.

[62] W. Kozek, "On the transfer function calculus for underspread LTV channels," *IEEE Trans. Signal Process.*, vol. 45, pp. 219–223, 1997.

[63] G. Matz, F. Hlawatsch, and W. Kozek, "Generalized evolutionary spectral analysis and the Weyl specturm of nonstationary processes," *IEEE Trans. Signal Process.*, vol. 45, pp. 1520–1534, 1997.

[64] K. N. Stevens, *Acoustic Phonetics*, Cambridge, MA: MIT Press, 1998.

[65] L. Deng and D. O'Shaughnessy, *A Dynamic and Optimization-Oriented Approach*, Marcel Dekker Inc., New York, NY, 2003.

[66] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Am.*, vol. 55, pp. 1304–1312, 1974.

[67] S. McCandless, "An algorithm for automatic formant extraction using linear prediction spectra," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 22, pp. 134–141, 1974.

[68] G. Rigoll, "A new algorithm for estimation of formant trajectories directly from the speech signal based on an extended Kalman filter," *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.*, pp. 1229–1232, 1988.

[69] G. Kopec, "Formant tracking using hidden Markov models and vector quantization," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 34, pp. 709–729, 1986.

[70] D. Talkin, "Speech formant trajectory estimation using dynamic programming with modulated transition costs," *J. Acoust. Soc. Am.*, vol. S1, pp. S55, 1987.

[71] T. M. Nearey, "Static, dynamic, and relational properties in vowel perception," *J. Acoust. Soc. Am.*, vol. 85, pp. 2088–2113, 1989.

[72] M. Niranjan, I. J. Cox, and S. Hingorani, "Recursive tracking of formants in speech signals," *Proc. IEEE Intl. Conf. on Acoust. Speech Signal Process.*, pp. 205–208, 1994.

[73] L. Welling and H. Ney, "Formant estimation for speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 6, pp. 36–48, 1998.

[74] Y. Zheng and M. Hasegawa-Johnson, "Formant tracking by mixture state particle filter," *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.*, vol. 1, pp. 565–568, 2004.

[75] S. Manocha and C. Y. Espy-Wilson, "Knowledge-based formant tracking with confidence measure using dynamic programming," *J. Acoust. Soc. Am.*, p. Abstract: 2pSC5, 2005.

[76] D. T. Toledano, J. G. Villardebó, and L. H. Gómez, "Initialization, training, and context-dependency in HMM-based formant tracking," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, pp. 511–524, 2006.

[77] K. Mustafa and I. C. Bruce, "Robust formant tracking for continuous speech with speaker variability," *IEEE Trans. on Audio Speech Lang. Process.*, vol. 14, pp. 435–444, 2006.

[78] L. Deng, X. Cui, R. Pruvenok, J. Huang, S. Momen, Y. Chen, and A. Alwan, "A database of vocal tract resonance trajectories for research in speech processing," *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.*, vol. 1, pp. 369–372, 2006.

[79] D. Rudoy, D. Spendley, and P. J. Wolfe, "Conditionally linear Gaussian models for tracking of vocal tract resonances," in *Proc. 8th Ann. Conf. Intl. Speech Commun. Ass.*, 2007, pp. 526–529.

[80] J. Vargas and S. McLaughlin, "Cascade prediction filters with adaptive zeros to track the time-varying resonances of the vocal tract," *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, pp. 1–6, 2008.

[81] Alberto de Castro, Daniel Ramos, and Joaquin Gonzalez-Rodriguez, "Forensic speaker recognition using traditional features comparing automatic and human-in-the-loop formant tracking," in *Proc. of Interspeech*, 2009, pp. 2343–2346.

[82] C. Gläser, M. Heckmann, F. Joublin, and C. Goerick, "Combining auditory preprocessing and Bayesian estimation for robust formant tracking," *IEEE Trans. on Audio Speech Lang. Process.*, vol. 18, pp. 224–237, 2010.

[83] T. T. Wang and T. F. Quatieri, "High-pitch formant estimation by exploiting temporal change of pitch," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, pp. 171–187, 2010.

[84] H. M. Hanson and E. S. Chuang, "Glottal characteristics of male speakers: acoustic correlates and comparison with female data," *J. Acoust. Soc. Am.*, vol. 106, pp. 1064–1077, 1999.

[85] M. Iseli, Y.-L. Shue, and A. Alwan, "Age, sex, and vowel dependencies of acoustical measures related to the voice source," *J. Acoust. Soc. Am.*, vol. 121, pp. 2283–2295, 2007.

[86] A. V. Oppenheim, R. W. Schafer, and T. G. Stockham, "Nonlinear filtering of multiplied and convolved signals," *Proc. IEEE*, vol. 56, pp. 1264–1291, 1968.

[87] S. J. Julier and J. K. Uhlmann, "New extension of the Kalman filter to nonlinear systems," *Proc. SPIE*, pp. 182–194, 1997.

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

[89] R. H. Shumway and D. S. Stoffer, "An approach to time series smoothing and forecasting using the EM algorithm," *J. Time Ser. Analysis*, vol. 3, pp. 255–264, 1982.

[90] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc. Ser. B*, vol. 39, pp. 1–31, 1977.

[91] L. Ljung, *System Identification*, Upper Saddle River, NJ: Prentice-Hall, 1999.

[92] L. A. Liporace, "Linear estimation of non-stationary signals," *J. Acoust. Soc. Am.*, vol. 58, pp. 1268–1295, 1975.

[93] J. Turner and B. Dickinson, "Linear prediction applied to time-varying all-pole signals," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.*, 1977, pp. 750–753.

[94] M. G. Hall, A. V. Oppenheim, and A. S. Willsky, "Time-varying parametric modeling of speech," *Signal Process.*, vol. 5, pp. 267–285, 1983.

[95] Y. Grenier, "Autoregressive models with time-dependent log area ratios," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 36, pp. 1602–1612, 1988.

[96] K. S. Nathan, Y. T. Lee, and H. F. Silverman, "A time-varying analysis method for rapid transitions in speech," *IEEE Trans. Signal Process.*, vol. 39, pp. 815–824, 1991.

[97] K. S. Nathan and H. F. Silverman, "Time-varying feature selection and classification of unvoiced stop consonants," *IEEE Trans. Speech Audio Process.*, vol. 2, pp. 395–405, 1994.

[98] J. J. Rajan and P. J. W. Rayner, "Generalized feature extraction for time-varying autoregressive models," *IEEE Trans. Signal Process.*, vol. 44, pp. 2498–2507, 1996.

[99] J. J. Rajan, P. J. W. Rayner, and S. J. Godsill, "Bayesian approach to parameter estimation and interpolation of time-varying autoregressive processes using the Gibbs sampler," *IEE Vision, Image, Signal Process.*, vol. 144, pp. 249–256, 1997.

[100] J. Vermaak, C. Andrieu, A. Doucet, and S. J. Godsill, "Particle methods for Bayesian modeling and enhancement of speech signals," *IEEE Trans. Speech Audio Process.*, vol. 10, pp. 173–185, 2002.

[101] J. R. Hopgood and P. J. W. Rayner, "Blind single channel deconvolution using nonstationary signal processing," *IEEE Trans. Speech Audio Process.*, vol. 11, pp. 476–488, 2003.

[102] K. Schnell and A. Lacroix, "Time-varying linear prediction for speech analysis and synthesis," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.*, 2008, pp. 3941–3944.

[103] J. R. Hopgood, "Bayesian blind MIMO deconvolution of nonstationary autoregressive sources mixed through all-pole channels," *Proc. IEEE Workshop Statist. Signal Process.*, vol. 11, pp. 422–425, 2003.

[104] C. Evers and J. R. Hopgood, "Block-based TVAR models for single-channel blind dereverberation of speech from a moving speaker," in *Proc. IEEE 14th Worksh. Stat. Signal Process.*, 2007, pp. 274–278.

[105] C. Evers, J. R. Hopgood, and J. Bell, "Acoustic models for online blind source dereverberation using sequential Monte Carlo methods," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.*, 2008, pp. 4597–4600.

[106] C. Evers and J. R. Hopgood, "Parametric modelling for single-channel blind dereverberation of speech from a moving speaker," *IET Signal Process.*, vol. 2, pp. 59–74, 2008.

[107] K. B. Eom, "Analysis of acoustic signatures from moving vehicles using time-varying autoregressive models," *Multidim. Systems Signal Process.*, vol. 10, pp. 357–378, 1999.

[108] Y. I. Abramovich, N. K. Spencer, and M. D. E. Turley, "Order estimation and discrimination between stationary and time-varying (TVAR) autoregressive models," *IEEE Trans. on Signal Process.*, vol. 55, pp. 2861–2876, 2007.

[109] J. P. Kaipio and P. A. Karjaleinen, "Estimation of event-related synchronization changes by a new TVAR method," *IEEE Trans. Biomed. Eng.*, vol. 44, pp. 649–656, 1997.

[110] M. Juntunen, I. J. Zervo, and J. P. Kaipio, "Root modulus constraints in autoregressive model estimation," *Circuit System Signal Process.*, vol. 17, pp. 709–718, 1998.

[111] M. Juntunen, J. Tervo, and J. P. Kaipio, "Stabilization of Subba Rao-Liporace models," *Circuit System Signal Process.*, vol. 17, pp. 395–406, 1989.

[112] P. Shan and A. A. Beex, "High-resolution instantaneous frequency estimation based on time-varying AR modeling," in *Proc. IEEE Intl. Symp. Time-Frequency Time-Scale Analysis*, 1998, pp. 109–112.

[113] P. Shan and A. A. Beex, "FM interference suppression in spread spectrum communications using time-varying autoregressive model based instantaneous frequency estimation," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.*, 1999, pp. 2559–2562.

[114] P. Shan and A. A. Beex, "Time-varying filtering using full spectral information for soft-cancellation of FM interference in spread spectrum communications," in *Proc. IEEE Worksh. Signal Process. Adv. Wir. Comm.*, 1999, pp. 321–324.

[115] M. K. Tsatsanis and G. B. Giannakis, "Time-varying system identification and model validation using wavelets," *IEEE Trans. Signal Process.*, vol. 41, pp. 3512–3523, 1993.

[116] T. S. Rao, "The fitting of non-stationary time series models with time dependent parameters," *J. Roy. Stat. Soc. Ser. B*, vol. 32, pp. 312–322, 1970.

[117] D. Rudoy, T. F. Quatieri, and P. J. Wolfe, "Time-varying autoregressive tests for multiscale speech analysis," in *Proc. 10th Ann. Conf. Intl. Speech Commun. Ass.*, 2009, pp. 2839–2842.

[118] G. Kitagawa and W. Gersch, "A smoothness priors time-varying AR coefficient modeling of nonstationary covariance time series," *IEEE Trans. Automat. Control*, vol. 30, pp. 48–56, 1985.

[119] C. Andrieu, M. Davy, and A. Doucet, "Efficient particle filtering for jump Markov systems. Application to time-varying autoregressions," *IEEE Trans. Signal Process.*, 2003.

[120] Q. Huang, J. Yang, and S. Wei, "Variational Bayesian learning for speech modeling and enhancement," *Signal Process.*, vol. 88, pp. 2350–2356, 2007.

[121] T. Hsiao, "Identification of time-varying autoregressive systems using maximum a posteriori estimation," *IEEE Trans. Signal Process.*, vol. 56, pp. 3497–3509, 2008.

[122] M. Niedzwiecki, *Identification of Time-Varying Processes*, West Sussex, UK: Wiley, 2002.

[123] Y. I. Abramovich, N. K. Spencer, and M. D. E. Turley, "Time-varying autoregressive (TVAR) models for multiple radar observations," *IEEE Trans. on Signal Process.*, vol. 55, pp. 1298–1311, 2007.

[124] S. J. Godsill and P. J. W. Rayner, *Digital Audio Restoration*, Springer, Norwell, MA, 1998.

[125] H. Akaike, "Block-Toeplitz matrix inversion," *SIAM J. Appl. Math.*, vol. 24, pp. 234–241, 1973.

[126] S. M. Kay, "Noise compensation for autoregressive spectral estimates," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 28, pp. 292–303, 1983.

[127] C. E. Davila, "A subpace approach to estimation of autoregressive parameters from noisy measurements," *IEEE Trans. Signal Process.*, vol. 46, pp. 531–534, 1998.

[128] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithms," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 4, pp. 373–385, 1998.

[129] J. Vermaak, *Bayesian Modeling and Enhancement of Speech Signals*, Ph.D. thesis, University of Cambridge, UK, 2000.

[130] Y. Cho and L. K. Saul, "Learning dictionaries of stable autoregressive models for audio scene analysis," in *Proc. 26th Intl. Conf. Machine Learning*, 2009, pp. 169–176.

[131] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming (web page and software)," June 2009. [Online]. Available: http://stanford.edu/ boyd/cvx.

[132] C. Martinelli, G. Orlandi, L Prina-Ricotti, and S. Racazzi, "Identification of stable nonstationary lattice predictors by linear programming," *Proc. IEEE*, vol. 74, pp. 759–760, 1986.

[133] R. L. Moses and D. Liu, "Determining the closest stable polynomial to an unstable one," *IEEE Trans. Signal Process.*, pp. 901–906, 1991.

[134] M. Jachan, G. Matz, and F. Hlawatsch, "TFARMA models: Order estimation and stabilization," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.*, 2005, pp. 301–304.

[135] S. Dasgupta, M. Fu, and C. Schwarz, "Robust relative stability of time-invariant and time-varying lattice filters," *IEEE Trans. on Signal Process.*, vol. 46, pp. 2088–2100, 1998.

[136] S. Dasgupta, M. Fu, and C. Schwarz, "Exponential asymptotic stability of time-varying inverse prediction error filters," *IEEE Trans. on Signal Process.*, vol. 48, pp. 1928–1936, 2000.

[137] W. Fong, S. Godsill, A. Doucet, and M. West, "Monte Carlo smoothing with applications to audio signal enhancement," *IEEE Trans. on Signal Process.*, vol. 50, pp. 438–449, 2002.

[138] E. Punskaya, C. Andrieu, A. Doucet, and W. J. Fitzgerald, "Bayesian curve fitting using MCMC with applications to signal segmentation," *IEEE Trans. Signal Process.*, vol. 50, pp. 747–758, 2002.

[139] B. G. Quinn and D. F. Nicholls, "The estimation of random coefficient autoregressive models I," *J. Time Ser. Anal.*, vol. 1, pp. 37–46, 1980.

[140] B. G. Quinn and D. F. Nicholls, "The estimation of random coefficient autoregressive models II," *J. Time Ser. Anal.*, vol. 2, 1981.

[141] H. Ohlsson, L. Ljung, and S. Boyd, "Segmentation of ARX-models using sum-of-norms regularization," *Automat.*, vol. 46, pp. 1107–1111, 2010.

[142] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *J. Roy. Stat. Soc. B*, vol. 58, pp. 267–288, 1996.

[143] R. M. Gray, A. Buzo, A. H. Gray, and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 28, pp. 3993–4003, 1980.

[144] M. Basseville and A. Benveniste, "Sequential detection of abrupt changes in spectral characteristics of digitial signals," *IEEE Trans. Inf. Theory*, vol. 29, pp. 709–724, 1983.

[145] F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelihood method," *Proc. 6th Intl. Congr. Acoust. Tokyo, Japan*, pp. C17–C20, 1968.

[146] T. T. Georgiou, "Distances between power spectral densities," *IEEE Trans. Signal Process.*, vol. 55, pp. 3993–4003, 2007.

[147] A. Von Brandt, "Detecting and estimating parameter jumps using ladder algorithms and likelihood ratio tests," *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.*, vol. 8, pp. 1017–1020, 1983.

[148] R. Andre-Obrecht, "A new statistical approach for the automatic segmentation of continuous speech signals," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 36, pp. 29–40, 1988.

[149] E. Moulines and R. Di Francesco, "Detection of the glottal closure by jumps in the statistical properties of the speech signal," *Speech Commun.*, vol. 9, pp. 401–418, 1990.

[150] F. Kozin and F. Nakajima, "The order determination problem for linear time-varying AR models," *IEEE Trans. Automat. Control*, vol. 25, pp. 250–257, 1980.

[151] S. M. Kay, "A new nonstationarity detector," *IEEE Trans. Signal Process.*, vol. 56, pp. 1440–1451, 2008.

[152] S. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory*, Upper Saddle River, NJ: Prentice-Hall, 1998.

[153] D. Rudoy, P. Basu, T. F. Quatieri, B. Dunn, and P. J. Wolfe, "Adaptive short-time analysis-synthesis for speech enhancement," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.*, 2008, pp. 4905–4908.

[154] D. Rudoy, P. Basu, and P. J. Wolfe, "Superposition frames for adaptive time-frequency analysis and fast reconstruction," *IEEE Trans. Signal Process.*, vol. 58, pp. 2581–2596, 2010.

[155] R. C. Hendriks, R. Heusdens, and J. Jensen, "Adaptive time segmentation for improved speech enhancement," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, pp. 2064–2074, 2006.

[156] V. Tyagi, H. Bourlard, and C. Wellekens, "On variable-scale piecewise stationary spectral analysis of signals for ASR," *Speech Commun.*, vol. 48, pp. 1182–1191, 2006.

[157] G. S. Morrison, "Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs," *J. Acoust. Soc. Am.*, vol. 125, pp. 2387–2397, 2009.

[158] D. Y. Wong, J. D. Markel, and A. H. Gray, "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 27, pp. 350–355, 1979.

[159] M. Brookes, P. A. Naylor, and J. Gudnasson, "A quantitative assessment of group delay methods of identifying glottal closures in voiced speech," *IEEE Trans. Audio Speech Lang. Process.*, vol. 8, pp. 1017–1020, 2006.

[160] D. G. Childers and J. N. Larar, "Electroglottography for laryngeal function assessment and speech analysis," *IEEE Trans. on Biomed. Eng.*, vol. 31, pp. 807–817, 1984.

[161] P. A. Naylor, A. Kounoudes, J. Gudnasson, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, pp. 34–43, 2007.

[162] M Morf, B. Dickinson, T. Kailath, and A. Vieira, "Efficient solution of covariance equations for linear prediction," *IEEE Trans. Acoust. Speech Signal Process.*, pp. 429–433, 1977.

[163] J. W. Brewer, "Kronecker products and matrix calculus in system theory," *IEEE Trans. Circuits Syst.*, vol. 25, pp. 772–781, 1978.

[164] S. Dasgupta and M. D. Perlman, "Power of the noncentral F-test: Effect of additional variates on Hotelling's t^2 test," *J. Am. Statist. Ass.*, vol. 69, pp. 174–180, 1974.

[165] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, and M. Davies, "A tutorial on onset detection in music signals," *IEEE Trans. Speech Audio Process.*, vol. 13, pp. 1035–1048, 2005.

[166] S. Jarifia, D. Pastora, and O. Rosec, "A fusion approach for automatic speech segmentation of large corpora with application ot speech synthesis," *Speech Commun.*, vol. 50, pp. 67–80, 2008.

[167] R. E. Quandt, "Tests of the hypothesis that a linear regression system obeys two separate regimes," *J. Amer. Stat. Assoc.*, vol. 55, pp. 324–330, 1960.

[168] T. Drugman and T. Dutoit, "Glottal closure and opening instant detection from speech signals," in *Proc. 10th Ann. Conf. Intl. Speech Commun. Ass.*, 2009.

[169] M. P. Thomas, J. Gudnason, and P. A. Naylor, "Detection of glottal closing and opening instants using an improved DYPSA framework," in *Proc. 17th Eur. Signal Process. Conf. Glasgow, Scotland*, 2009.

[170] N. Henrich, C. d'Alessandro, B. Doval, and M. Castellengo, "On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation," *J. Acoust. Soc. Am.*, vol. 115, pp. 1321–1332, 2004.

[171] M. Huckvale, "Speech filing system: Tools for speech research," 2000.

[172] M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," 2006.

[173] M. B. Priestley and T. Subba-Rao, "A test for non-stationarity of time-series," *J. Roy. Stat. Soc. B*, vol. 31, pp. 140–149, 1969.

[174] T. W. Epps, "Testing that a Gaussian process is stationary," *Ann. Statist.*, vol. 16, pp. 1667–1683, 1988.

[175] M. Grasse, M. R. Frater, and J. F. Arnold, "Testing vbr video traffic for stationarity," *IEEE Trans. Circ. Sys. Video Tech.*, vol. 10(3), pp. 448–459, 2000.

[176] Y. Dwivedi and S. Subba Rao, "A test for second order stationarity based on the discrete Fourier transform," *J. Time Ser. Anal.*, 2010.

[177] P. Borgnat, P. Flandrin, P. Honeine, C. Richard, and J. Xiao, "Testing stationarity with surrogates: a time-frequency approach," *IEEE Trans. Signal Process.*, vol. 57, pp. 2256–2268, 2010.

[178] E. Paparoditis, "Testing temporal constancy of the spectral structure of a time series," *Bernoulli*, vol. 15, pp. 1190–1221, 2009.

[179] D. J. Thomson, "Spectrum estimation and harmonic analysis," *Proc. IEEE*, vol. 70, pp. 1055–1096, 1982.

[180] D. L. Jones and R. G. Baraniuk, "A simple scheme for adapting time-frequency representations," *IEEE Trans. Signal Process.*, vol. 42, pp. 3530–3535, 1994.

[181] D. L. Jones and T. W. Parks, "A high resolution data-adaptive time-frequency representation," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 38, pp. 2127–2135, 1990.

[182] D. L. Jones and R. G. Baraniuk, "An adaptive optimal-kernel time-frequency representation," *IEEE Trans. Signal Process.*, vol. 43, pp. 2361–2371, 1995.

[183] R. N. Czerwinski and D. L. Jones, "Adaptive short-time Fourier analysis," *IEEE Signal Process. Lett.*, vol. 4, pp. 42–45, 1997.

[184] M. M. Goodwin, *Adaptive Signal Models: Theory, Algorithms and Audio Applications*, Kluwer Academic, Norwell, MA, 1998.

[185] S. Mallat, *A Wavelet Tour of Signal Procesing*, Academic, London, second edition, 1999.

[186] H. K. Kwok and D. L. Jones, "Improved instantaneous frequency estimation using an adaptive short-time Fourier transform," *IEEE Trans. Signal Process.*, vol. 48, pp. 2964–2972, 2000.

[187] P. J. Wolfe, S. J. Godsill, and M. Dörfler, "Multi-Gabor dictionaries for audio time-frequency analysis," in *Proc. IEEE Worksh. Appl. Signal Process. Audio Acoust.*, 2001, pp. 43–46.

[188] M. Dörfler, *Gabor Analysis for a Class of Signals called Music*, Ph.D. thesis, University of Vienna, July 2002.

[189] P. J. Wolfe, S. J. Godsill, and W.-J. Ng, "Bayesian variable selection and regularization for time-frequency surface estimation (with discussion)," *J. Roy. Stat. Soc. B*, vol. 66, pp. 575–589, 2004.

[190] F. Jaillet and B. Torresani, "Time-frequency jigsaw puzzle: Adaptive multiwindow and multilayered Gabor expansions," *Intl. J. Wavel. Multires. Inf. Process.*, vol. 5, pp. 293–316, 2007.

[191] I. Djurovic and L. J. Stankovic, "Adaptive windowed Fourier transform," *Signal Process.*, vol. 83, pp. 91–100, 2003.

[192] A. Nesbit, E. Vincent, and M. D. Plumbley, "Benchmarking flexible adaptive time-frequency transforms for underdetermined audio source separation," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.*, 2009.

[193] P. J. Durka and K. J. Blinowska, "A unified time-frequency parametrization of EEGs," *IEEE Eng. Med. Biol. Mag.*, vol. 20, pp. 47–53, 2001.

[194] E. J. Rothwell, K. M. Chen, and D. P. Nyquist, "An adaptive-window-width short-time Fourier transform for visualization of radar target substructure resonances," *IEEE Trans. Antennas Propag.*, vol. 46, pp. 1393–1395, 1998.

[195] H. S. Malvar, "Lapped transforms for efficient transform/subband coding," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 38, pp. 969–978, 1990.

[196] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Trans. Inf. Theory*, vol. 38, pp. 713–718, 1992.

[197] E. Wesfried and M. V. Wickerhauser, "Adapted local trigonometric transforms and speech processing," *IEEE Trans. Signal Process.*, vol. 41, pp. 3596–3600, 1993.

[198] Z. Xiong, K. Ramchandran, C. Herley, and M. Orchard, "Flexible tree-structured signal expansions using time-varying wavelet packets," *IEEE Trans. Signal Process.*, vol. 45, pp. 333–345, 1997.

[199] P. Prandoni and M. Vetterli, "R/D optimal linear prediction," *IEEE Trans. Speech Audio Process.*, vol. 8, pp. 646–655, 2000.

[200] O. A. Niamut and R. Heusdens, "Optimal time segmentation for overlap-add systems with variable amount of window overlap," *IEEE Signal Process. Lett.*, vol. 12, pp. 665–668, 2005.

[201] R. Heusdens and J. Jensen, "Jointly optimal time segmentation, component selection and quantization for sinusoidal coding of audio and speech," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.*, 2005, vol. 3, pp. 193–196.

[202] C. A. Rødbro, J. Jensen, and R. Heusdens, "Rate-distortion optimal time-segmentation and redundancy selection for VoIP," *IEEE Trans. Audio Speech Lang. Process.*, vol. 8, pp. 752–763, 2000.

[203] Z. Cvetkovic and M. Vetterli, "Oversampled filter banks," *IEEE Trans. Signal Process.*, vol. 46, pp. 1245–1255, 1998.

[204] H. Bolcskei, F. Hlawatsch, and H. G. Feichtinger, "Frame-theoretic analysis of oversampled filter banks," *IEEE Trans. Signal Process.*, vol. 46, pp. 3256–3268, 1998.

[205] J. Kovačević and A. Chebira, "Life beyond bases: The advent of frames (Part II)," *IEEE Signal Process. Mag.*, vol. 24, pp. 115–125, 2007.

[206] M. Zibulski and Y. Y. Zeevi, "Discrete multiwindow Gabor-type transforms," *IEEE Trans. Signal Process.*, vol. 45, pp. 1428–1442, 1997.

[207] S. Li, "Discrete multi-Gabor expansions," *IEEE Trans. Inf. Theory*, vol. 45, pp. 1954–1967, 1999.

[208] F. Jaillet, P. Balasz, and M. Dörfler, "Nonstationary Gabor frames," in *Proc. 8th Intl. Conf. Sampling Theory Appl.*, 2009.

[209] S. Qiu and H. G. Feichtinger, "Discrete Gabor structures and optimal representations," *IEEE Trans. Signal Process.*, vol. 43, pp. 2258–2268, 1995.

[210] O. Christensen, *An Introduction to Frames and Riesz Bases*, Birkhäuser, Boston, MA, 2003.

[211] C. Heil and D. Walnut, "Continuous and discrete wavelet transforms," *SIAM Rev.*, vol. 31, pp. 628–666, 1989.

[212] S. H. Nawab and T. F. Quatieri, "Short-time Fourier transform," in *Advanced topics in signal processing*, J. S. Lim and A. Oppenheim, Eds., pp. 289–337. Prentice Hall, 1988.

[213] I. Daubechies, A. Grossmann, and Y. Meyer, "Painless nonorthogonal expansions," *J. Math. Phys.*, vol. 27, pp. 1271–1283, 1986.

[214] C. Herley, J. Kovačević, K. Ramchandran, and M. Vetterli, "Tilings of the time-frequency plane: Construction of arbitrary orthogonal bases and fast tiling algorithms," *IEEE Trans. Signal Process.*, vol. 41, pp. 3341–3359, 1993.

[215] G. Wang, "The most general time-varying filter bank and time-varying lapped transforms," *IEEE Trans. Signal Process.*, vol. 45, pp. 3775–3789, 2006.

[216] R. S. Orr, "The order of computation for finite discrete Gabor transforms," *IEEE Trans. Signal Process.*, vol. 41, pp. 122–130, 1993.

[217] R. A. Wiggins, "Minimum entropy deconvolution," *Geoexplorat.*, vol. 16, pp. 21–35, 1978.

[218] P. Basu, D. Rudoy, and P. J. Wolfe, "A nonparametric test for stationarity based on local Fourier analysis," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.*, 2009, pp. 3005–3008.

[219] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Commun.*, vol. 9, pp. 351–356, 1990.

[220] A. El-Jaroudi and J. Makhoul, "Discrete all-pole modeling," *IEEE Trans. Signal Process.*, vol. 39, pp. 411–423, 1991.

[221] A. Yasmin, P. Fieguth, and Li Deng, "Speech enhancement using voice source models," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.*, 1999, pp. 797–800.

[222] D. Sengupta and S. M. Kay, "Parameter estimation and GLRT detection in colored non-Gaussian autoregressive processes," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 38, pp. 1661–1675, 1990.

[223] R. T. Behrens and L. L. Scharf, "Signal processing applications of oblique projection operators," *IEEE Trans. Signal Process.*, vol. 42, pp. 1413–1424, 1994.

[224] T. W. Anderson and H. Rubin, "Estimation of the parameters of a single equation in a complete system of stochastic equations," *Ann. Math. Stat.*, vol. 20, pp. 46–63, 1949.

[225] T. W. Anderson and H. Rubin, "The asymptotic properties of estimates of the parameters of a single equation in a complete system of stochastic equations," *Ann. Math. Stat.*, vol. 21, pp. 570–582, 1950.

[226] J. Durbin, "Estimation of parameters in time-series regression models," *J. Roy. Stat. Soc. Ser. B*, vol. 22, pp. 139–153, 1960.

[227] D. Vincent, O. Rosec, and T. Chonavel, "A new method for speech synthesis and transformation based on an ARX-LF source-filter decomposition and HNM modeling," in *Proc. Intl. Conf. Speech Audio Acoust.*, 2007, pp. 525–528.

[228] P. Speckman, "Kernel smoothing in partial linear models," *J. Roy. Stat. Soc. Ser. B*, vol. 50, pp. 413–436, 1988.

[229] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*, Chapman and Hall, 2003.

[230] R. K. Mehra, "Optimal input signals for parameter estimation in dynamic systems—survey and new results," *IEEE Trans. Automat. Control*, pp. 753–769, 1974.

[231] S. Mallat, *A Wavelet Tour of Signal Procesing: The Sparse Way*, Academic, London, third edition, 2009.

[232] J. Gudnason, M. R. P. Thomas, P. A. Naylor, and D. P. W. Ellis, "Voice source waveform analysis and synthesis using principal component analysis and Gaussian mixture modelling," in *Proc. INTERSPEECH*, 2009, pp. 108–112.

[233] J. M. Fadili and E. Bullmore, "Penalized partially linear models using sparse representations with an application to fMRI time series," *IEEE Trans. Signal Process.*, vol. 53, pp. 3436–3449, 2005.

[234] P. Green, C. Jennison, and A. Seheult, "Analysis of field experiments by least squares smoothing," *J. Roy. Stat. Soc. Ser. B*, vol. 47, pp. 299–315, 1985.

[235] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, 1994.

[236] P. J. Murphy, K. G. McGuigan, M. Walsh, and M. Colreavy, "Investigation of a glottal related harmonics-to-noise ratio and spectral tilt as indicators of glottal noise in synthesized and human voice signals," *J. Acoust. Soc. Am.*, vol. 123, pp. 1642–1652, 2008.

[237] P. C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, 2007.

[238] N. Li and P. C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: implications for noise reduction," *J. Acoust. Soc. Am.*, vol. 123, pp. 1673–1682, 2008.

[239] G. Kim, Y. Lu, Y. Hu, and P. C. P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *J. Acoust. Soc. Am.*, vol. 126, pp. 1486–1494, 2009.

[240] J. Hillenbrand and R. A. Houde, "Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech," *J. Speech Hearing Research*, vol. 39, pp. 311–321, 1996.

[241] J. Kreiman, B. R. Gerratt, G. B. Kempster, A. Erman, and G. S. Berke, "Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research," *J. Speech Hearing Research*, vol. 36, pp. 21–40, 1993.

[242] B. R. Gerratt and J. Kreiman, "Measuring vocal quality with speech synthesis," *J. Acoust. Soc. Am.*, vol. 110, pp. 2560–2566, 2001.

[243] R. Shrivastav and C. M. Sapienza, "Objective measures of breathy voice quality obtained using an auditory model," *J. Acoust. Soc. Am.*, vol. 114, pp. 2217–2224, 2003.

[244] P. Alku, C. Magi, S. Yrttiaho, T. Backstrom, and B. Story, "Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering," *J. Acoust. Soc. Am.*, vol. 125, pp. 3289–3305, 2009.