# Improved Statistical Signal Processing of Nonstationary Random Processes Using Time-Warping

Scott Thomas Wisdom

A thesis submitted in partial fulfillment of the
requirements for the degree of

Master of Science in Electrical Engineering

University of Washington

2014

Reading Committee:

Les Atlas, Chair

James Pitton

Henrique Malvar

Program Authorized to Offer Degree:
Electrical Engineering

University of Washington

**Abstract**

Improved Statistical Signal Processing of Nonstationary Random Processes Using Time-Warping

Scott Thomas Wisdom

Chair of the Supervisory Committee:
Professor Les Atlas
Electrical Engineering

A common assumption used in statistical signal processing of nonstationary random signals is that the signals are locally stationary. Using this assumption, data is segmented into short analysis frames, and processing is performed using these short frames. Short frames limit the amount of data available, which in turn limits the performance of statistical estimators.

In this thesis, we propose a novel method that promises improved performance for a variety of statistical signal processing algorithms. This method proposes to estimate certain time-varying parameters of nonstationary signals and then use this estimated information to perform a time-warping of the data that compensates for the time-varying parameters. Since the time-warped data is more stationary, longer analysis frames may be used, which improves the performance of statistical estimators.

We first examine the spectral statistics of two particular types of nonstationary random processes that are useful for modeling ship propeller noise and voiced speech. We examine the effect of time-varying frequency content on these spectral statistics, and in addition show that the cross-frequency spectral statistics of these signals contain significant additional information that is not usually exploited using a stationary assumption. This information, combined with our proposed method, promises improvements for a wide variety of applications in the future. We then describe and test an implementation of our time-warping method, the fan-chirp transform. We apply our method to two applications,

detection of ship noise in a passive sonar application and joint denoising and dereverberation of speech. Our method yields improved results for both applications compared to conventional methods.

## TABLE OF CONTENTS

## LIST OF FIGURES

## ACKNOWLEDGMENTS

I would like to thank my advisors, Les Atlas and Jim Pitton, for their time, creativity, support, and encouragement through my Master's studies. You have both helped me see problems in new ways and I have benefited immensely from your expertise. Thanks also to Rico Malvar of Microsoft Research; your advice and support have been invaluable, and I have greatly benefited from your deep knowledge and extensive experience. I would also like to thank Josh Smith, Jeff Bilmes, Don Percival, Maryam Fazel, Marina Meila, and Mehran Mesbahi for all the effort they have put in to their excellent courses at the University of Washington, which have been crucial to my own development. Thanks to Shannon Hughes, Francois Meyer, Tom Mullis, Peter Mathys, and Robert Erickson from the University of Colorado Boulder for their early guidance during my undergraduate career. I have also had the privilege to know and work with many talented collaborators, lab mates, and fellow grad students, including Greg Okopal, Tommy Powers, Alanson Sample, Ben Waters, Elliot Saba, Bill Kooiman, Kai Wei, Brian King, Xing Li, Pascal Clark, Jessica Tran, Nicole Nichols, Xingbo Peng, Renshu Gu, Brian Hutchinson, John Halloran, Rishabh Iyer, De Meng, Karthik Mohan, Reza Eghbali, Amin Jalali, Andrew Haddock, Mike Imhof, David Perlmutter, Aaron Parks, Yi Zhao, and Vamsi Talla. I am also very grateful to the Office of Naval Research for financial support throughout my studies. Thanks to my parents and grandparents, my sister Rosie, and the Rogers family for always believing in me and providing encouragement and support through the ups and the downs. And most of all, thanks to my amazing girlfriend Lindsay, who makes my life immeasurably happy and keeps me going every day.

## DEDICATION

To my mom, dad, and sister Rosie.

# Chapter 1
# INTRODUCTION

## *1.1 Overview*

This thesis describes a novel paradigm for processing nonstationary random signals that promises improved performance for a wide variety of statistical estimators. Historically, many nonstationary signals of interest, such as speech, neural signals, and ship noise, have been processed frame-by-frame, where certain parameters of the signal are assumed to be stationary within the frame duration. For example, speech processing often employs analysis frames of duration 20 milliseconds or less, because speech is assumed to be approximately stationary over this duration [1, §7.2.3]. The requirement of short-term stationarity limits the amount of data within an analysis frame, which in turn limits the performance of statistical estimators used for tasks such as enhancement, detection, or classification.

We propose to extend the duration of these short frames while still keeping the data approximately stationary by altering the time base of the analysis. This alteration is easily implemented as a time-warping. For certain signals of interest, such a time-warping can yield longer analysis frames that are still approximately stationary. Longer frame durations mean more data is available to statistical estimators, which improves the performance of these estimators by decreasing their variance and allowing them to be more effective in the presence of noise. If resynthesis is desired—for example, after applying modifications to remove noise—from processed short frames, it is a simple matter to undo the time-warping and perform reconstruction of the signal.

We also show that our nonstationary signals of interest, which are simple models of ship propeller noise and voiced speech, are cyclostationary when they have stationary frequency content, and close to cyclostationary when the frequency content varies linearly over the analysis window. Using a time-warping, it is possible to compensate these "near-cyclostationary" signals to be cyclostationary. The cyclostationary property of these non-

stationary signals is interesting, because many useful algorithms have been developed for cyclostationary signals that exploit their particular properties. Perhaps the most important property about cyclostationary signals is that they are correlated with frequency-shifted copies of themselves, which is a consequence of the cross-frequency correlation in the spectral representation of such signals. Applications that exploit the information from cross-frequency correlations in cyclostationary signals include blind source separation [2] and beamforming [3]–[5], as well as improved minimum mean-square error filtering [6], and detection [7]. Such algorithms have been primarily developed for man-made communications signals in the past, and the idea of compensating real-world signals such as ship propeller noise or voiced speech to make them cyclostationary seems to have little coverage in the literature.

To complete this introduction chapter, we will provide a survey of the related literature and describe the contents of this thesis.

## *1.2 Survey of the Literature*

### *1.2.1 Testing stationarity*

Since one of our goals is to find transformations that stationarize data (that is, make their statistics less dependent on time), we here examine prior work on testing for stationarity. There are several tests of stationarity in the literature. One of the first tests for stationarity is due to Priestley and Subba Rao [8]. Their method takes a time series, divides it into frames, computes a spectral estimate for each frame, and the test statistic looks for differences between the spectral estimates of the frames. Borgnat and Flandrin along with their coauthors proposed using the method of surrogates to test stationarity, where surrogates are generated by reconstructing from one-dimensional spectra or two-dimensional time-frequency displays that have had their phase randomized [9], [10]. Borgnat and Flandrin also showed how the surrogate method can be used to detect specific nonstationary trends in data, which for their example was dynamic light scattering of a living cell nucleus [9]. Wavelet-based tests of stationarity have been proposed by von Sachs and Neumann [11] and Nason [12].

### *1.2.2 Stationarization of random processes and time-warping*

Both stationarization and analysis of nonstationary random processes by means of time-warping, or time-deformation, have been proposed by several authors. In 1988, Gray and Zhang introduced the class of $M$-stationary processes, which are continuous random processes whose periodic structure vary linearly over time (which means frequency changes over time are proportional to $1/t$) [13]. Such temporal variation means $M$-stationary processes are stationary on a log-spaced time scale [13]. In 2005, Gray et al. provided a discrete-time formulation of $M$-stationary processes, and applied the model to bat echolocation data [14]. Building on this work, Jiang et al. introduced the class of $G$-stationary processes, which is composed of stationary processes $y(u)$ warped by some monotonic time-warping function $u = g(t)$: $x(t) = y(g^{-1}(u))$ [15]. In this same paper, Jiang et al. also suggested a method of fitting a $G$-stationary process to data using a comparison between sample autocorrelation functions of the first and second halves of the data and doing an exhaustive search over a range of parameters. Most recently, Xu et al. proposed using the $G$-stationary process model to perform time-varying filtering of nonstationary random processes, and demonstrated this method on a few simple examples [16]. For the application of activity recognition, Veeraraghavan et al. used an optimization over warping functions to normalize measured sequences of activities [17]. By using this optimization, they were able to improve their recognition performance and improve computation efficiency by clustering activity data.

In 2013, Andén and Mallat proposed the deep scattering spectrum, which they show is stable to time-warpings of the input signal in the sense of a Lipschitz continuity condition using the $\ell_2$-norm [18].

### *1.2.3 Other notions of stationarization*

In 1978, Gardner described how cyclostationary process can be stationarized by random translations. That is, if $x(t)$ is cyclostationary with cycle period $T$, then the random translated version $\widetilde{x}(t) = x(t + \theta)$, where $\theta$ is a uniformly distributed random variable on $[-T/2, T/2]$, then $\widetilde{x}(t)$ is wide-sense stationary [19]. In the same paper, Gardner went on

to characterize wider classes of nonstationary processes that are stationarizable, including almost-cyclostationary processes. The motivation for this stationarization is to allow processing of cyclostationary signals using time-invariant filters, which incurs a cost of usually modest higher mean-squared error. Our paradigm presented in this thesis, however, promises better performance of time-invariant filters applied to certain types of nonstationary signals, because we intend to keep track of the nature of the time-deformation so that it can be undone after processing.

Another notion of stationarization of cyclostationary processes is the well-known fact that any discrete-time cyclostationary process whose cycle period is $M$ samples long can be transformed to an $M-$dimensional vector-valued wide-sense stationary random process. This property is originally due to Gladyshev in 1961 [20]. A continuous version of this property was proposed by Gladyshev in 1963 [21].

### *1.3 Contents of the Thesis*

This thesis is made up of two parts. The first part, which consists of chapters 2 and 3, is more theoretical. Chapter 2 describes the structure of the spectral second-order statistics of two subclasses of nonstationary random processes and the effect that a time-varying fundamental frequency has on these statistics. This analysis shows that cross-frequency spectral correlations contain a substantial amount of information that is not usually exploited using a stationary assumption about the data, and that our proposed paradigm is necessary to prevent smearing of this additional information. Chapter 3 describes methods of estimating and applying time-warping functions to signals with time-varying fundamental frequency such that their fundamental frequency becomes constant. After the time-warping is performed, these signals are more (cyclo-)stationary.

The second part of this thesis, consisting of chapters 4 and 5, presents applications that make use of the theoretical ideas and issues raised in part one. Chapter 4 describes an improved detection method for modulated wide-sense stationary random processes that exploits time-varying modulation frequency structure over the analysis frame to increase coherence time. That is, by exploiting time-varying modulation frequency content in the signal of interest, the method allows for longer analysis windows, which provides a better-

performing detector. Detection of ship propeller noise in passive sonar is used as an example. Chapter 5 presents a novel algorithm for joint noise reduction and dereverberation for single-channel speech signals. This chapter takes a similar approach to chapter 4, in that the time-varying spectral structure of voiced speech is estimated and compensated to allow for longer analysis frames.

Chapter 6 concludes the thesis, summarizes our contributions, and suggests promising avenues for future work.

Chapter 2

# SPECTRAL SECOND-ORDER STATISTICS OF NONSTATIONARY RANDOM PROCESSES

## *2.1 Introduction*

It is well-known that all wide-sense stationary (WSS) random processes possess complex-valued spectral increments that are uncorrelated between different frequencies [22], [23]. Loosely speaking, the spectral increments of a random process are the spectral representation of the process. For WSS processes, the well-known one-dimensional power spectral density (PSD) characterizes the second-order statistical behavior of the process's spectral representation [22]. These complex-valued spectral increments also have the property that they are uncorrelated with their conjugates [23, pp 199]. This means their real parts are uncorrelated with their imaginary parts and that the variances of their real and imaginary parts are equal.

Nonstationary random processes, on the other hand, possess correlated spectral increments [23, Result 9.2], which necessitates a two-dimensional representation that conveys both auto- and cross-correlations between spectral increments [23], [24]. Also unlike WSS processes, the complex-valued spectral increments of nonstationary random processes can exhibit correlation with their conjugates, which necessitates an additional second-order, complementary statistic.

In this chapter we will examine the exact structure of the two-dimensional spectral correlation for two specific types of nonstationary random processes. These two types of random processes are chosen here because they are useful for modeling real-world signals of interest. These processes are amplitude-modulated wide-sense stationary (AM-WSS) processes, which are a good model for ship propeller noise in passive sonar, and filtered jittered pulse train (F-JPT) processes, which we present as a simplified model for voiced speech.

Both AM-WSS and F-JPT processes have a notion of "fundamental frequency"; for AM-WSS, this fundamental frequency is the lowest frequency of a deterministic and possibly harmonic modulator $m(t)$ that multiplicatively modulates a WSS random process. The deterministic modulator $m(t)$ is the sole source of nonstationarity in the AM-WSS signal. When an AM-WSS process is used to model ship noise, this fundamental frequency corresponds to the speed of a ship's propeller. For F-JPT processes, the fundamental frequency describes the frequency of the driving pulse train, which is akin to the frequency of glottal pulses in voiced speech. In this chapter, we will examine the two-dimensional spectral second-order statistics for two cases: 1) the fundamental frequency of these processes is constant over the finite duration of the process, and 2) the fundamental frequency varies linearly over a finite duration of the process. In the first case, we will see that both processes are cyclostationary. In the second case, the processes are no longer cyclostationary, and we will examine the effect of time-varying fundamental frequency on the second-order two-dimensional spectral statistics.

Though we will not explicitly estimate these spectral statistics in this chapter, nor in this thesis, examining these statistics gives insight about the types of signals we are interested in processing. We provide this preliminary work in the hope that the exact structure of the spectral correlation of these processes can be exploited in the future to improve performance of various statistical signal processing algorithms. We hope the reader will be convinced of the following three points after reading this chapter:

1. Our nonstationary signals of interest exhibit substantial cross-frequency spectral correlation, which is information not usually exploited when an assumption of stationarity is used.

2. The spectral increments of our nonstationary signals of interest exhibit substantial correlation with their conjugates, which is additional information that is not usually exploited by standard processing

3. Variation of the fundamental frequency of our nonstationary signals of interest over a finite analysis interval smears out the potential additional information, which ne-

cessitates some way of correcting or compensating for this time-varying fundamental frequency. In later chapters we will see this method of correction is a time-warping of the data.

Regarding point 1, if we know the structure of correlation between the spectral increments at different frequencies, then frequency-shifted copies of a nonstationary signal can be used to achieve better performance for statistical estimators. Regarding point 2, if the spectral increments are correlated with their conjugates, then further performance gains may be had by employing widely-linear processing, which is processing that uses both complex-valued data and the conjugate of the complex-valued data as inputs [23], [25]–[27].

This chapter proceeds as follows. Section 2.2 provides required background on the statistics of complex-valued random data, the spectral representation of random processes, and cyclostationary random processes. Section 2.3 describes the second-order spectral statistics of AM-WSS processes . Section 2.4 describes the second-order spectral statistics of F-JPT processes. Section 2.5 summarizes our contributions and suggests promising directions for future work.

## *2.2 Background*

### *2.2.1 Complex-valued random data*

Since we will be examining statistics of the complex-valued spectral increments of random processes, we need to give some background on the statistics of complex-valued random data. There has been a recent resurgence of interest in processing complex-valued data, starting with Picinbono [28]–[30] and leading to books by Schreier and Scharf [23] and Mandic and Goh [27]. The main idea motivating this research is that complex-valued random data requires not one, but two second-order moments to completely specify its second-order statistical behavior. For many developments and applications in the past, the complementary second-order moment has been ignored. Ignoring the complementary second-order moment is valid if and only if the data is *second-order circular*, or *proper*, which is often not true of complex-valued random data. The meaning behind these terms will be explained shortly, and we will see that improper random data exhibits correlation

between itself and its conjugate. This also means there may be correlation between the data's real and imaginary components and/or imbalance between the variance of its real and imaginary components.

A scalar, complex-valued random variable $x$ has two second-order moments, the Hermitian variance $R_{xx}$ and the complementary variance $\widetilde{R}_{xx}$, which are defined as

$$\begin{aligned} R_{xx} &= E\left\{(x-\mu_x)(x-\mu_x)^*\right\} \\ \widetilde{R}_{xx} &= E\left\{(x-\mu_x)(x-\mu_x)\right\}. \end{aligned} \tag{2.1}$$

Note that $R_{xx}$ is always nonnegative and real-valued, while $\widetilde{R}_{xx}$ is generally complex-valued.

The random variable $x$ is said to be *proper*[1] if its complementary variance vanishes: $\widetilde{R}_{xx} = 0$. Otherwise, $x$ is called *improper* [23, Definition 2.1]. A more restrictive property of a complex random variable is *circular*, which implies that the probability distribution of $x$ in the complex plane is rotationally invariant, which in general corresponds to all $N$th-order moments being nonzero only if the number of conjugated and nonconjugated terms are equal [23, Result 2.13]. The distinction between impropriety and noncircularity is analogous to the distinction between wide-sense stationarity and strict-sense stationarity.

The *impropriety coefficient* of $x$ (sometimes referred to as "circularity coefficient" in the literature) is defined as

$$\rho_x = \frac{\widetilde{R}_{xx}}{R_{xx}} \tag{2.2}$$

and is a complex-valued quantity that carries information about the shape of the random variable's constellation in the complex plane. The magnitude of $\rho_x$ is known as the *degree of noncircularity*, and if the complex-valued random variable $x$ is Gaussian distributed, $|\rho_x|$ describes the eccentricity of the ellipse in the complex plane. The angle of $\rho_x$ is equal to $2\phi$, where $\phi$ is known as the *polarization angle*. Realizations of proper (i.e., $|\rho_x| = 0$) and improper (for $|\rho_x| = 0.8$, $\angle\rho_x = 2\phi$, and $\phi = 3\pi/8$) complex-valued Gaussian random variables are shown in figure 2.1.

---

[1]The term "proper" may seem a bit strange. Its origin, according to [23], lies in a 1993 paper by Neeser and Massey [31], where the authors were considering applications of proper random vectors in the context of communications and information theory. In these applications, circular distributions were considered the correct, or "proper," signals. The term has since stuck to refer to second-order circularity.

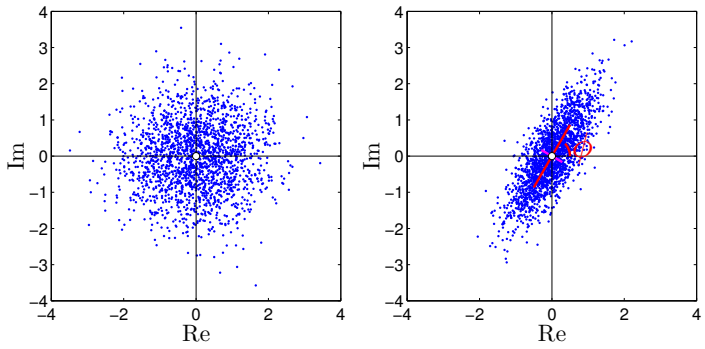


Figure 2.1: Proper complex-valued Gaussian (left) and improper complex-valued Gaussian (right) with degree of noncircularity $|\rho_x| = 0.8$, variance of major axis in red, variance of minor axis in magenta, and polarization angle $\phi = 3\pi/8$ indicated.

Notice that for a zero-mean, complex-valued random variable $x = a + jb$, where $a$ is the real part and $b$ is the imaginary part, the complementary variance factors as follows:

$$\widetilde{R}_{xx} = E\left\{(a+jb)(a+jb)\right\} = E\{a^2\} + 2jE\{ab\} - E\{b^2\}. \tag{2.3}$$

This expression highlights the fact that the complementary variance will be nonzero if the real and imaginary parts are correlated (i.e., the $2jE\{ab\}$ term in (2.3) is non-zero) or if there is an imbalance between the variances of the real and imaginary parts (i.e., the $E\{a^2\} - E\{b^2\}$ terms in (2.3) have a non-zero result).

### *2.2.2 Spectral second-order statistics of nonstationary random processes*

The Cramér-Loève spectral representation [24] of a harmonizable random process $x(t)$ is

$$x(t) = \int e^{j2\pi ft} d\xi(f), \tag{2.4}$$

where $\xi(f)$ is a complex-valued spectral process with increments $d\xi(f)$.

In this chapter, to simplify our notation, we will assume all signals of interest are finite-energy[2], which means we can assume $d\xi(f) = X(f)df$, where $X(f)$ is a complex-valued

[2]See appendix A in Clark's thesis [32] or §1.1.2 in Napolitano's book [33] for a detailed treatment of this assumption.

random process. We are interested in the Hermitian spectral correlation

$$S_{XX}(f_1, f_2) = E\left[X(f_1)X^*(f_2)\right] \tag{2.5}$$

and complementary spectral correlation[3]

$$\widetilde{S}_{XX}(f_1, f_2) = E\left[X(f_1)X(f_2)\right]. \tag{2.6}$$

We also want to know the Hermitian spectral covariance

$$R_{XX}(f_1, f_2) = E\left[X(f_1)X^*(f_2)\right] - E\left[X(f_1)\right]E\left[X^*(f_2)\right] \tag{2.7}$$

and the complementary spectral covariance

$$\widetilde{R}_{XX}(f_1, f_2) = E\left[X(f_1)X(f_2)\right] - E\left[X(f_1)\right]E\left[X(f_2)\right]. \tag{2.8}$$

To gain intuition about these two-dimensional spectral correlations, let us first examine the two-dimensional spectral correlations of a real-valued WSS random process $x(t)$. For such a process, the spectral representation $X(f)$ is zero-mean at all frequencies, and is uncorrelated at different frequencies. Since $x(t)$ is real-valued, its spectral representation is conjugate-symmetric: $X(f) = X^*(-f)$. Thus, the spectral representation $X(f)$ has two-dimensional second order moments

$$S_{XX}(f_1, f_2) = E[X(f_1)X^*(f_2)] = \begin{cases} S_x(f) & \text{if } f_1 = f_2 = f \\ 0 & \text{otherwise} \end{cases} \tag{2.9}$$

$$\widetilde{S}_{XX}(f_1, f_2) = E[X(f_1)X(f_2)] = E[X(f_1)X^*(-f_2)] = \begin{cases} S_x(f) & \text{if } f_1 = -f_2 = f \\ 0 & \text{otherwise} \end{cases} \tag{2.10}$$

---

[3]Note that this definition of spectral correlation differs from that proposed by Schreier and Scharf [23], which is $\widetilde{S}_{xx}(f_1, f_2)df_1df_2 = E\left[d\xi(f_1)d\xi(-f_2)\right]$ [23, (9.49)]. The additional minus sign on $f_2$ comes from Schreier and Scharf's desire for the Cramér-Loève representation for nonstationary random processes to simplify to that of WSS random processes, which they define as $\widetilde{P}_{xx}(f)df = E\left[d\xi(f)d\xi(-f)\right]$ [23, Result 8.1]. The reason for the extra minus sign in the WSS case is because Schreier and Scharf consider *complex*-valued harmonizable random processes, where $x^*(t) = \left(\int_{-\infty}^{\infty} e^{j2\pi ft}d\xi(f)\right)^* = \int_{-\infty}^{\infty} e^{j2\pi ft}d\xi^*(-f)$ [23, (8.12)]. Here, our definition differs, because we are only concerned with real-valued processes $x(t)$, and we are treating the complex-valued spectral representation $X(f)$ as a complex-valued random process. Furthermore, since we desire to eventually examine the impropriety of the complex-valued random process $X(f)$, we thus require its Hermitian and complementary covariances.

Here, $S_x(f)$ is the familiar one-dimensional power spectral density (PSD), which is the Fourier transform in $\tau$ of the autocovariance function of $x(t)$:

$$S_x(f) = \int r_{xx}(\tau) e^{-j2\pi\tau f} d\tau. \tag{2.11}$$

Note that the complementary moment $\widetilde{S}_{XX}(f_1, f_2)$ in (2.10) is zero everywhere except at $f = 0$ where it equals $S_x(0)$, since this is the only value of $f$ that satisfies $f_1 = -f_2 = f$. In this case, we can totally disregard the complementary spectral correlation $\widetilde{S}_{XX}(f_1, f_2)$, because it carries no additional information about $x(t)$. This confirms our statement in the introduction, that the spectral increments of a WSS process are uncorrelated with their conjugates, and are thus always proper. This justifies the traditional assumption that the PSD fully characterizes the spectral behavior of a WSS random process.

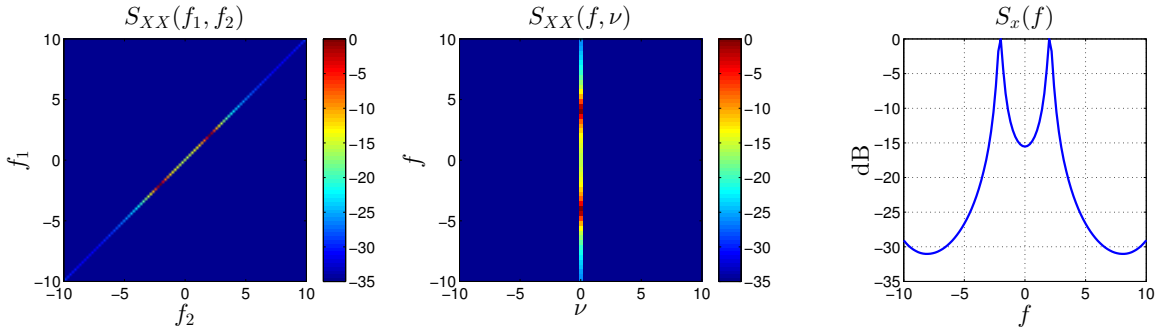


Figure 2.2: Left: bifrequency spectral correlation of a wide-sense stationary (WSS) random process $x(t)$. Center: global/local spectral correlation of $x(t)$. Right: conventional one-dimensional power spectral density (PSD) of $x(t)$. Notice that along the diagonal $f_1 = f_2$ in the bifrequency spectral correlation (left panel) and along the vertical line $\nu = 0$ in the global/local spectral correlation (center panel) are both equal to the one-dimensional PSD shown in the right panel. Also notice that the bifrequency and global/local spectral correlations are zero everywhere else.

Here a picture is useful to gain intuition. The left panel in figure 2.2 shows the bifrequency two-dimensional spectral correlation $S_{XX}(f_1, f_2)$ of a zero-mean, real-valued, WSS, autoregressive process $x(t)$. A portion of the one-dimensional PSD $S_x(f)$ is shown in the

right panel of figure 2.2.

The center panel of figure 2.2 shows an alternate way to view the two-dimensional spectral correlations. While the definitions above use a bifrequency formulation (i.e., $(f_1, f_2)$, where the units of $f_1$ and $f_2$ are equal), an alternate formulation is a "global/local" frequency formulation, where the two variables are $f$, a global frequency, and $\nu$, a local frequency [23, pp 235]. The global frequency $f$ and the local frequency $\nu$ are equivalent to the notions of acoustic frequency and modulation frequency, respectively [34]. The global/local representation is a simple one-to-one variable transformation between $(f_1, f_2)$ and $(f, \nu)$, where $f = f_1$ and $\nu = f_1 - f_2$. The global/local spectral correlation $S_{XX}(f, \nu)$ has an additional nice property: it is the Fourier transform in $t$ of the second-order time-frequency representation $V_{xx}(t, f)$ [23], which is a time-frequency distribution such as a Wigner-Ville or Rihaczek distribution [23, pp 238].

Note that along the diagonal $f_1 = f_2$, the bifrequency spectral correlation in the left panel of figure 2.2 is equal to the PSD $S_x(f)$, shown in the center panel of figure 2.2. Similarly, along the vertical line $\nu = 0$ in the right panel of figure 2.2, the global/local spectral correlation is also equal to the PSD $S_x(f)$. The line that the conventional one-dimensional PSD for WSS processes lies along is sometimes referred to as the *stationary manifold* [23, pp 233]. For the bifrequency formulation, the stationary manifold is the diagonal line $f_1 = f_2$. In the global/local formulation, the stationary manifold is the vertical line $\nu = 0$.

Throughout the rest of this chapter, we will make use of both the bifrequency and global/local formulations of two-dimensional spectral correlation, as the most intuitive formulation depends on the signal of interest.

*Connection to spectral impropriety*

Recall the impropriety coefficient of a complex-valued random variable from (2.2). Clark [32] and Clark et al. [35] examined the spectral impropriety of real-valued signals. Here, our formulation of bifrequency spectral covariances leads to an elegant geometric formulation

of spectral impropriety. Define the spectral impropriety of a random process as

$$\rho_X(f) = \frac{\widetilde{R}_{XX}(f,f)}{R_{XX}(f,f)}. \tag{2.12}$$

From this equation, it is easy to see that the spectral impropriety at any frequency can be found by simply taking the bifrequency complementary spectral covariance (2.8) along the diagonal $f_1 = f_2 = f$, $\widetilde{R}_{XX}(f,f)$, and dividing it by the bifrequency Hermitian spectral covariance (2.7) along the diagaonal $f_1 = f_2 = f$, $R_{XX}(f,f)$.

### *2.2.3 Cyclostationary random processes*

We briefly review cyclostationary signals here, as well as their spectral correlation. We will see later in this chapter that when the fundamental frequency of our signals of interest is constant, such signals are cyclostationary.

A zero-mean, real-valued random process $x(t)$ is said to be *cyclostationary*, or periodically-correlated [21], with cycle period $T_0$ if and only if its time-varying autocovariance function is periodic with fundamental period $T_0$ [36]. That is,

$$r_{xx}(t,\tau) = E\left[x(t)x(t+\tau)\right] = r_{xx}(t+T_0,\tau) \tag{2.13}$$

for all $t$ and $\tau$.

Cyclostationary signals are nonstationary because their autocovariance function varies periodically over time. However, since the nature of the nonstationarity is simple, it is relatively easy to characterize the two-dimensional spectral correlation of cyclostationary random processes. In fact, the exact structure of the correlation between spectral increments has been explored for only a few small subclasses of the class of nonstationary random process, which includes cyclostationary signals and some generalizations of cyclostationary signals.

#### *Spectral correlation of cyclostationary random processes*

The two-dimensional bifrequency Hermitian spectral correlation $R_{XX}(f_1,f_2)$ of cyclostationary processes has support on parallel lines with unit slope with intercepts at integer multiples of $1/T_0$ in the bifrequency plane, [36], [37]. The two-dimensional bifrequency

complementary spectral correlation $\widetilde{R}_{XX}(f_1, f_2)$ of cyclostationary processes has support on parallel lines with slope of $-1$ with intercepts at integer multiples of $1/T_0$ in the bifrequency plane. For visualizations of these spectral correlations, see the center panels of figure 2.3 and the right panels of figure 2.4.

*Generalizations of cyclostationary random processes*

Several generalizations of cyclostationary processes have been described in the literature. Though we will not use results for these generalizations in this thesis, we mention them here for completeness. A sum of at least two uncorrelated cyclostationary random processes with cycle periods that are coprime is said to be *almost-cyclostationary* (ACS) [19]. This means that $r_{xx}(t, \tau)$ is an almost-periodic function of $t$ [33]. Other generalizations include generalized almost-cyclostationary (GACS) processes [33, ch 2] and spectrally-correlated (SC) processes [33, ch 4].

ACS processes have spectral correlations have support on parallel lines with unit slope and intercepts at integer multiples of $1/T_i$ in the bifrequency plane, where $T_i$ is one of the $N$ coprime cycle periods for $i = 0, 1, ...N - 1$ [19]. The spectral correlations of GACS have an impulsive component corresponding to the underlying ACS process and a continuous component corresponding to the purely GACS contribution [33, Theorem 2.2.22]. Spectrally-correlated (SC) processes have spectral correlations with support on a countable set of curves in the bifrequency plane [33].

In the next two sections, we will describe the Hermitian and complementary, two-dimensional second-order spectral statistics of two subclasses of nonstationary random processes, AM-WSS and F-JPT. We will see that when the fundamental frequency of either of these processes is constant, these processes are cyclostationary.

## *2.3 Amplitude-Modulated Wide-Sense Stationary (AM-WSS) Processes*

This section examines the subclass of nonstationary signals that are composed of a WSS random process $w(t)$ amplitude-modulated by a deterministic signal $m(t)$:

$$x(t) = m(t)w(t). \tag{2.14}$$

We will provide mathematical expressions and visualizations of the two-dimensional spectral statistics of these processes. We will also see that time-varying frequency content in $m(t)$ smears the spectral correlation, which motivates our desire later in this thesis to compensate for time-varying frequency content.

These AM-WSS processes abound in communications, where for example $m(t)$ is a high-frequency narrowband carrier and $w(t)$ is a random sequence of symbols to be transmitted. Such signals are also useful for modeling real-world signals such as machinery noise, particularly propeller noise in passive sonar.

When used as a model of propeller noise, $m(t)$ is a deterministic, real-valued, periodic modulator of the form

$$m(t) = \mathrm{Re}\left\{\sum_{k=1}^{K} a_k \exp\left(j2\pi k\phi_0(t)\right)\right\} \tag{2.15}$$

where $a_k \in \mathbb{C}$ and $\phi_0(t)$ is the fundamental instantaneous phase of $m(t)$ which is the integral of $f_0(t)$, the slowly time-varying instantaneous frequency: $\phi_0(t) = \int_0^t f_0(w)dw$. The component $w(t)$ is a zero-mean WSS random process. The number of harmonics $K$ corresponds to the number of blades on the propeller [38].

The signal in (2.14) is a nonstationary random process, because its time-varying autocovariance function is

$$\begin{aligned} r_{xx}(t,\tau) &= E\left[x(t)x(t-\tau)\right] = m(t)m(t-\tau)E\left[w(t)w(t-\tau)\right] \\ &= m(t)m(t-\tau)r_{ww}(\tau) \end{aligned} \tag{2.16}$$

If $w(t)$ is white noise with variance $\sigma_w^2$, $r_{xx}(t,\tau)$ has an even simpler form:

$$r_{xx}(t,\tau) = \begin{cases} m^2(t)\sigma_w^2, & \text{if } \tau = 0. \\ 0, & \text{otherwise.} \end{cases} \tag{2.17}$$

The $m(t)m(t-\tau)$ factor in (2.16) or, in the case of a white $w(t)$, $m^2(t)$ in (2.17), causes the autocovariance function to vary periodically over time. Thus, all nonstationarity in an AM-WSS signal is due entirely to the amplitude modulation of $m(t)$. If $f_0(t)$ is constant over all time, (2.14) is a cyclostationary [36] signal with cycle period $T_0 = 1/\mathcal{F}_0$, since $\mathcal{F}_0$ is the fundamental frequency of $m(t)$ and $m(t)$ caused periodic variation of the autocovariance

function $r_{xx}(t,\tau)$. If $f_0(t)$ is such that $m(t)$ is an almost-periodic function (that is, contains periodic components with coprime periods), (2.14) is an almost-cyclostationary signal [19].

### *2.3.1 Spectral second-order statistics of AM-WSS with constant fundamental modulation frequency*

The Fourier transform of an AM-WSS process $x(t)$ is

$$X(f) = \int M(f-u)dZ(u) \tag{2.18}$$

which is just the time-frequency dual to (2.14), where $dZ(f)$ is the complex-valued spectral increment process for $w(t)$ [32, (3.8)]. We desire to calculate the spectral Hermitian covariance $R_{XX}(f_1,f_2)$ and the spectral complementary covariance $\tilde{R}_{XX}(f_1,f_2)$. Note that $E[X(f)] = 0$ because $E[dZ(f)] = 0$, so

$$R_{XX}(f_1,f_2) = S_{XX}(f_1,f_2) = E[X(f_1)X^*(f_2)] \tag{2.19}$$

$$\widetilde{R}_{XX}(f_1,f_2) = \widetilde{S}_{XX}(f_1,f_2) = E[X(f_1)X(f_2)]. \tag{2.20}$$

We will compute analytic expressions for $R_{XX}(f_1,f_2)$ and $\widetilde{R}_{XX}(f_1,f_2)$ for two cases: first, when $w(t)$ is white noise with variance $\sigma_w^2$, and second, when $w(t)$ is an arbitrary zero-mean WSS process with autocovariance function $r_{ww}(\tau)$ and spectral density $S_w(f)$.

If $w(t)$ is white noise with variance $\sigma_w^2$, the spectral Hermitian covariance is

$$R_{XX}(f_1,f_2) = \sigma_w^2 \int m^2(t)e^{-j2\pi t(f_1-f_2)}dt. \tag{2.21}$$

Likewise, the spectral complementary covariance is

$$\widetilde{R}_{XX}(f_1,f_2) = \sigma_w^2 \int m^2(t)e^{-j2\pi t(f_1+f_2)}dt. \tag{2.22}$$

See appendix A for proofs.

Equations (2.21) and (2.22) are interesting, because they correspond to the DEMON ("demodulated noise") spectrum evaluated at $f_1-f_2$ and $f_1+f_2$, respectively. The DEMON spectrum is a commonly-used tool in the sonar community to examine the frequency content of $m(t)$ (that is, the modulation frequency content of $x(t)$). The DEMON spectrum of a

real-valued signal $x(t)$ is the Fourier transform of the squared signal:

$$D_x(f) = \int x^2(t) e^{-j2\pi ft} dt \tag{2.23}$$

The squaring operation is a nonlinearity that performs a demodulation operation (essentially demodulating the signal with itself). If $m(t)$ has constant fundamental frequency $\mathcal{F}_0$ and consists of harmonics, $|D_x(f)|$ exhibits spectral peaks at multiples of $\mathcal{F}_0$.

When $w(t)$ is not white noise (i.e., when $S_w(f)$ is not constant for all $f$), the expressions for the bifrequency spectral covariances become more complicated, and there is no elegant form that emerges as in (2.21) and (2.22). These expressions are

$$R_{XX}(f_1, f_2) = \int M(f_1 - u) M(f_2 - u) S_w(u) du \tag{2.24}$$

and

$$\widetilde{R}_{XX}(f_1, f_2) = \int M(f_1 - u) M(f_2 + u) S_w(u) du. \tag{2.25}$$

Again, see appendix A for proofs.

The PSD $S_w(f)$ acts as a kernel within the integrals (2.24) and (2.25), which complicates the expressions. However, the expressions for the spectral covariances in the global/local representation yield elegant forms. Using the variable substitution $f_1 \leftarrow f$ and $f_2 \leftarrow (f - \nu)$, we can define the two-dimensional global/local Hermitian and complementary spectral covariances:

$$R_{XX}(f, \nu) = E\left\{X(f) X^*(f - \nu)\right\} \tag{2.26}$$

and

$$\widetilde{R}_{XX}(f, \nu) = E\left\{X(f) X(f - \nu)\right\} = E\left\{X(f) X^*(-f + \nu)\right\}. \tag{2.27}$$

Since $M(f) = \frac{1}{2} \sum_{k=1}^{K} a_k \left[\delta(f + k\mathcal{F}_0) + \delta(f - k\mathcal{F}_0)\right]$, and we use $W(f)$ to represent the spectrum of $w(t)$[4], we can write $X(f)$ as

$$X(f) = M(f) * W(f) = \frac{1}{2} \sum_{k=1}^{K} a_k \left[W(f + k\mathcal{F}_0) + W(f - k\mathcal{F}_0)\right]. \tag{2.28}$$

[4]Note we are using the finite-energy assumption here. WSS signals can never have finite energy, so this is more of a notational convenience. See appendix A in Clark's thesis [32] for an explanation of this.

Thus, $X(f)$ is a sum of shifted copies of $W(f)$. Note that these copies of $W(f)$ may overlap. Now when we evaluate the Hermitian or complementary spectral covariance, it is easy to see that $X^*(f)$ will be shifted by $\nu$:

$$R_{XX}(f,\nu) = E\left\{X(f)X^*(f-\nu)\right\} \tag{2.29}$$

and

$$\widetilde{R}_{XX}(f,\nu) = E\left\{X(f)X(f-\nu)\right\} = E\left\{X(f)X^*(-f+\nu)\right\}. \tag{2.30}$$

For simplicity of derivation here, assume $a_k = 1$ for $k = 1, ..., K$. When a shifted copy of $W(f)$, $W(f-\nu)$, lines up with a conjugate shifted copy of $W(f)$, $W^*(f-\nu)$, the PSD of $w(t)$ results: $E\left\{W(f)W^*(f)\right\} = S_w(f)$. Since $w(t)$ is real-valued, its spectrum is conjugate symmetric: $W(f) = W^*(-f)$, which means that $E\left\{W(f)W(-f)\right\} = E\left\{W(f-\nu)W^*(f-\nu)\right\} = S_w(f-\nu)$. For the current example, shifted copies of $W(f-\nu)$ and conjugate shifted copies $W^*(f-\nu)$ only line up when $\nu$ is a multiple of $\mathcal{F}_0$. Thus, the only values of $\nu$ for which $R_{XX}(f,\nu)$ and $\widetilde{R}_{XX}(f,\nu)$ are non-zero are integer multiples of $\mathcal{F}_0$. Thus, for $\nu = n\mathcal{F}_0$, $n$ an integer, $R_{XX}(f,\nu)$ and $\widetilde{R}_{XX}(f,\nu)$ consist of the sum of shifted copies of $S_w(f)$.

The global/local Hermitian and complementary spectral correlation are shown in figure 2.3 for the case when $w(t)$ is white noise and in figure 2.4 when $w(t)$ is a WSS process (specifically in this case, $w(t)$ is an autoregressive process).

Notice that figure 2.3 indicates that the one-dimensional DEMON spectrum is sufficient when $w(t)$ is white (i.e., has a constant PSD), because the DEMON spectrum is equal to the Hermitian and complementary global/local spectral correlation along a horizontal line $f = f_a$, where $f_a$ is any frequency, or along the anti-diagonal $f_1 = -f_2$ of bifrequency Hermitian and the diagonal $f_1 = f_2$ of bifrequency complementary. This is emphasized in figure 2.3, where the value of the bifrequency and global/local spectral covariances along the one-dimensional slices denoted by the dotted gray lines is shown in the right panel.

However, when $w(t)$ is colored WSS, the DEMON spectrum is no longer a perfect estimator of a one-dimensional slice through the spectral correlation. Note that in figure 2.4, not all horizontal slices through global/local spectral correlations are the same (left panels of figure 2.4). Similarly, not all anti-diagonal slices through Hermitian bifrequency spectral correlation are the same (top right panel of figure 2.4), and not all diagonal slices through

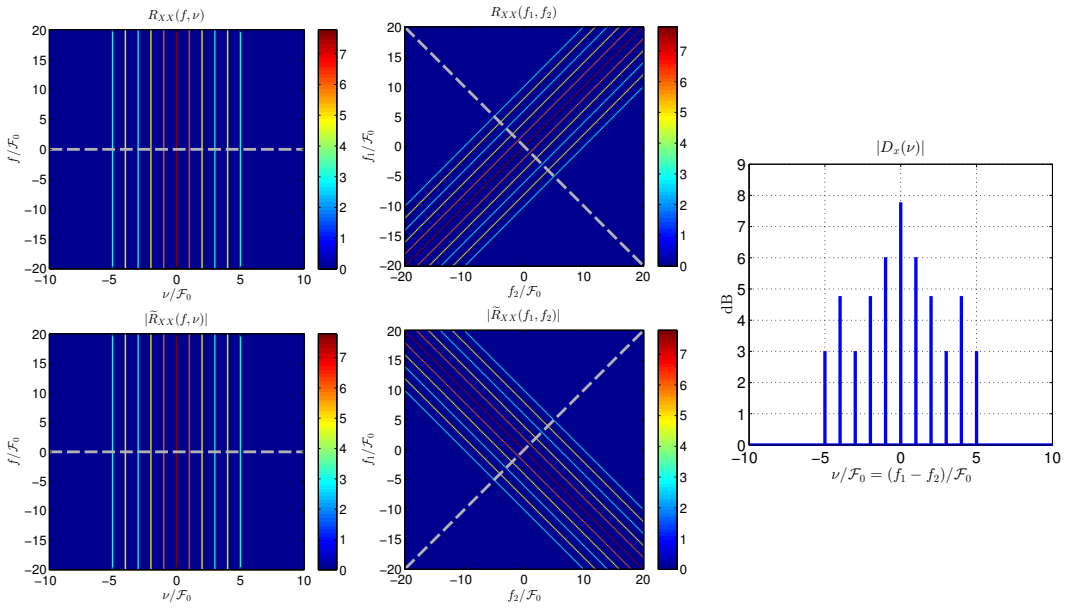


Figure 2.3: Left panels: global/local two-dimensional spectral correlation of an amplitude-modulated wide-sense stationary (AM-WSS) process when $w(t)$ is white, unit-variance Gaussian noise and $m(t)$ has $K = 3$ harmonics. Center panels: bifrequency two-dimensional spectral correlation for the same process. Top panels: Hermitian spectral correlation. Bottom panels: complementary spectral correlation. Far right panel: the DEMON spectrum $|D_x(\nu)|$ of the AM-WSS signal. Notice that the DEMON spectrum is equal to the spectral correlations in the left 4 panels along the dotted gray lines, and peaks occur at integer multiples of $\mathcal{F}_0$, the fundamental frequency of the modulator $m(t)$.

complementary bifrequency correlation (bottom right panel of figure 2.4) are the same. Thus, the equivalence of the DEMON spectrum and the spectral correlations only holds when $w(t)$ is white. Clark et al. proposed a multiband version of the DEMON estimator to address this problem [39].

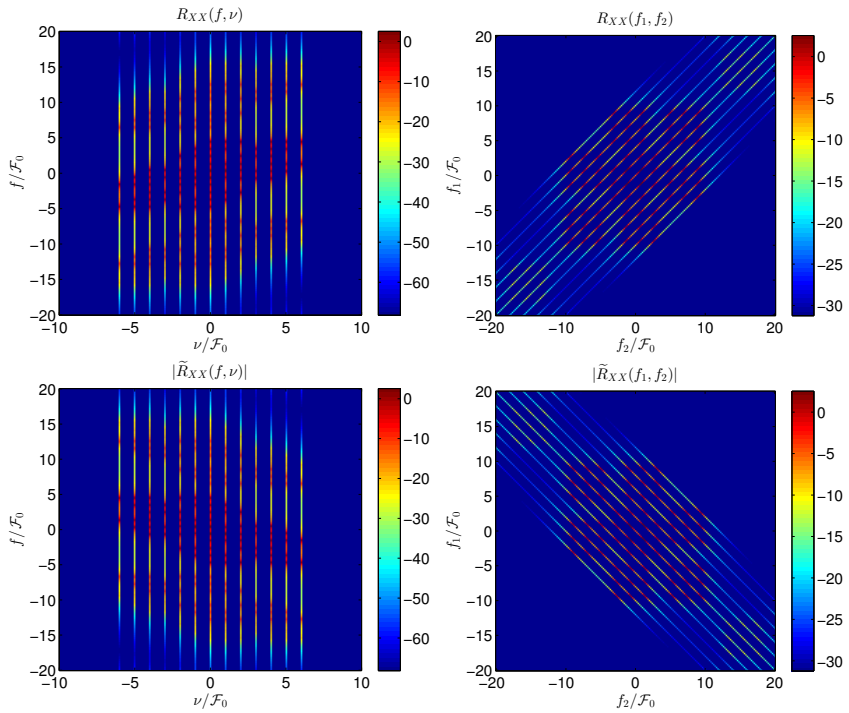


Figure 2.4: Global/local two-dimensional spectral correlation (left panels) and bifrequency two-dimensional spectral correlation (right panels) of an amplitude-modulated wide-sense stationary (AM-WSS) process when $w(t)$ is a WSS autoregressive process and $m(t)$ has $K = 3$ harmonics and constant fundamental frequency: $f_0(t) = \mathcal{F}_0$. Top panels are Hermitian spectral correlation, bottom panels are complementary spectral correlation. Lines occur at integer multiples of $\mathcal{F}_0$, the fundamental frequency of the modulator $m(t)$.

### *2.3.2 Spectral second-order statistics of AM-WSS with time-varying fundamental modulation frequency*

Most prior work using the DEMON spectrum has assumed that $f_0(t) = \mathcal{F}_0$; that is, the fundamental modulation frequency is constant over the duration of the signal $x(t)$, and thus $m(t)$ is perfectly periodic. In this case, an AM-WSS process is cyclostationary, with a cycle period of $T_0 = 1/\mathcal{F}_0$. Real-world data, however, often has time-varying modulation frequency (e.g., a ship that is speeding up or slowing down), and thus does not always have perfectly periodic modulation.

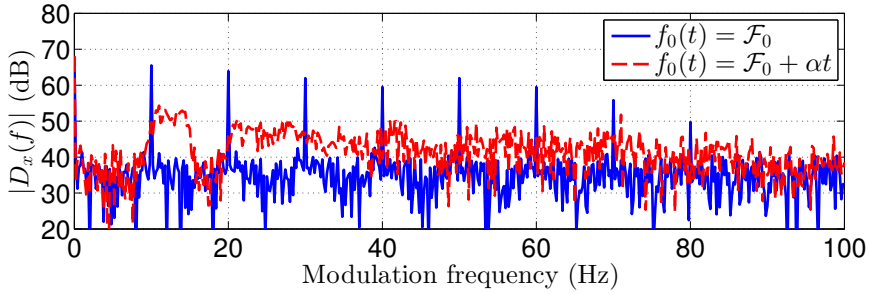


Figure 2.5: Illustration of smearing of the estimated DEMON spectrum when $m(t)$ has either constant or linearly-varying fundamental modulation frequency $f_0(t)$. In both cases $w(t)$ is white noise, and $m(t)$ has $K = 3$ harmonics and constant fundamental frequency $f_0(t) = \mathcal{F}_0$. These lines are equivalent to an estimate of one side of the far right panel of figure 2.3.

Time-varying frequency content in $m(t)$ smears the spectral correlation function, which also corresponds to smearing energy in the DEMON spectrum. This effect is illustrated in figure 2.5, where the estimated DEMON spectrum of two types of AM-WSS signals are shown. For both AM-WSS signals, $w(t)$ is unit-variance white noise. The blue curve corresponds to constant fundamental modulation frequency, $\mathcal{F}_0 = 10$ Hz, and $K = 3$ harmonics for $m(t)$. This results in narrow peaks in the DEMON spectrum at multiples 10 Hz. In the second case, the fundamental modulation frequency of $m(t)$ is $f_0(t) = 10 + \frac{5}{T}t$. That is, $m(t)$ has a starting frequency of 10 Hz and total frequency change of 5 Hz over the duration $T$. Because of the time-varying frequency content in $m(t)$, the DEMON spectrum is smeared out, resulting in a reduced magnitude at the spectral locations at multiples of 10 Hz. As we will see in chapter 4, to correct this smearing we can perform the equivalent of a time-warping of the squared signal before taking the Fourier transform, which restores the narrow peaks as long as the time-warping uses the correct instantaneous frequency trajectory.

Though we will not plot two-dimensional spectral correlations for the linearly-varying $f_0(t)$ case here, smearing similar to the one-dimensional DEMON spectrum in figure 2.5 will occur in the two-dimensional spectral correlation functions.

We will see more of the AM-WSS processes in chapter 4, where we present an improved

detection scheme that takes advantage of time-varying $f_0(t)$.

### *2.4 Filtered Jittered Pulse Train (F-JPT) Processes*

This section defines a filtered jitter pulse train (F-JPT) process and uses it is a simplified model for voiced speech. Then we give both mathematical expressions and visualizations of the spectral correlations of the F-JPT under different conditions, which demonstrates the additional information present in the spectral correlation that is often disregarded by using the stationary assumption (i.e. that short segments of voiced speech are stationary, and thus have uncorrelated spectral increments).

Define a finite-duration F-JPT process as

$$x(t) = h(t) * d(t) = h(t) * \sum_{n=-N}^{N} g(t - n\mathcal{T}_0 + k_n) \tag{2.31}$$

where $h(t)$ is the impulse response of an linear time-invariant (LTI) filter and $d(t)$ is a jittered pulse train with constant fundamental frequency $\mathcal{F}_0 = 1/\mathcal{T}_0$. Here we will take the pulse shape $g(t)$ to be a unit-energy rectangular pulse of width $T_g$:

$$g(t) = \begin{cases} \frac{1}{\sqrt{T_g}}, & \text{for } -\frac{T_g}{2} \leq t \leq \frac{T_g}{2} \\ 0 & \text{otherwise.} \end{cases} \tag{2.32}$$

Ultimately, we will be interested in the limit case when $T_g \to 0$. The signal $x(t)$ is assumed to be of finite duration $T$, and we assume that $2N + 1$ pulses of duration $\mathcal{T}_0 = 1/\mathcal{F}_0$ fit within this interval. If $x(t)$ is used as a model of voiced speech, $h(t)$ can be used to model all-pole formant filter that simulates the vocal tract and/or a glottal pulse shape. The pulse train $d(t)$ models the timing of periodic glottal excitation. Using a pulse train to model glottal pulses is a common method in speech modeling [1]. Jittering of the pulses has also been proposed to model small fluctuations in glottal cycle length [40], and the estimated jitter of glottal pulses has also been proposed as features for automatic speech recognition [41]. The random variable $k_n$ controls the jitter, and is uniformly-distributed with a range of $[-D/2, D/2]$, where $D$ defines the allowable range of jitter of an individual pulse. The $k_n$ are independent and identically distributed for all $n$.

### *2.4.1 Spectral second-order statistics of F-JPT with constant fundamental frequency*

In this subsection, we will give mathematical expressions for the Hermitian and complementary spectral covariances, (2.7) and (2.8), of the F-JPT when the fundamental frequency is constant: $f_0(t) = \mathcal{F}_0$. After we give these mathematical expressions, we will present a series a visualizations that highlight the differences between the spectral covariance functions and demonstrate the importance of accounting for the spectral covariance of voiced speech, especially cross-frequency covariance.

The Fourier transform of $x(t)$ is

$$X(f) = H(f) \sum_{n=-N}^{N} G(f) e^{-j2\pi f(n\mathcal{T}_0 - k_n)}. \tag{2.33}$$

The Hermitian spectral covariance[5] of a F-JPT process when the pulse width $T_g$ is infinitely narrow is:

$$\begin{aligned} R_{XX}(f_1, f_2) = H(f_1)H^*(f_2) &\left( \frac{\sin\left(\pi D(f_2 - f_1)\right)}{\pi D(f_2 - f_1)} - \frac{\sin(\pi D f_1)}{\pi D f_1} \frac{\sin(\pi D f_2)}{\pi D f_2} \right) \\ &\cdot \frac{\sin\left[\pi(2N+1)\mathcal{T}_0(f_2 - f_1)\right]}{\sin(\pi \mathcal{T}_0(f_2 - f_1)} \end{aligned} \tag{2.34}$$

and

$$\begin{aligned} \widetilde{R}_{XX}(f_1, f_2) = H(f_1)H(f_2) &\left( \frac{\sin\left(\pi D(f_2 + f_1)\right)}{\pi D(f_2 + f_1)} - \frac{\sin(\pi D f_1)}{\pi D f_1} \frac{\sin(\pi D f_2)}{\pi D f_2} \right) \\ &\cdot \frac{\sin\left[\pi(2N+1)\mathcal{T}_0(f_2 + f_1)\right]}{\sin(\pi \mathcal{T}_0(f_2 + f_1)}. \end{aligned} \tag{2.35}$$

Note that in (2.34) and (2.35), the jitter causes a multiplication of the spectral correlation by a combination of sinc functions.

We will now provide visualizations of the expressions (2.34) and (2.35) and compare them to the two-dimensional bifrequency spectral covariances of a comparable WSS process. We will see that the Hermitian and complementary spectral covariances of the nonstationary

[5]Here, we give mathematical expressions and visualizations of the spectral covariances. Note that for F-JPT processes, the mean of the spectral representation $X(f)$ is not zero, as it was for AM-WSS processes. This means that the spectral correlations are not equal to the spectral covariances, i.e. $S_{XX}(f_1, f_2) \neq R_{XX}(f_1, f_2)$ and $\widetilde{S}_{XX}(f_1, f_2) \neq \widetilde{R}_{XX}(f_1, f_2)$. See appendix B for mathematical expressions for the bifrequency Hermitian and complementary spectral correlations $S_{XX}(f_1, f_2)$ and $\widetilde{S}_{XX}(f_1, f_2)$ that include these mean terms.

F-JPT provide additional information beyond that of the standard one-dimensional power spectral density $S_x(f)$.

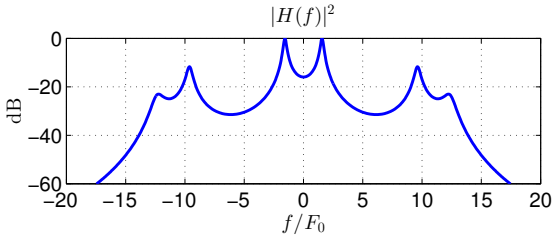


Figure 2.6: Magnitude-squared response of formant filter used for F-JPT.

The parameters of the F-JPT process we will use are $\mathcal{F}_0$=210Hz, which yields a period of $\mathcal{T}_0 = 1/\mathcal{F}_0 = 4.76$ms. The covariances are evaluated on an evenly-spaced frequency grid $\{f_1, f_2\}$ that has 20 evenly-spaced frequencies between each harmonic of $F_0$. For example, $f_i$ has 20 points between $\mathcal{F}_0$ and $2\mathcal{F}_0$. A maximum number of 20 harmonics was chosen, yielding a cross-frequency grid of $401 \times 401$ points. The duration is $T = 33.3$ms (which allows exactly $2N + 1 = 7$ pulses in the duration). A jitter of 10% of the fundamental period $\mathcal{T}_0$ is used. This jitter amount is used because informal listening tests indicated that this is the maximum amount of jitter for which a F-JPT still sounds somewhat like real voiced speech, which indicates that the jitter is modeling the randomness of glottal pulses. The magnitude-squared response of the formant filter $h(t)$ is shown in figure 2.6.

The bifrequency Hermitian spectral covariances are visualized in figure 2.7 for two signals, one signal which is WSS and the other signal which is a nonstationary F-JPT process. The WSS signal is the formant filter from figure 2.6 driven by white Gaussian noise with unit variance, instead of a jittered pulse train. The important thing to notice about these visualizations is that the two-dimensional bifrequency Hermitian spectral covariance $R_{XX}(f_1, f_2)$ of F-JPT in the right panel exhibits substantial covariance off the diagonal $f_1 = f_2$, while $R_{XX}(f_1, f_2)$ of a WSS signal (white noise driving the formant filter) is only non-zero on the diagonal $f_1 = f_2$ and 0 everywhere else. This non-zero off-diagonal covariance is information that is not usually exploited when the data is assumed to be WSS.

F-JPT processes also exhibit substantial complementary spectral covariance. Figure 2.8

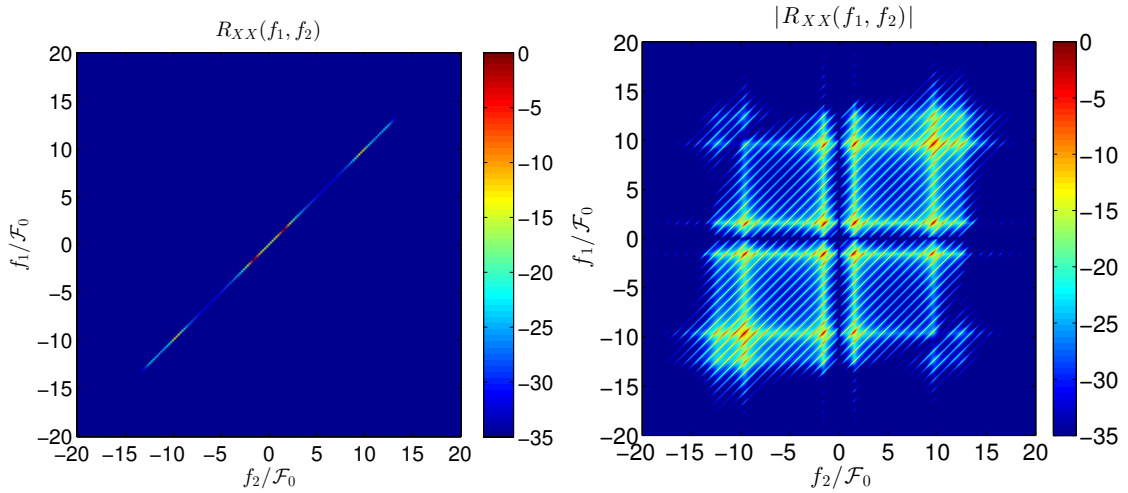


Figure 2.7: Left: two-dimensional bifrequency Hermitian spectral covariance $R_{XX}(f_1, f_2)$ of a wide-sense stationary signal (unit-variance white noise driving a formant filter). Right: two-dimensional bifrequency Hermitian spectral covariance $R_{XX}(f_1, f_2)$ of a finite-duration, nonstationary F-JPT signal (a model of voiced speech) using the same formant filter and a constant fundamental frequency $f_0(t) = \mathcal{F}_0$. Note that the nonstationary F-JPT exhibits substantial spectral covariance off the diagonal $f_1 = f_2$ (right panel), which is additional information not usually exploited when using an assumption of stationarity.

illustrates this point. The left panel of figure 2.8 shows the magnitude of the bifrequency complementary spectral covariance $|\widetilde{R}_{XX}(f_1, f_2)|$ of a WSS signal that is the formant filter driven by unit-variance white noise. The right panel shows $|\widetilde{R}_{XX}(f_1, f_2)|$ for the same nonstationary F-JPT process as in figure 2.7. Notice that the parallel lines corresponding to correlations between pulse train harmonics go the opposite direction from the parallel lines in the Hermitian spectral covariance in the right panel of figure 2.7 (anti-diagonal, instead of diagonal). As described in section 2.2.2, the bifrequency complementary spectral covariance of wide-sense stationary random processes is only non-zero on the line $f_1 = -f_2$. The standard PSD $S_x(f)$ lies along this line, which means that the complementary spectral covariance is completely redundant for wide-sense stationary random processes (this anti-diagonal orientation of the PSD also justifies the conventional approach of only

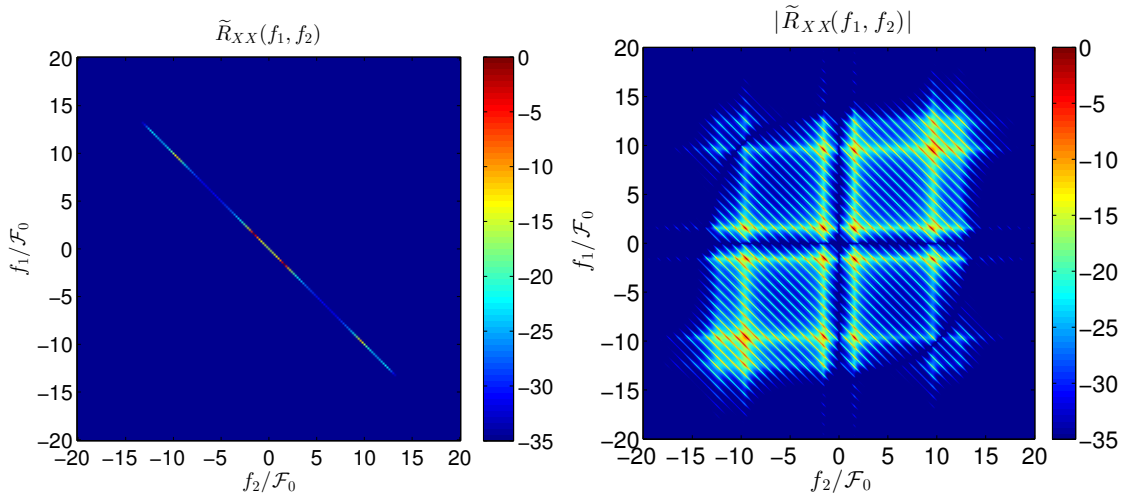


Figure 2.8: Left: magnitude of two-dimensional bifrequency complementary spectral covariance $|\widetilde{R}_{XX}(f_1, f_2)|$ of a wide-sense stationary signal (unit-variance white noise driving a formant filter). Right: two-dimensional bifrequency complementary spectral covariance $|\widetilde{R}_{XX}(f_1, f_2)|$ of a nonstationary F-JPT signal (a model of voiced speech) using the same formant filter and a constant fundamental frequency $f_0(t) = \mathcal{F}_0$. Note that the nonstationary F-JPT exhibits substantial spectral covariance off the anti-diagonal $f_1 = -f_2$, which is additional information not usually exploited when using an assumption of stationarity.

examining the spectral covariance along the stationary manifold $f_1 = f_2$ when processing WSS signals). However, for nonstationary F-JPT processes, the bifrequency complementary spectral covariance does carry additional information about the process because it is not just a rotation of the Hermitian spectral covariance, shown in the right panel of figure 2.7.

### *2.4.2 Spectral second-order statistics of F-JPT with time-varying fundamental frequency*

Now consider a finite-duration F-JPT process when the fundamental frequency varies linearly over time, achieving an instantaneous fundamental frequency of $\mathcal{F}_0 - \alpha\frac{T}{2}$ at time $t = -T/2$, $\mathcal{F}_0$ at time $t = 0$, and $\mathcal{F}_0 + \alpha\frac{T}{2}$ at time $t = T/2$. For this case, define the time-varying period

$$T_0[n] = \frac{n}{\alpha\frac{n}{\mathcal{F}_0} + \mathcal{F}_0} \tag{2.36}$$

where $\alpha$ plays the role of a slope, or chirp rate, in units of Hertz per second. Note that if $\alpha = 0$, (2.36) simplifies to $T_0[n] = n\frac{1}{\mathcal{F}_0} = n\mathcal{T}_0$, which is equivalent to the constant fundamental frequency case in (2.31). For convenience, we assume here that exactly $2N+1$ periods with the varying lengths given in (2.36) fit perfectly within the duration $T$. Thus, we have the following expression for a finite-duration $x(t)$:

$$x(t) = h(t) * \sum_{n=-N}^{N} \delta(t - T_0[n] + k_n) \tag{2.37}$$

Since the fundamental period is now a function of time, there is no clean expression for the equivalent last multiplicative term in the spectral covariances, which for a constant-frequency, infinite-duration F-JPT was a Dirac pulse train in (**??**) and (**??**), and for a constant-frequency, finite-duration F-JPT was a Dirichlet kernel in (2.34) and (2.35). Here, the cleanest expression is a finite sum of time-varying complex-exponentials:

$$R_{XX}(f_1, f_2) = H(f_1)H^*(f_2)\left(\frac{\sin\left(\pi D(f_2 - f_1)\right)}{\pi D(f_2 - f_1)} - \frac{\sin(\pi D f_1)}{\pi D f_1}\frac{\sin(\pi D f_2)}{\pi D f_2}\right) \cdot \sum_{n=-N}^{N} e^{j2\pi T_0[n] f} \tag{2.38}$$

and

$$\widetilde{R}_{XX}(f_1, f_2) = H(f_1)H(f_2)\left(\frac{\sin\left(\pi D(f_2 + f_1)\right)}{\pi D(f_2 + f_1)} - \frac{\sin(\pi D f_1)}{\pi D f_1}\frac{\sin(\pi D f_2)}{\pi D f_2}\right) \cdot \sum_{n=-N}^{N} e^{j2\pi T_0[n] f}. \tag{2.39}$$

*Empirical measurement of chirp rates $\alpha$ in real-world speech*

If the F-JPT is used as a model for voiced speech, one might wonder what range of $\alpha$s are encountered in the real-world. As a lower bound on this range, we empirically examined time-varying fundamental frequency data from the CSTR pitch-detection corpus [42]. This corpus provides time-varying fundamental frequency estimates for read speech from a male and a female speaker using a super-resolution pitch detection algorithm on laryngograph data (a laryngograph consists of electrodes attached to the neck that precisely measure vocal fold activity). The results are shown in figure 2.9. These histograms are generated

by finding all voiced segments of the desired duration $T$ by reading data from the provided ground-truth text files. Then a least-squares linear fit is performed on each segment. For each duration $T$ and each speaker gender, a histogram of the slopes of the 3000 segments with the highest signal-to-error ratio

$$\text{SER} = 10\log_{10}\frac{\|f_0[n]\|_2^2}{\|\hat{f}_0[n] - f_0[n]\|_2^2}, \tag{2.40}$$

where $f_0[n]$ is the true fundamental frequency and $\hat{f}_0[n] = a\frac{n}{N} + b$, $n = 1, ..., N$, is the linear fit to $f_0[n]$. The smallest SER for the chosen slopes was about 15dB. The slopes are normalized such that they have units of Hz/s: $\alpha = a/T$. The ranges of $\alpha$ represented by these histograms are a lower bound because read speech tends to have less variation of the fundamental frequency (compared to everyday conversational speech, which due to prosody and other effects can have a much more variable fundamental frequency).

*Visualizations of spectral second-order statistics*

We now show visualizations of the Hermitian and complementary spectral covariances when the fundamental frequency varies linearly over the finite duration $T$: $f_0(t) = \alpha t + \mathcal{F}_0$, for $t = -T/2$ to $t = T/2$. A chirp rate of $\alpha = \frac{10\text{Hz}}{33.3\text{ms}} = 300$ Hz/s is used.

Figure 2.10 compares the Hermitian bifrequency spectral covariance for an F-JPT with a constant fundamental frequency (left panel) with the Hermitian bifrequency spectral covariance of an F-JPT with a linearly-varying fundamental frequency (right panel). The effect of a linearly-varying fundamental frequency is a smearing of the diagonal lines. Notice that for the Hermitian covariance, the lines tend to be more smeared further away from the diagonal $f_1 = f_2$. These smeared lines tend to become attenuated, which indicates that linearly-varying fundamental frequency does not have too extreme of an effect on the Hermitian spectral covariance.

Figure 2.11 compares the complementary bifrequency spectral covariance for an F-JPT with a constant fundamental frequency (left panel) with the complementary bifrequency spectral covariance of an F-JPT with a linearly-varying fundamental frequency (right panel). Notice that the smearing of the anti-diagonal lines occurs more along the diagonal $f_1 = f_2$ in the complementary spectral covariance. The smearing here for the complementary

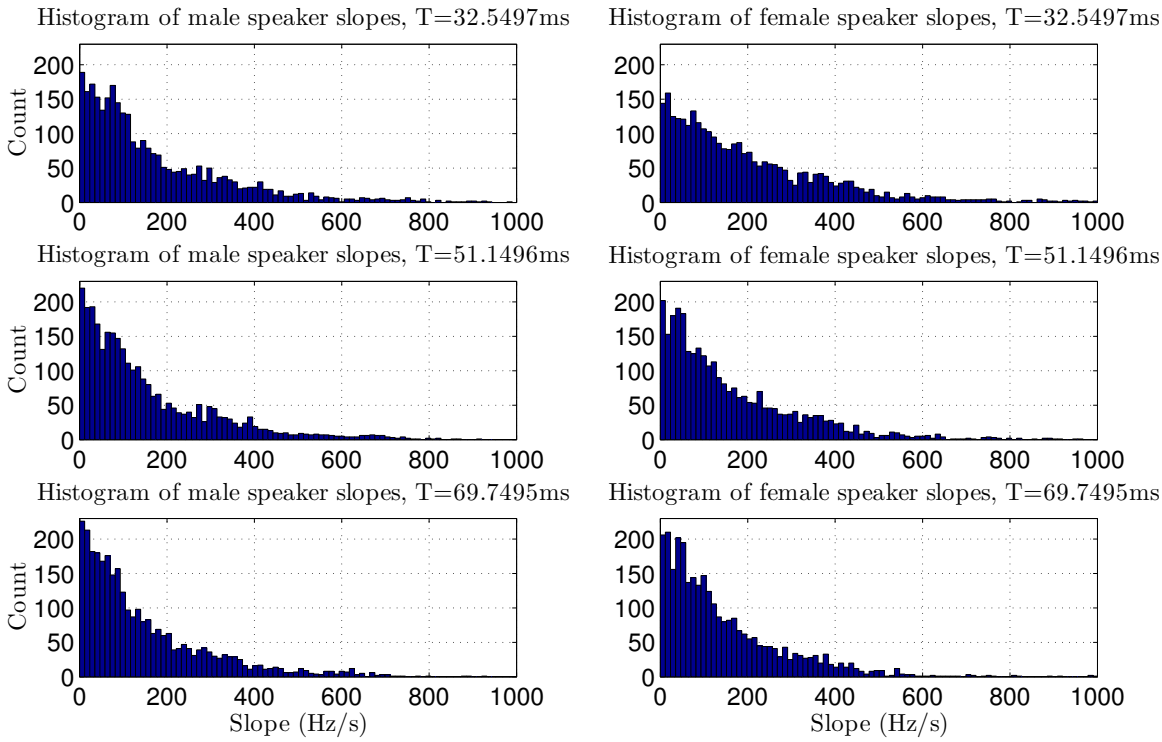


Figure 2.9: Histograms of the slopes $\alpha$ of linear fits to the fundamental frequency of real-world vowels from the CSTR pitch detection corpus [42]. These histograms show that the fundamental frequency of real-world voiced read speech exhibits a fair amount of variation over the specified durations $T$.

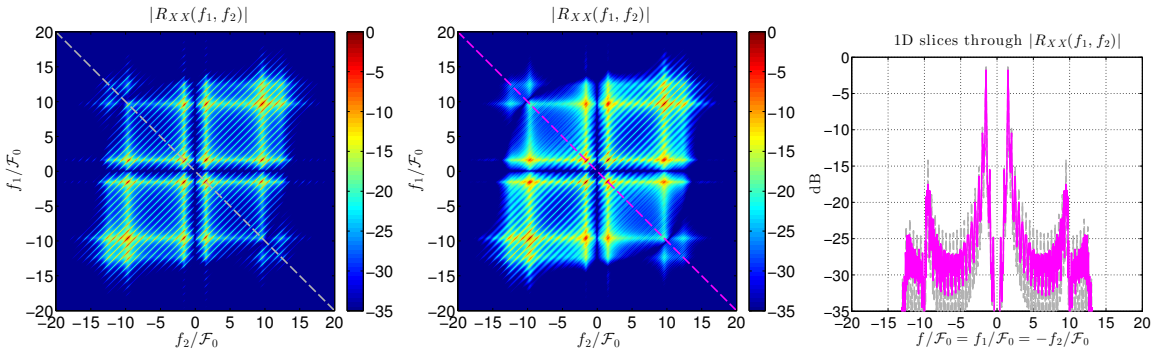


Figure 2.10: Left: two-dimensional bifrequency Hermitian spectral covariance $R_{XX}(f_1, f_2)$ of a F-JPT with constant fundamental frequency. Center: two-dimensional bifrequency Hermitian spectral covariance $R_{XX}(f_1, f_2)$ of a F-JPT with linearly-varying fundamental frequency. Right: 1D slices along the anti-diagonal of the left and center panels. Notice the broadening and smearing of diagonal lines that occurs in the center panel (magenta 1D slice) relative to the left panel (grey 1D slice) caused by the time-varying fundamental frequency.

covariance is more sensitive to linearly-varying fundamental frequency than the smearing in the Hermitian covariance (right panel of figure 2.10.

This smearing is an undesirable effect, because it attenuates the value of the peaks along the diagonal and spreads energy to adjacent points. As we will see in chapter 3, this smearing can be corrected by estimating the time-varying fundamental frequency and compensating it via time-warping. This effectively makes the smeared out lines in the spectral covariance narrow again.

### *2.5 Summary and Future Work*

In this chapter we have examined the two-dimensional spectral covariance functions of two subclasses of nonstationary random processes. Two dimensions are necessary because nonstationary random processes have spectral increments are correlated between different frequencies. Furthermore, the spectral increments may be correlated with the conjugates of themselves, which requires both Hermitian and complementary moments to fully specify the

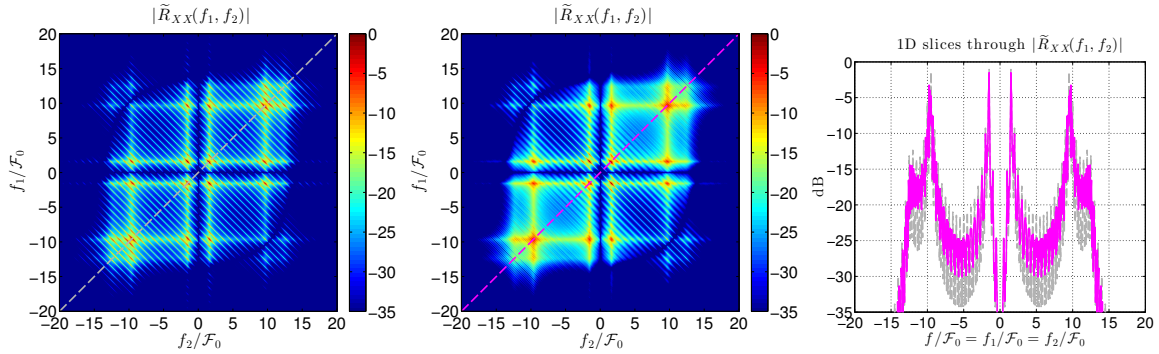


Figure 2.11: Left: magnitude of two-dimensional bifrequency complementary spectral covariance $|\widetilde{R}_{XX}(f_1, f_2)|$ of a F-JPT with constant fundamental frequency. Center: two-dimensional bifrequency complementary spectral covariance $|\widetilde{R}_{XX}(f_1, f_2)|$ of a F-JPT with linearly-varying fundamental frequency. Right: 1D slices along the diagonal of the left and center panels. Notice the broadening and smearing of anti-diagonal lines that occurs in the center panel (magenta 1d slice) relative to the left panel (grey 1D slice) caused by the time-varying fundamental frequency. Also, note the smearing occurs along the stationary manifold $f_1 = f_2$ in the center panel, in contrast to the anti-diagonal smearing in the Hermitian spectral correlation in the center panel of figure 2.10.

behavior of the complex-valued spectral increments. We also examined the effect of linearly-varying fundamental frequency on these spectral covariance functions, and we found that any time variation in fundamental frequency smears out narrow lines in the two-dimensional spectral covariances.

Narrow lines in the spectral covariances are important, especially if we desire to use these covariances for improved statistical signal processing. This is because making the lines in the spectral covariances more narrow makes the signal of interest more cyclostationary, which could allow the vast array of useful algorithms designed for cyclostationary signals to be applied for the first time to natural signals, such as speech and ship noise in passive sonar. Algorithms designed for cyclostationary signals exploit the cross-frequency spectral correlation of such signals at a finite number of frequency shifts (which are the cycle frequencies), and thus these algorithms work best when cross-frequency spectral correlations are narrow

and maximized. Blind versions of these algorithms only require *a priori* knowledge of the cycle frequencies of a desired signal. These algorithms have tantalizing applications, such as blind beamforming [3]–[5], blind source separation [2], improved minimum mean-square error filtering [6], and improved source location and detection [7]. Furthermore, because the spectral increments of our processes of interest are correlated with their conjugates, widely-linear versions of these algorithms [5], [27], [43] could improve performance even more.

# Chapter 3

# ESTIMATION AND COMPENSATION OF TIME-VARYING FREQUENCY CONTENT

## *3.1 Introduction*

This section describes a method for estimating the time-varying frequency content of random processes that have the structure described in chapter 2. We then describe a transform that then uses this information to compensate signals such that they are more stationary. We also present some measurements of how effective this method is.

Section 3.2 describes an existing method, the fan-chirp transform, that we can use to implement our novel method. This transform forms the basis for our approach in chapter 5. Section 3.3 proposes a measure for the stationarity of speech-like data, and demonstrates the effectiveness of time-warping to make short frames of data approximately stationary.

## *3.2 Analysis and Synthesis Using the Fan-Chirp Transform*

In this section, we review the forward and inverse fan-chirp transform. The forward fan-chirp transform uses a linear model of time-varying fundamental frequency to estimate the chirp rate (i.e., the slope of the linear fit) of the fundamental frequency and then uses a time-warping to compensate for this time-varying fundamental frequency. The combination of the forward and inverse fan-chirp transforms provides an analysis-synthesis framework that provides almost perfect reconstruction, which we will find useful for speech processing in chapter 5.

### *3.2.1 The forward fan-chirp transform*

We adopt the fan-chirp transform formulation used by Cancela et al. [44]. The forward fan-chirp transform is defined as

$$X(f,\alpha) = \int x(t)\phi'_\alpha(t)e^{-j2\pi f\phi_\alpha(t)}dt \tag{3.1}$$

where $\phi_\alpha(t) = \left(1 + \frac{1}{2}\alpha t\right) t$ and $\phi'_\alpha(t) = 1 + \alpha t$. The variable $\alpha$ is an analysis chirp rate. Using a change of variable $\tau \leftarrow \phi_\alpha(t)$, (3.1) can be written as the Fourier transform of a time-warped signal:

$$X(f, \alpha) = \int_{-\infty}^{\infty} x(\phi_\alpha^{-1}(\tau)) e^{-j2\pi f \tau} d\tau. \tag{3.2}$$

The short-time fan-chirp transform (STFChT) of $x(t)$ is defined as the fan-chirp transform of the $d$th short frame of $x(t)$:

$$X_d(f, \hat{\alpha}_d) = \int_{-T_w/2}^{T_w/2} w(\tau) x_d(\phi_{\hat{\alpha}_d}^{-1}(\tau)) e^{-j2\pi f \tau} d\tau \tag{3.3}$$

where $w(t)$ is an analysis window, $\hat{\alpha}_d$ is the analysis chirp rate for the $d$th frame given by (3.7), and $x_d(t)$ is the $d$th short frame of the input signal of duration $T$:

$$x_d(t) = \begin{cases} x(t - dT_{hop}), & -T/2 \le t \le T/2 \\ 0, & \text{otherwise.} \end{cases} \tag{3.4}$$

$T$ is the duration of the pre-warped short-time duration, $T_{hop}$ is the frame hop, $T_w$ is the post-warped short-time duration, and $w(t)$ is a $T_w$-long analysis window. The analysis window is applied after time-warping so as to avoid warping of the window, which can cause unpredictable smearing of the Fourier transform.

Implementing the fan-chirp transform as a time-warping followed by a Fourier transform allows efficient implementation, consisting simply as an interpolation of the signal followed by an FFT. In the implementation provided by Cancela et al. [44], the interpolation used in the forward fan-chirp transform is linear.

Kèpesi and Weruaga [45] provide a method for determination of the analysis chirp rate $\alpha$ using the gathered log spectrum (GLogS). The GLogS is defined as follows:

$$\rho(f_0, \alpha) = \frac{1}{N_h} \sum_{k=1}^{N_h} \ln |X(k f_0, \alpha)| \tag{3.5}$$

where $N_h$ is the maximum number of harmonics that fit within the analysis bandwidth. That is,

$$N_h = \left\lfloor \frac{f_s}{2 f_0 \left(1 + \frac{1}{2}|\alpha| T_w\right)} \right\rfloor. \tag{3.6}$$

Cancela et al. [44] proposed several enhancements to the GLogS. First, they observed improved results by replacing $\ln|\cdot|$ with $\ln(1+\gamma|\cdot|)$. Cancela et al. note that this expression approximates a $p$-norm, with $0 < p < 1$, where lower values of $\gamma$ with $\gamma \geq 1$ approach the 1-norm, while higher values approaches the 0-norm. Cancela et al. note that $\gamma = 10$ gave good results for their application.

Additionally, Cancela et al. propose modifications that suppress multiples and submultiples of the current $f_0$. Also, they propose normalizing the GLogS such that is has zero mean and unit variance. This is necessary because the variance of the GLogS increases with increasing fundamental frequency. For mean and variances measured over all frames in a database, a polynomial fit is determined and the GLogS are compensated using these polynomial fits.

Let $\bar{\rho}_d(f_0, \alpha)$ be the GLogS of the $d$th frame with these enhancements applied. For practical implementation, finite sets $\mathcal{A}$ of candidate chirp rates and $\mathcal{F}_0$ of candidate fundamental frequencies are used, and the GLogS is exhaustively computed for every chirp rate in $\mathcal{A}$ and fundamental frequency in $\mathcal{F}_0$. The analysis chirp rate $\hat{\alpha}_d$ for the $d$th frame is thus found by

$$\hat{\alpha}_d = \underset{\alpha \in \mathcal{A}}{\operatorname{argmax}} \max_{f_0 \in \mathcal{F}_0} \bar{\rho}_d(f_0, \alpha). \tag{3.7}$$

### *3.2.2 The inverse fan-chirp transform*

Inverting the fan-chirp transform is a matter of reversing the steps used in the forward transform. Thus, the inverse fan-chirp transform for a short-time frame consists of an inverse Fourier transform, removal of the analysis window, and an inverse time-warping. The removal of the analysis window $w(t)$ from the $T_w$-long warped signal limits the choice of analysis windows to positive functions only, such as a Hamming window, so the window can be divided out. Also, since the warping is nonuniform, it is possible that the sampling interval between points may exceed the Nyquist sampling interval. To combat this, the data should be oversampled before time-warping, which means the data must be downsampled after undoing the time-warping.

The choice of post-warped duration $T_w$ and the method of interpolation used in the inverse time-warping affect the reconstruction error of the inverse fan-chirp transform. There

is a trade-off between reconstruction performance and computational complexity, because interpolation error decreases as interpolation order increases. Kèpesi and Weruaga [46] analyzed fan-chirp reconstruction error with respect to order of the time-warping interpolation and oversampling factor, and found that for cubic Hermite splines and an oversampling factor of 2, a signal-to-error ratio of over 30dB can be achieved. For our application, we choose an oversampling factor of 8 and cubic-spline interpolation.

### *3.3 Measuring the Effect of Time-Warping*

Recall the definition of a direct spectral estimator of the power spectral density (PSD) $S_x(k)$ of discrete data $x[n]$ using the analysis window $h[n]$ [22], which is the magnitude-squared of the discrete Fourier transform (DFT) of windowed data:

$$S_x(k) = \left| \sum_{n=0}^{N-1} x[n]h[n]e^{j2\pi \frac{k}{K} n} \right|^2 . \tag{3.8}$$

Here we are interested in measuring the "peakiness" of the PSD, because if there is time-varying frequency content within an analysis frame, narrow peaks will become smeared, which indicates that the data within the frame has time-varying fundamental frequency. The goal of time-warping the frame is to compensate for the time-varying frequency content, which should make the spectrum more peaky.

We elect to use information-theoretic entropy as a measure of "peakiness", which intuitively corresponds to a measure of uncertainty [47]. If we normalize the nonnegative function $S_x(k)$ to be a probability distribution, entropy will be higher when the PSD has more evenly-distributed values across frequency. That is, the PSD resembles a uniform distribution. Conversely, if the PSD is more peaky, entropy will be smaller, as this case approaches a more deterministic distribution.

We will use a pulse train with a linearly time-varying fundamental frequency. We choose this test signal because it resembles glottal excitations. For simplicity, we do not impose a vocal tract filter, which is a simple multiplication of the vocal tract shape in the frequency domain (note that this application of a vocal tract filter decreases the entropy of the PSD). This means we are also assuming that the vocal tract filter does not change over the analysis window. Of course, for real-world speech, the assumption of a stationary vocal tract filter

is only valid for short frames on the order of 20-30ms [1]. Here, we justify not modeling the vocal tract filter by arguing that a slowly time-varying inverse filter or could remove the effect of a vocal tract filter, leaving only the speech harmonics behind.

The test signal, a sampled pulse train of duration $T$ with a frequency of $f_c$ in the middle of the $T$-long duration and a maximum frequency change of $\Delta f$ from $t = 0$ to $t = T$ is generated as

$$x[n] = \frac{1}{N_h} \sum_{k=1}^{N_h} \sin \left[ 2\pi \left( f_c - \frac{\Delta f}{2} + \frac{\Delta f}{T} \right) \frac{n}{f_s} \right], \tag{3.9}$$

where $n$ ranges from 0 to $\lfloor T f_s \rfloor$ and $N_h = \left\lfloor \frac{f_s/2}{f_c + \Delta f/2} \right\rfloor$ is the maximum number of harmonics that fit within the Nyquist rate.

The normalized entropy measure we will use is

$$\hat{H}(y) = \frac{H(y)}{\log K} = -\frac{1}{\log K} \sum_{k=0}^{K-1} y[k] \log y[k], \tag{3.10}$$

where $y[k]$ is a discrete probability mass function, and $\hat{H}(y)$ achieves a maximum of 1 when $y[k] = \frac{1}{K}$ for all $k$ (recall that the entropy $H(y)$ achieves a maximum of $\log K$ when $y[k]$ is the uniform distribution [47]).

For the experiment, synthetic signals were generated with six different total frequency changes over the analysis frame, ranging from 0 Hz to 35 Hz, which are chosen because they cover the range of frequency changes in real speech (for example, $\frac{35\,\text{Hz}}{30\,\text{ms}} = 1167$ Hz/s is the extreme of the range covered by the empirically-determined bounds in figure 2.9). For each of these total frequency changes, four different center frequencies ranging from 160 Hz to 220 Hz were used (these correspond to typical fundamental frequencies in real-world speech), and the normalized entropies are averaged over these center frequencies. The results are shown in figure 3.1. Notice that the normalized entropy of the PSD of the original data (left panel) decreases until about 30ms, which is the commonly accepted upper limit of the duration of analysis frames for speech. As frame duration increases beyond 30ms, the normalized entropy of the PSD tends to increase for the original frames (left panel), which indicates that unless frames are short, spectral peaks corresponding to voiced speech harmonics are smeared out across several frequency bins. Note that for the yellow line, where $\Delta f = 0$, the normalized entropy decreases monotonically as $T$ increases.

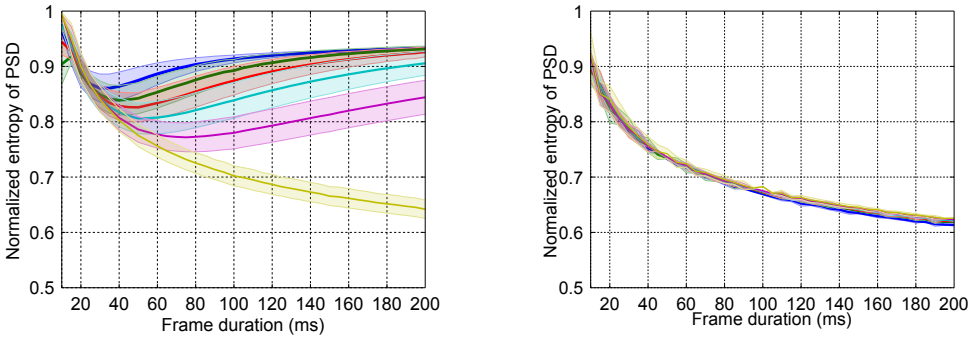


Figure 3.1: Normalized entropy of the estimated power spectral density (PSD) of the original (left) and of the time-warped (right) analysis frame versus duration of the frame for different chirp rates. Lower normalized entropy corresponds to more "peaky" PSDs. Shaded regions correspond to error bars that indicate standard deviation across 4 different center frequencies $f_c$ ranging from 160 Hz to 220 Hz.

However, in the right panel of figure 3.1, the normalized entropy of time-warped frames tends to only decrease as the frame duration increases. This monotonic decrease of the normalized entropy means that time-warping the signal concentrates spectral peaks. Concentrating energy into fewer FFT bins is important when an estimator is attempting to "classify" FFT bins as containing either signal-plus-noise or noise. In chapter 4, our detection statistic sums up harmonically-related peaks in the DEMON spectrum representing modulation frequency. For such a detection application, we desire peaks that have the highest amplitude and are as narrow as possible. For discrete sampled speech data, concentrating spectral peaks in acoustic frequency means that energy from speech harmonics is concentrated into fewer FFT bins, which we will see is a useful feature in chapter 5. In chapter 5, our speech enhancement estimator works by applying a gain that is a monotonically-increasing function of the signal-to-noise ratio (SNR). For such estimators, it is better to have a few bins with high SNR (i.e., that contain most of the signal energy) than to have many bins with low SNR (i.e., the signal energy spread out across many bins).

### *3.4 Summary*

In this chapter, we reviewed an existing implementation of our proposed method, the fan-chirp transform. The fan-chirp transform consists of a linear time-warping of the input signal followed by a Fourier transform, which allows efficient implementation using an interpolation followed by a FFT. The linear time-warping of the signal relies on estimation of the chirp rate $\alpha$, and we described an existing method of estimating $\alpha$.

We proposed a measure, normalized entropy, of the effectiveness of the fan-chirp transform, where the effectiveness corresponds to the ability of the fan-chirp transform to produce concentrated peaks in the power spectral density. Using synthetic data with varying degrees of time-varying fundamental frequency, we demonstrated that standard processing concentrates the peaks for increasing data duration $T$ only until a certain point (about $T = 30$ms); after this point the peaks quickly become smeared. The fan-chirp transform, on the other hand, continues to concentrate spectral peaks as the number of data samples increases, which shows the fan-chirp transform implements our method on synthetic data with similar parameters to that of voiced speech.

Chapter 4

# IMPROVED DETECTION OF AMPLITUDE-MODULATED WIDE-SENSE STATIONARY (AM-WSS) PROCESSES[1]

## *4.1 Introduction*

This chapter describes an application of our proposed paradigm to detecting ship propeller noise in passive sonar. The AM-WSS signals we described in chapter 2 are often used to model ship propeller noise. However, the common assumption used is that the speed of the propeller does not vary over the analysis interval. In reality, propeller speed often changes, especially in port situations as ships maneuver. We demonstrate improved detection performance on both synthetic and real passive sonar data, especially when the speed of the propeller varies during the analysis interval. In particular, we accomplish up to a 6dB increase in the detection statistic on real-world passive sonar data.

## *4.2 Prior Work*

Using the assumptions that $w(t)$ is white and that $m(t) = \mathrm{Re}\left\{\sum_{k=0}^{\infty} a_k e^{-j2\pi k\mathcal{F}_0 t}\right\}$, Lourens and du Preez [38] demonstrated that the DEMON spectrum (first introduced in section 2.3.1) is an approximate maximum-likelihood estimator (MLE) of the constant fundamental frequency of modulation, $f_0$. Clark et al. [39] demonstrated that for real-world sonar signals, $w(t)$ is often colored, and relaxed the assumption that $w(t)$ is white. They proposed a multiband version of the DEMON spectrum, which showed improved performance. Tao et al. [49] relaxed the assumption of constant fundamental frequency, assuming a linear chirp model for $m(t)$, and derived a MLE for propeller acceleration rate.

Similar approaches have been taken recently to modeling acoustic frequency variation, with most of the methods focused on speech processing. Omer and Torrésani [50] considered a model that consists of a frequency-modulated complex-valued WSS process and derived

[1]The contents of this chapter also appear in Wisdom et al. [48]. This paper is ©2014 IEEE, and portions are reused with permission.

an approximate ML estimator using Gabor frames. Kaewtip et al. [51] used time-warping to achieve a similar effect and used the approach to improve automatic speech recognition. Kepési and Weruaga [45] used the fan-chirp transform to fit linear instantaneous frequency (IF) trajectories to short frames of voiced speech.

Our approach differs from these acoustic frequency methods in two ways: first, we are looking at modulation frequency instead of acoustic frequency, and second, our method allows for an arbitrary model of trajectories (i.e., not just constant or linear). Furthermore, our ultimate goal differs somewhat: rather than examine estimator performance, our goal is extension of the coherence time, which is the amount of time during which an analysis method is coherent with the signal. Here, "coherent" refers to the stationarity of modulation frequency within the analysis interval.

In this chapter, we generalize the DEMON spectrum to allow for frequency variation in the modulator itself. We also show the relationship between the DEMON spectrum and the theory of spectral impropriety, namely that the DEMON spectrum is equivalent to a scaled spectral impropriety coefficient. Incorporating a model of the frequency variation into a GLRT detection statistic increases the coherence time of the estimator, allowing for the use of longer analysis windows and thus providing greater SNR.

## *4.3 Background*

This section covers the definition of the signal model we are interested in, the time-varying DEMON spectrum, and the generalized DEMON spectrum.

### *4.3.1 Time-varying DEMON spectrum*

If $m(t)$ has frequency content that varies over time, a time-variant version of the DEMON spectrum can be used for analysis. The time-varying DEMON spectrum is defined as follows:

$$D_x(d, f) = \int x_d^2(t) e^{-j2\pi ft} dt \tag{4.1}$$

with $x_d(t) = g(t)x(t - dT)$ is the $d$th frame of $x(t)$, where $g(t)$ is a data taper of duration $T$. Just as a short-time Fourier transform (STFT) is used to examine time-varying acoustic

frequency content, the time-varying DEMON spectrum is used to examine time-varying modulation frequency content.

### *4.3.2 Generalized DEMON spectrum*

To correct the "smearing" effect caused by time-varying frequency content in $m(t)$, as seen in figure 2.5, a generalized DEMON spectrum can be defined that is more coherent with the signal. If the DEMON spectrum is taken along an instantaneous frequency trajectory that is approximately equal to the true instantaneous frequency of $m(t)$, then the peaks at multiples of this trajectory in the generalized DEMON spectrum will be sharper and maximized.

Formally, define a generalized DEMON spectrum along an IF trajectory $f(t)$, where $f(t)$ is a function over the duration $T$ of $x(t)$. Let $\phi(t) = \int_0^t f(w)dw$ be the instantaneous phase corresponding to the trajectory $f(t)$. Then the generalized DEMON spectrum is defined to be

$$D_x^{(G)}(\phi) = \int x^2(t)e^{-j2\pi\phi(t)}dt \tag{4.2}$$

This formulation can use an arbitrary model of trajectories. However, to simplify derivation and implementation, the remainder of this chapter will use a linear model of IF trajectories, where $f(t) = f_c + \beta t$. In this model, $f_c$ is a center frequency, and $\beta$ is a chirp rate.

We can write an alternate formulation of (4.2) that shows it is equivalent with the fan-chirp transform (3.1) from chapter 3 of a time-dependent scaling of the squared signal. Let $\phi(t) = f_c\phi_\alpha(t)$ with $\phi_\alpha(t) = \left(\frac{1}{2}\alpha t + 1\right)t = \left(\frac{1}{2}\frac{\beta}{f_c}t + 1\right)t$, and $y(t) = x^2(t)/\phi'_\alpha(t) = x^2(t)/\left(\frac{\beta}{f_c}t + 1\right)$. Note that in this formulation, $\alpha = \frac{\beta}{f_c}$. It is easy to see that

$$D_x^{(G)}(\phi(t)) = \int_{-\infty}^{\infty} y(t)\phi'_\alpha(t)e^{-j2\pi f_c\phi_\alpha(t)} \tag{4.3}$$

is equal to the fan-chirp transform $Y\left(f_c, \frac{\beta}{f_c}\right)$ (3.1) in chapter 3 of $y(t) = x^2(t)/\phi'_\alpha(t) = x^2(t)/\left(\frac{\beta}{f_c}t + 1\right)$. This equality indicates that the generalized DEMON spectrum can also be expressed as a warping of a time-dependent scaling of the squared signal followed by a

Fourier transform. Though in this chapter, we will use the formulation in (4.2), as it makes the derivation of the optimum detection statistic easier.

### *4.4 Detection of AM-WSS Signals*

To illustrate the advantage of using the generalized DEMON spectrum, we consider the problem of detecting an AM-WSS signal $x(t)$ embedded in additive white noise $v(t)$. The noisy measurements are

$$y(t) = x(t) + v(t) = m(t)w(t) + v(t) \tag{4.4}$$

Let the variance of $v(t)$ be $\sigma_v^2$ and let $w(t)$ be white noise with variance 1. Assume $m(t)$, $w(t)$, and $v(t)$ are zero-mean.

The following derivation represents discrete data sampled at $f_s$ as $N$-length vectors, for example $\mathbf{y}$ where the $n$th element is $y[n]$ for $n \in [0, N-1]$. Here, we have the following two hypotheses:

$$\begin{aligned} \mathcal{H}_0 &: \mathbf{y} = \mathbf{v} \\ \mathcal{H}_1 &: \mathbf{y} = \mathbf{x} + \mathbf{v} \end{aligned} \tag{4.5}$$

We will use a generalized likelihood ratio test (GLRT) [52] as a detection statistic, which is defined as

$$\mathcal{L}(\mathbf{y}) \stackrel{\Delta}{=} \frac{\max_{\theta_\mathbf{1}} p(\mathbf{y}; \theta_\mathbf{1}, \mathcal{H}_1)}{\max_{\theta_\mathbf{0}} p(\mathbf{y}; \theta_\mathbf{0}, \mathcal{H}_0)} \tag{4.6}$$

where $\theta_\mathbf{i}$ are the unknown parameters under the hypothesis $\mathcal{H}_i$. Here, under $\mathcal{H}_1$, we take $\theta_\mathbf{1}$ to be the parameters that specify $f_0(t)$. Under $\mathcal{H}_0$, we have no unknown parameters.

If we examine the natural log of (4.6), we get

$$\ln \mathcal{L}(\mathbf{y}) = \ln \max_{\theta_\mathbf{1}} p(\mathbf{y}; \theta_\mathbf{1}, \mathcal{H}_1) - \ln p(\mathbf{y}; \mathcal{H}_0) \tag{4.7}$$

Since $\ln(\cdot)$ is a monotonically increasing function, we can push the max outside the log. Using the assumption $m^2[n] \ll \sigma_v^2$ for all $n$, and taking $p(\cdot)$ to be a Gaussian distribution,

simplification of (4.7) yields the following:

$$\ln \mathcal{L}(\mathbf{y}) \approx \max_{\boldsymbol{\theta_1}} \left[ -\frac{1}{2} \sum_{n=0}^{N-1} \ln \left( 1 + \frac{m^2[n]}{\sigma_v^2} \right) \dots + \frac{1}{2\left(\sigma_v^2\right)^2} \sum_{n=0}^{N-1} m^2[n] y^2[n] \right] \tag{4.8}$$

The first term inside the max of (4.8) can be approximated by $-\frac{N}{2}\text{SNR}$ using the assumption that $\frac{m^2[n]}{\sigma_v^2} \ll 1$ for all $n$ and the Taylor series expansion for $\ln(1+x)$. Since $-\frac{N}{2}\text{SNR}$ is a constant term, it can be pushed outside the max and absorbed into the likelihood threshold. The $\frac{1}{2(\sigma_v^2)^2}$ scaling factor on the second term in (4.8) is also data-independent and can be absorbed into the likelihood threshold. Thus, we can define the modified log-GLRT

$$L(\mathbf{y}) \triangleq \max_{\boldsymbol{\theta_1}} \sum_{n=0}^{N-1} m^2[n] y^2[n] \tag{4.9}$$

Under the conventional assumption that $f_0(t) = f_0$, Lourens and du Preez derived an expression for (4.9) [38, (17)-(21)] in terms of the DEMON spectrum, which yields

$$L(\mathbf{y}) = \max_{f_0} \sum_{k=0}^{\infty} |D(k f_0)|^2 \tag{4.10}$$

Likewise, using a linear model $f_0(t) = f_c + \beta t$, Tao et al. derived a similar expression for (4.9) in [49], which we can write in terms of the generalized DEMON spectrum, yielding

$$L^{(G)}(\mathbf{y}) = \max_{f_c, \beta} \sum_{k=0}^{\infty} \left| D_{\mathbf{y}}^{(G)} \left( k \left[ f_c + \frac{1}{2}\beta t \right] t \right) \right|^2 \tag{4.11}$$

Note that the $k$=0 terms in (4.10) and (4.11) are simply $\|\mathbf{y}\|_2^2$, so they are constant terms that are not affected by $\boldsymbol{\theta_1}$ and can thus be moved outside the maximization.

### *4.5 Results*

We now consider the effect of the length of the analysis window on detection performance and demonstrate that when fundamental modulation frequency varies within a window, $L^{(G)}(\mathbf{y}) \geq L(\mathbf{y})$ for both synthetic and real data. This result shows that our proposed method increases the coherence time and thus allows longer analysis windows.

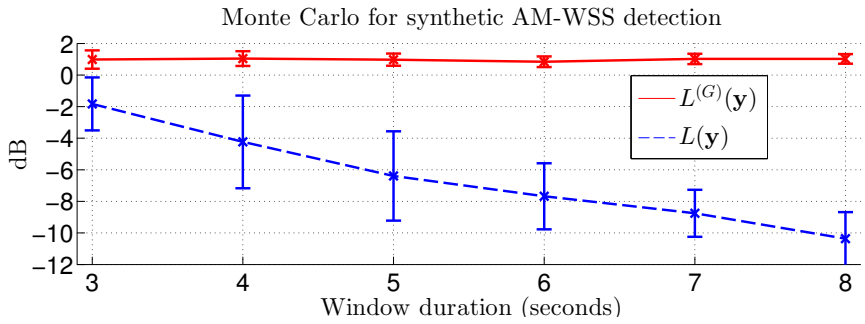


Figure 4.1: Monte Carlo experiment with synthetic data demonstrating increased coherence time using $L^{(G)}(\mathbf{y})$, the log-GLRT for our proposed model, versus $L(\mathbf{y})$, the log-GLRT for the conventional model.

### *4.5.1 Synthetic data*

A Monte Carlo analysis using synthetic data is performed. Realistic parameters that correspond to real-world passive sonar signals are used. The WSS noise carrier $w(t)$ is unit-variance white noise. The modulator $m(t)$ has 4 harmonics, center frequency $f_c = 2$ Hz, and we consider four possible chirp rates spaced equally from 0.05 to 0.2 Hz/s. For each chirp rate, we will consider a range of analysis window durations $T \in [3, 8]$ seconds. The sampling rate is $f_s = 4$ kHz.

For each $T$, $L(\mathbf{y})$ and $L^{(G)}(\mathbf{y})$ are computed and averaged over 400 trials, 100 trials for each $\beta$. Figure 4.1 shows the results. Notice that $L^{(G)}(\mathbf{y})$ remains approximately the same for all $\beta$ for all analysis durations $T$, while $L(\mathbf{y})$ decreases as $T$ increases. This happens because $L^{(G)}(\mathbf{y})$ accounts for linear frequency variations of $m(t)$, which means $L^{(G)}(\mathbf{y})$ is coherent with the signal over a longer duration. Said another way, $L^{(G)}(\mathbf{y})$ adds up the concentrated, higher-amplitude peaks produced by the generalized DEMON spectrum, while $L(\mathbf{y})$ adds up the smeared, lower-amplitude peaks of the conventional DEMON spectrum.

### *4.5.2 Real-world sonar data*

For a real-world example, we consider propeller noise that exhibits time-varying frequency content in its modulator. The example is a 12 second recording $\mathbf{z}$ of a Zodiac boat starting

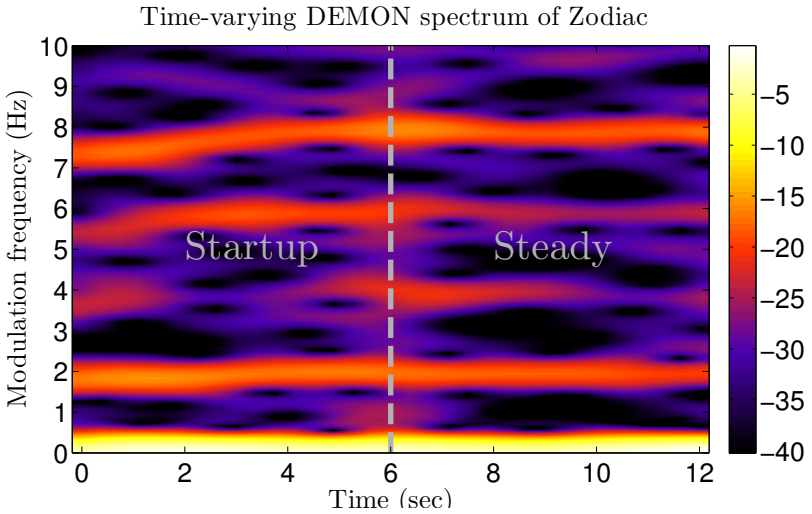


Figure 4.2: Time-varying DEMON spectrum of Zodiac boat data. The startup and steady portions are labeled. Notice the increasing modulation frequency in the startup portion.

up its engine and moving away[2]. The time-varying DEMON spectrum given by (4.1) of this signal is shown in figure 4.2. The first 6 seconds are labeled as "startup", during which the propeller rate is increasing as the boat starts its engine and moves away. During the last 6 seconds, labeled "steady", the Zodiac boat is underway and the propeller rate is relatively constant. We choose this example because the startup portion illustrates a real-world case when $f_0(t)$ is varying over the analysis window, while the steady portion illustrates the conventional assumption that $f_0(t)$ is constant over the analysis window.

Figure 4.3 shows a comparison between $L(\mathbf{z}_d)$ and $L^{(G)}(\mathbf{z}_d)$ for varying analysis window durations $T$. The expression $\mathbf{z}_d$ is the $d$th frame of duration $T$ of $\mathbf{z}$. The analysis windows are not overlapping. Notice that for all $T$ and $d$, and for both startup and steady, $L^{(G)}(\mathbf{z}_d)$ consistently achieves a higher value than $L(\mathbf{z}_d)$. During startup, $L^{(G)}(\mathbf{z}_d)$ is especially greater (by over 6dB) using a long analysis window ($T = 6$s), because the propeller rate is far from constant.

[2]Thanks to Brad Hanson and the NOAA Northwest Fisheries Science Center Marine Mammal Program for providing this data.

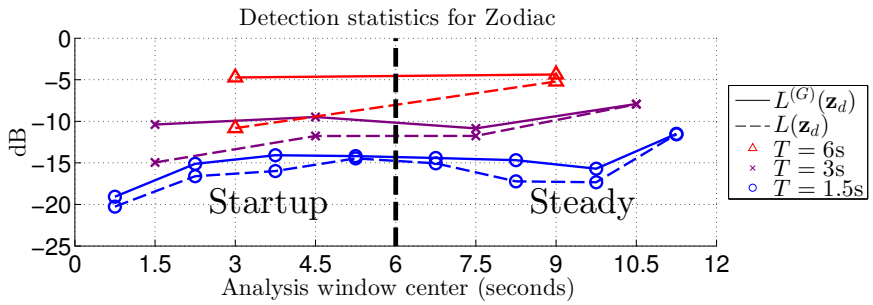


Figure 4.3: Detection statistics for real-world Zodiac boat data using different analysis window lengths. Analysis windows are non-overlapping. Notice that the statistic using the proposed generalized DEMON spectrum, $L^{(G)}(\mathbf{z}_d)$, is consistently higher than the conventional statistic $L(\mathbf{z}_d)$, especially as the analysis window duration $T$ increases.

### *4.6 Summary and Future Work*

In this chapter, we have examined a particular class of nonstationary signals, AM-WSS, and how its spectral statistics relate to the DEMON spectrum. Conventional approaches assume that the modulation frequency content is stationary within the window. However, this assumption often does not hold. By accounting for the variation, longer analysis windows can be used, providing higher SNR estimators.

We presented a method to extend the coherence time of such signals using the generalized DEMON spectrum. We validated this method using a linear model for modulation frequency trajectories to perform detection on synthetic data in a Monte Carlo experiment as well as on real-world passive sonar data. On real-world passive sonar data, our proposed method achieved up to a 6dB gain in the detection statistic over the conventional method.

Future work will investigate other models for time-varying modulation frequency (such as quadratic and higher-order polynomials), as well as incorporating our method into other statistical estimators used in beamforming, enhancement, and classification.

Chapter 5

# IMPROVED DEREVERBERATION AND NOISE REDUCTION FOR SPEECH[1]

## *5.1 Introduction*

The enhancement of speech signals in the presence of reverberation and noise remains a challenging problem with many applications. Many methods are prone to generating artifacts in the enhanced speech, and must trade off noise reduction against speech distortion.

In this chapter, we describe a new enhancement algorithm that suppresses both reverberation and background noise. We combine a statistically optimal single-channel enhancement algorithm that suppresses background noise and reverberation with an adaptive time-frequency transform domain that is coherent with speech signals over longer durations than the short-time Fourier transform (STFT). Thus, we are able to use longer analysis windows while still satisfying the assumptions of the optimal single-channel enhancement filter. Multichannel processing is made possible using a classic minimum variance distortionless response (MVDR) beamformer or, in the case of two-channel data, a delay-and-sum beamformer (DSB) preceding the single-channel enhancement.

First, we review the speech enhancement and dereverberation problem, as well as the enhancement algorithm we use proposed by Habets [55], which suppresses both noise and late reverberation based on a statistical model of reverberation. Then, we describe the fan-chirp transform, proposed by Weruaga and Kèpesi [45], [56] and improved upon by Cancela et al. [44], which provides an enhancement domain, the short-time fan-chirp transform (STFChT), that better matches time-varying frequency content of voiced speech. We discuss why performing the enhancement in the STFChT domain gives superior results compared to the STFT domain. Finally, we present our results on the REVERB challenge dataset [54],

[1]The contents of this chapter also appear in a paper by Wisdom et al. [53] for the 2014 REVERB challenge workshop [54].

which shows that our new method achieves superior results versus conventional STFT-based processing in terms of objective measures.

Our basic multichannel architecture of single-channel enhancement preceded by beamforming is not unprecedented. Gannot and Cohen [57] used a similar architecture for noise reduction that consists of a generalized sidelobe cancellation (GSC) beamformer followed by a single-channel post-filter. Maas et al. [58] employed a similar single-channel enhancement algorithm for reverberation suppression and observed promising speech recognition performance in even highly reverberant environments.

There have been several dereverberation and enhancement approaches that estimate and leverage the time-varying fundamental frequency $f_0$ of speech. Nakatani et al. [59] proposed a dereverberation method using inverse filtering that exploits the harmonicity of speech to build an adaptive comb filter. Kawahara et al. [60] used adaptive spectral analysis and estimates of $f_0$ to perform manipulation of speech characteristics.

Droppo and Acero [61] observed how the fundamental frequency of speech can change within an analysis window, and proposed a new framework that could better predict the energy of voiced speech. Dunn and Quatieri [62] used the fan-chirp transform for sinusoidal analysis and synthesis of speech, and Dunn et al. [63] also examined the effect of various interpolation methods on reconstruction error. Pantazis et al. [64] proposed an analysis/synthesis domain that uses estimates of instantaneous frequency to decompose speech into quasi-harmonic AM-FM components. Degottex and Stylianou [65] proposed another analysis/synthesis scheme for speech using an adaptive harmonic model that they claim is more flexible than the fan-chirp, as it allows nonlinear frequency trajectories.

To our knowledge, single-channel enhancement has not been attempted in these new related transform domains. Here, we demonstrate improved performance using the STFChT instead of the STFT.

### *5.2 Optimal Single-Channel Suppression of Noise and Late Reverberation*

In this section, we review the speech enhancement problem and a popular statistical speech enhancement algorithm, the minimum mean-square error log-spectral amplitude (MMSE-LSA) estimator, which was originally proposed by Ephraim and Malah [66], [67] and later

improved by Cohen [68]. We review the application of MMSE-LSA to both noise reduction and joint dereverberation and noise reduction (the latter of which was proposed by Habets [55]).

### *5.2.1 Noise reduction using MMSE-LSA*

A classic speech enhancement algorithm is the minimum mean-square error (MMSE) short-time spectral amplitude estimator proposed by Ephraim and Malah [66]. They later refined the estimator to minimize the MSE of the log-spectra [67]. We will refer to this algorithm as LSA (log-spectral amplitude). Minimizing the MSE of the log-spectra was found to provide better enhanced output because log-spectra are more perceptually meaningful. Cohen [68] suggested improvements to Ephraim and Malah's algorithm, which he referred to as "optimal modified log-spectral amplitude" (OM-LSA).

Given samples of a noisy speech signal

$$y[n] = s[n] + v[n], \tag{5.1}$$

where $s[n]$ is the clean speech signal and $v[n]$ is additive noise, the goal of an enhancement algorithm is to estimate $s[n]$ from the noisy observations $y[n]$. The LSA estimator yields an estimate $\hat{A}(d,k)$ of the clean STFT magnitudes $|S(d,k)|$ (where $S(d,k)$ are assumed to be normally distributed) by applying a frequency-dependent gain $G_{\mathrm{LSA}}(d,k)$ to the noisy STFT magnitudes $|Y(d,k)|$:

$$\hat{A}(d,k) = G_{\mathrm{LSA}}(d,k)|Y(d,k)|. \tag{5.2}$$

Given these estimated magnitudes, the enhanced speech is reconstructed from STFT coefficients combining $\hat{A}(d,k)$ with noisy phase: $\hat{S}(d,k) = \hat{A}(d,k)e^{j\angle Y(d,k)}$. The LSA gains are computed as [67, (20)]:

$$G_{\mathrm{LSA}}(d,k) = \frac{\xi(d,k)}{1+\xi(d,k)} \exp\left\{\frac{1}{2}\int_{v(d,k)}^{\infty} \frac{e^{-t}}{t} dt\right\}, \tag{5.3}$$

where $\xi(d,k)$ is the *a priori* signal-to-noise ratio (SNR) for the $k$th frequency bin of the $d$th frame, and is defined to be $\xi(d,k) \triangleq \frac{\lambda_s(d,k)}{\lambda_v(d,k)}$, where $\lambda_s(d,k) = E\left\{|S(d,k)|^2\right\}$ is the variance of $S(d,k)$ and $\lambda_v(d,k) = E\left\{|V(d,k)|^2\right\}$ is the variance of $V(d,k)$. The variable

$v(d,k) = \frac{\xi(d,k)}{1+\xi(d,k)}\gamma(d,k)$, where $\gamma(d,k)$ is the *a posteriori* SNR for the $k$th frequency bin of the $d$th frame, defined as $\gamma(d,k) \triangleq \frac{|Y(d,k)|^2}{\lambda_v(d,k)}$.

Cohen [68] refined Ephraim and Malah's approach to include a lower bound $G_{\min}$ for the gains as well as an *a priori* speech presence probability (SPP) estimator $p(d,k)$. Cohen's estimator is as follows [68, (8)]:

$$G_{\text{OM-LSA}} = \{G_{\text{LSA}}(d,k)\}^{p(d,k)} \cdot G_{\min}^{1-p(d,k)}. \tag{5.4}$$

Cohen also derived an efficient estimator for the SPP $p(d,k)$ [68] that exploits the strong interframe and interfrequency correlation of speech in the STFT domain.

### *5.2.2 Joint dereverberation and noise reduction*

Habets [55] proposed a MMSE-LSA enhancement algorithm that uses a statistical model of reverberation to suppress both noise and late reverberation. The signal model he uses is

$$y[n] = s[n] * h[n] + v[n] = x_e[n] + x_\ell[n] + v[n], \tag{5.5}$$

where $s[n]$ is the clean speech signal, $h[n]$ is the room impulse response (RIR), and $v[n]$ is additive noise. The terms $x_e[n]$ and $x_\ell[n]$ correspond to the early and late reverberated speech signals, respectively. The partition between early and late reverberations is determined by a parameter $n_e$, which is a discrete sample index. All samples in the RIR before $n_e$ are taken to cause early reflections, and all samples after $n_e$ are taken to cause late reflections [55]. Thus,

$$h[n] = \begin{cases} 0, & \text{if } n < 0 \\ h_e[n], & \text{if } 0 \le n < n_e \\ h_\ell[n] & \text{if } n_e \le n. \end{cases} \tag{5.6}$$

Using these definitions, $x_e[n] = s[n] * h_e[n]$ and $x_\ell[n] = s[n] * h_\ell[n]$.

Habets proposed a generalized statistical model of reverberation that is valid both when the source-microphone distance is less than or greater than the critical distance [55]. This model divides the RIR $h[n]$ into a direct-path component $h_d[n]$ and reverberant component $h_r[n]$. Both direct-path and reverberant components are taken to be white, zero-mean,

stationary Gaussian noise sequences $b_d[n]$ and $b_r[n]$ with variances $\sigma_d^2$ and $\sigma_r^2$ enveloped by an exponential decay,

$$h_d[n] = b_d[n]e^{-\bar{\zeta}n} \text{ and } h_r[n] = b_r[n]e^{-\bar{\zeta}n}, \tag{5.7}$$

where $\bar{\zeta}$ is related to the reverberation time $T_{60}$ by [55]:

$$\bar{\zeta} = \frac{3\ln(10)}{T_{60}f_s}. \tag{5.8}$$

Using this model, the expected value of the energy envelope of $h[n]$ is

$$E\left[h^2[n]\right] = \begin{cases} \sigma_d^2 e^{-2\bar{\zeta}n}, & \text{for } 0 \leq n < n_d \\ \sigma_r^2 e^{-2\bar{\zeta}n}, & \text{for } n \geq n_d \\ 0 & \text{otherwise.} \end{cases} \tag{5.9}$$

Under the assumptions that the speech signal is stationary over short analysis windows (i.e., duration much less than $T_{60}$), Habets proposed [55, (3.87)] the following model of the spectral variance of the reverberant component $x_r[n]$:

$$\begin{aligned} \lambda_{x_r}(d,k) =& e^{-2\bar{\zeta}(k)R}\lambda_{x_r}(d-1,k)... \\ &+ \frac{E_r}{E_d}\left(1 - e^{-2\bar{\zeta}(k)R}\right)\lambda_{x_d}(d-1,k), \end{aligned} \tag{5.10}$$

where $R$ is the number of samples separating two adjacent analysis frames and $E_r/E_d$ is the inverse of the direct-to-reverberant ratio (DRR). Thus, the spectral variance of the reverberant component in the current frame $d$ is composed of scaled copies of the spectral variance of the reverberation and the spectral variance of the direct-path signal from the previous frame $d-1$.

Using this model, the variance of the late reverberant component can be expressed as [55, (3.85)]:

$$\lambda_{x_\ell}(d,k) = e^{-2\bar{\zeta}(k)(n_e - R)}\lambda_{x_r}\left(d - \frac{n_e}{R} + 1, k\right), \tag{5.11}$$

which is quite useful in practice, because the variance of the late-reverberant component can be computed from the variance of the total reverberant component.

To suppress both noise and late reverberation, the *a priori* and *a posteriori* SNRs $\xi(d,k)$ and $\gamma(d,k)$ from the previous section become *a priori* and *a posteriori* signal-to-interference ratios (SIRs), given by [55, (3.25), (3.26)]:

$$\xi(d,k) = \frac{\lambda_{x_e}(d,k)}{\lambda_{x_\ell}(d,k) + \lambda_v(d,k)} \tag{5.12}$$

and

$$\gamma(d,k) = \frac{|Y(d,k)|^2}{\lambda_{x_\ell}(d,k) + \lambda_v(d,k)}. \tag{5.13}$$

The gains are computed by plugging the SIRs in (5.12) and (5.13) into (5.3) and (5.4). Habets suggested an additional change to (5.4), which makes $G_{\min}$ time- and frequency-dependent. This is done because the interference of both noise and late reverberation is time-varying. The modification is [55, (3.29)]

$$G_{\min}(d,k) = \frac{G_{\min,x_\ell}\hat{\lambda}_{x_\ell}(d,k) + G_{\min,v}\hat{\lambda}_v(d,k)}{\hat{\lambda}_{x_\ell}(d,k) + \hat{\lambda}_v(d,k)}. \tag{5.14}$$

Notice that two parameters in (5.8) and (5.10) are not known *a priori*; namely, $T_{60}$ and the DRR. These parameters must be blindly estimated from the data. For $T_{60}$ estimation, Löllmann et al. [69] propose an algorithm, which we found to be effective. As for the DRR, Habets suggested an online adaptive procedure [55, §3.7.2].

### *5.3 MMSE-LSA in the Fan-Chirp Domain*

In this section we analyze performing joint dereverberation and noise reduction using MMSE-LSA in the STFChT domain and provide an example for why the STFChT (3.3), which is a domain that is more coherent with speech signals, is a more appropriate enhancement domain than the STFT.

The MMSE-LSA framework for joint dereverberation and noise reduction implicitly assumes that the the frequency content of speech does not change very much over the analysis duration. Such an assumption relies on the local stationarity of speech signals within the analysis frame. For voiced speech, this is essentially equivalent to the fundamental frequency $f_0$ being constant over the analysis frame, and the frequency variation of voiced speech limits analysis durations to 10-30ms.

Using shorter analysis frames means only a finite amount of approximately stationary data is available at any specific time, and this finite amount of data limits the performance of statistical estimators. To improve this situation, we propose to increase the analysis duration by changing the time base of the analysis such that there is less frequency variation within the frame, which makes the data more stationary. This time base modification is performed using the fan-chirp, which uses a linear model of frequency variation within the frame.

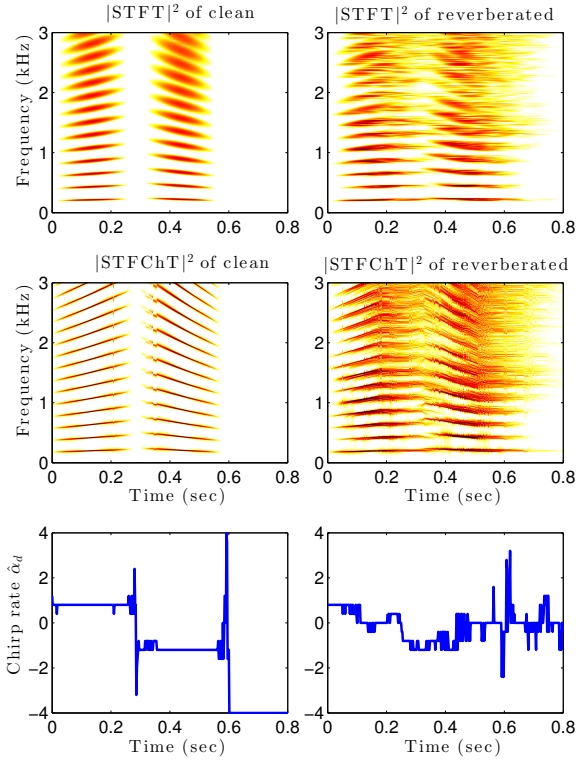


Figure 5.1: Simple example showing benefits of fan-chirp both for narrower harmonics across frequency and for better coherence with direct path signal in the presence of reverberation. Test signal is two consecutive synthetic speech-like harmonic stacks. All colormaps are identical with a dynamic range of 40dB. Left plots are the representations of the clean signal and right plots are the representations of the reverberated signal. The chosen analysis chirp rates $\hat{\alpha}$ are shown in the bottom plots.

To give intuition about the benefits of the fan-chirp in the presence of reverberation, we present a simple example in figure 5.1. Consider two successive Gaussian-enveloped harmonic chirps with duration 200 ms and spaced 100 ms apart. Let the $f_0$ of the first harmonic chirp start at 200 Hz and rise to 233 Hz, and let the $f_0$ of the second harmonic chirp have a range from 250 Hz falling to 200 Hz. Both chirps have 20 harmonics. This sequence of harmonic chirps has parameters that are typical of two successive voiced vowels (here we do not consider the spectral shape imposed by a vocal tract filter, for simplicity). Now, let us apply reverberation to this signal (we use the first channel of the `MediumRoom2_far_AnglA` RIR provided in the REVERB challenge development set [54]), and examine the clean and reverberated versions in both the STFT and STFChT domains. The result is shown in figure 5.1. STFT and STFChT parameters are exactly matched, with a sampling rate of $f_s = 16$kHz, a Hamming window of duration 2048 samples, a frame hop of 128 samples, and a 3262-length FFT.

Notice that at higher frequencies in the STFT of the clean signal (top left panel of figure 5.1), the harmonics become broader and more smeared across frequency. In contrast, the STFChT of the clean signal (center left panel of figure 5.1) exhibits narrow lines at all frequencies. When reverberation is applied, the STFT of the reverberated signal (top right panel of figure 5.1) exhibits smears that become wider with increasing frequency, and adjacent harmonics even become smeared together. In contrast, in the STFChT of the reverberated signal (center right panel of figure 5.1), the direct path signal shows up as narrow lines at all frequencies, while some smearing results in frames that contain reverberant energy. One cause of the smearing during reverberation-dominated frames seems to be errors in the estimation of $\hat{\alpha}_d$, the analysis chirp rate, caused by the additional reverberant components, which are shown in the bottom panels of figure 5.1. Despite these estimation errors, the STFChT still seems to give a better representation of the signal compared to the STFT, because the STFChT reduces smearing of higher-frequency components and achieves better coherence with the direct-path signal (i.e., direct-path signals show up as more narrow lines).

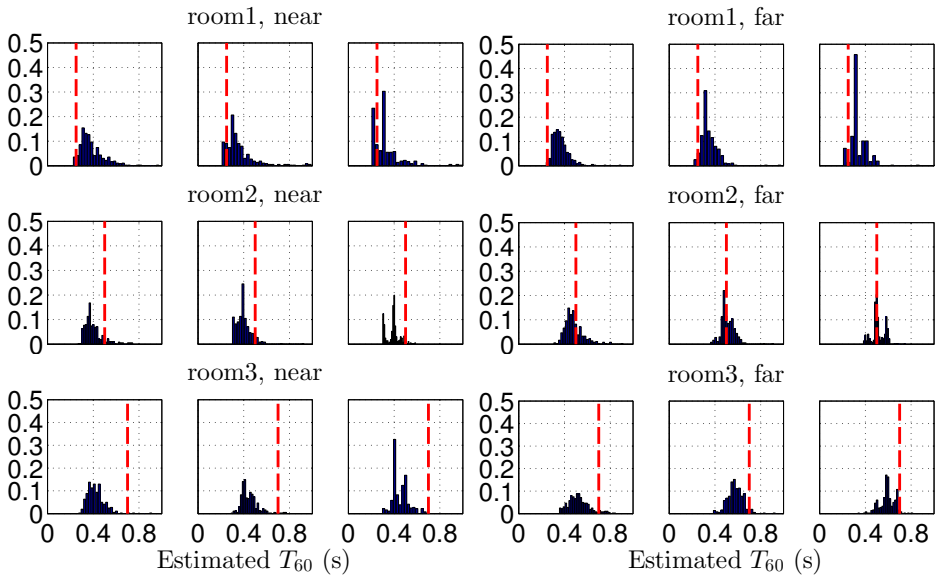


Figure 5.2: Histograms of estimated $T_{60}$ time measured on SimData evaluation dataset (these results were not used to tune the algorithm). For each condition, left plot is for 1ch data, center plot is for 2ch data, and right plot is for 8ch data. These plots show that $T_{60}$ estimation [69] precision generally improved with increasing amounts of data (i.e., with more channels), although for some conditions $T_{60}$ estimates were inaccurate. Dotted lines indicate approximate $T_{60}$ times given by REVERB organizers [54].

### *5.4 Implementation*

Our algorithms are implemented in MATLAB, and we use utterance-based processing. The algorithm starts by using the utterance data to estimate the $T_{60}$ time of the room using the blind algorithm proposed by Löllmann et al. [69]. Multichannel utterance input data is concatenated into a long vector, and as recommended by Löllmann et al., noise reduction is performed beforehand. We use Loizou's implementation [70] of Ephraim and Malah's LSA [67] for this pre-enhancement. Figure 5.2 shows histograms of the $T_{60}$ estimation performance using this approach.

For multichannel data, we estimate the direction of arrival (DOA) by cross-correlating oversampled data between channels. That is, we compute a $N_{ch}$-length vector of time delays

$\mathbf{d}$ with $d_1 = 0$ and $d_i$, $i$=2,...,$N_{ch}$ given by

$$d_i = \underset{k}{\text{argmax}} \frac{r_{1i}[k]}{U f_s}, \tag{5.15}$$

where $r_{1i}[k] = \sum_n x_1[n] x_i[n-k]$, $U$ is the oversampling factor, and $c = 340$ meters per second, the approximate speed of sound in air.

Given a time delay vector $\mathbf{d}$, the DOA estimate is given by the solution to $\mathbf{P}\hat{\mathbf{a}} = \frac{1}{c}\mathbf{d}$, where $\hat{\mathbf{a}}$ is a $3 \times 1$ unit vector representing the estimated DOA of the speech signal and $\mathbf{P}$ is a $N_{ch} \times 3$ matrix containing the Cartesian $(x, y, z)$ coordinates of the array elements. For example, for an eight-element uniform circular array, $P_{i1} = x_i = r\cos(i\pi/4)$, $P_{i2} = y_i = r\sin(i\pi/4)$, and $P_{i3} = z_i = 0$ for $i = 0, 1, ..., 7$, where $r$ is the array radius.

For the 8-channel case, the estimated DOA is used to form the steering vector $\mathbf{v}^H(f)$ for a frequency-domain MVDR beamformer applied to the multichannel signal. The weights $\mathbf{w}^H(d, f)$ for the MVDR are [71, (6.14-15)]

$$\mathbf{w}^H(d, f) = \frac{\mathbf{v}^H(f)\mathbf{S}_{\mathbf{yy}}^{-1}(d, f)}{\mathbf{v}^H(d, f)\mathbf{S}_{\mathbf{yy}}^{-1}(d, f)\mathbf{v}(d, f)}, \tag{5.16}$$

where $\mathbf{S}_{\mathbf{yy}}(d, f)$ is the spatial covariance matrix at frequency $f$ and frame $d$ estimated using $N$ snapshots $Y(d-n, f)$ for $-N/2 \leq n < N/2$ and $\mathbf{v}$ is given by

$$\mathbf{v}(f) = \exp\left(j\frac{2\pi f}{c}\mathbf{P}\hat{\mathbf{a}}\right). \tag{5.17}$$

The MVDR uses a 512-sample long Hamming window with 25% overlap, a 512-point FFT, and $N = 24$ snapshots for spatial covariance estimates. For 2-channel data, we use a delay-and-sum beamformer to enhance the signal with the delay given by the DOA estimate. Single-channel data is enhanced directly by the single-channel MMSE-LSA algorithm. A block diagram of these three cases is shown in figure 5.3.

We tried three analysis/synthesis domains for the MMSE-LSA enhancement algorithm: the STFT with a short window, the STFT with a long window, and the STFChT. The STFT with a short window uses 512-sample long ($T = 32$ms) Hamming windows, a frame hop of 128 samples, and an FFT length of 512. Short-window STFT processing is chosen to match conventional speech processing window lengths. The STFT with a long window uses 2048-sample long ($T = 128$ms) Hamming windows, a frame hop of 128 samples, and an

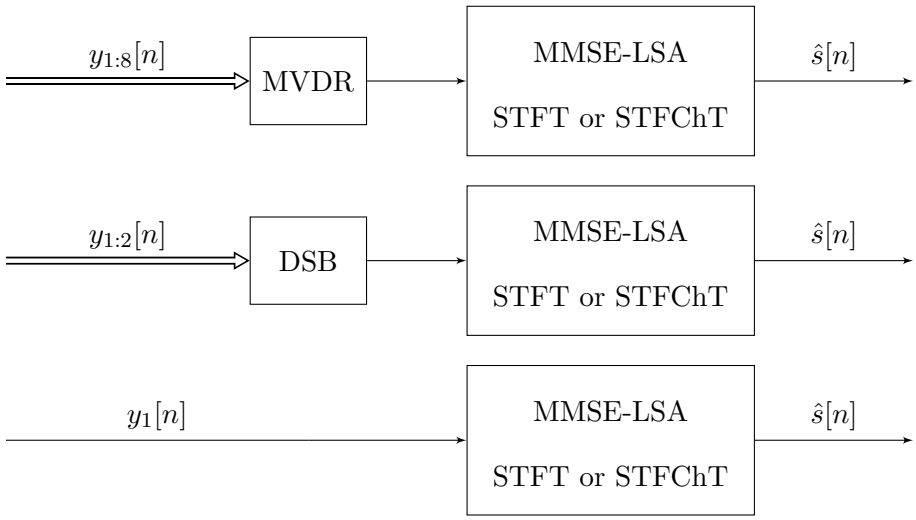


Figure 5.3: Block diagrams of processing for 8ch data using a minimum variance distortionless response (MVDR) beamformer (top), 2ch data using a delay-and-sum beamformer (DSB, middle), and 1ch data (bottom).

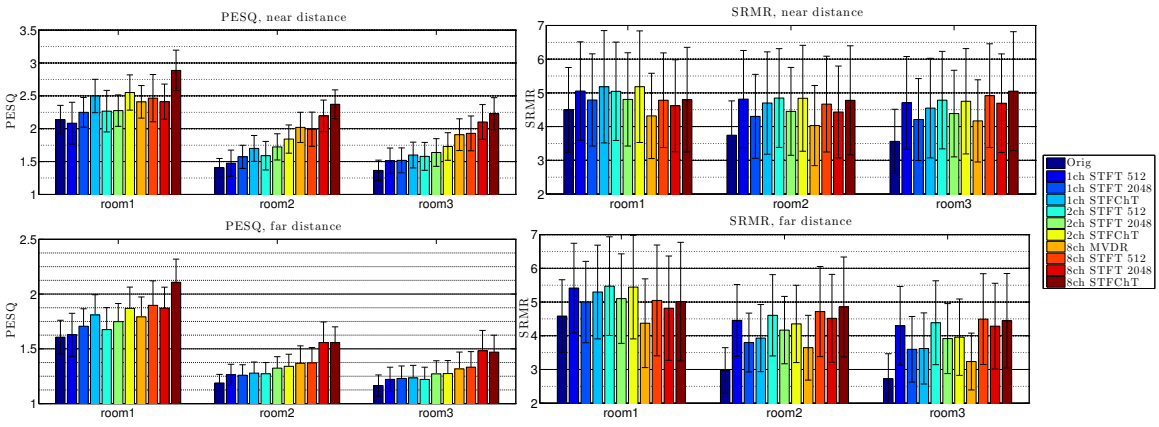


Figure 5.4: PESQ and SRMR results for SimData evaluation set. Upper plots are near distance condition, lower plots are far distance condition. Left plots are PESQ, right plots are SRMR.

FFT length of 3262. Long-window STFT processing is intended to match the parameters of STFChT processing for a direct comparison. STFChT processing uses an analysis duration of 2048 samples, a Hamming analysis window, a frame hop of 128 samples, an FFT length of 3262, oversampling factor of 8, and a set of possible analysis chirp rates $\mathcal{A}$ consisting of 21 equally spaced $\alpha$s from -4 to 4.

The forward STFChT, given by (3.3), proceeds frame-by-frame, estimating the optimal analysis chirp rate $\hat{\alpha}_d$ using (3.7), oversampling in time, warping, applying an analysis window, and taking the FFT. Then MMSE-LSA weights are estimated frame-by-frame and applied in the STFChT domain, and the enhanced speech signal is reconstructed using the inverse STFChT.

For all methods, noise estimation is performed with a decision-directed method and simple online updating of the noise variance. Voice activity detection to determine if a frame is noise-only or speech-plus-noise is done using Loizou's method, which compares the following quantity to a threshold $\eta_{\text{thresh}}$:

$$\eta(d) = \sum_k \ln \gamma(d,k) \frac{\xi(d,k)}{1+\xi(d,k)} - \ln(1+\xi(d,k)). \tag{5.18}$$

If $\eta(d) < \eta_{\text{thresh}}$, the frame is determined to be noise-only and the noise variance is updated as $\lambda_v(d,k) = \mu_v \lambda_v(d-1,k) + (1-\mu_v)|Y(d,k)|^2$, with $\mu_v = 0.98$ and $\eta_{\text{thresh}} = 0.15$.

For our implementation of Habets's joint dereverberation and noise reduction algorithm, we used Loizou's implementation [70] of Ephraim and Malah's LSA `logmmse` MATLAB implementation as a foundation. The forward STFChT code was written by Cancela et al. [44]. We wrote our own MATLAB implementation of the inverse STFChT.

Computation times for processing REVERB evaluation data are shown in figure 5.8. We measured reference wall clock times of 265.43s and 39.62s, respectively, for SimData and RealData. For 8-channel data, the MVDR and the STFChT require the most computation. For 1-channel and 2-channel data, the STFChT requires the most computation. For the STFChT, much of the computation is used to compute the GLogS for estimation of the analysis chirp rate $\hat{\alpha}_d$ (3.7) for each frame. Note that this computation could be easily parallelized in hardware.

### *5.5 Results*

We evaluate our algorithms on the REVERB challenge evaluation dataset [54]. The data consists of both simulated and real reverberated speech. Simulated data (SimData) are created by convolving utterances from the Wall Street Journal Cambridge read news (WSJ-CAM0) corpus [72] measured room impulse responses for three different reverberant rooms and at two distances: a near distance of about 0.5 meters and a far distance of about 2 meters. Recorded air conditioning noise is added at about 20dB signal-to-noise ratio (SNR). Real data (RealData) are actual recordings of male and female speakers from the multichannel Wall Street Journal audio-visual (MC-WSJ-AV) corpus [73] reading prompts in a noisy (air conditioning noise at about 20dB SNR) and reverberant room, at two distances: a near distance of 1 meter and a far distance of 2.5 meters.

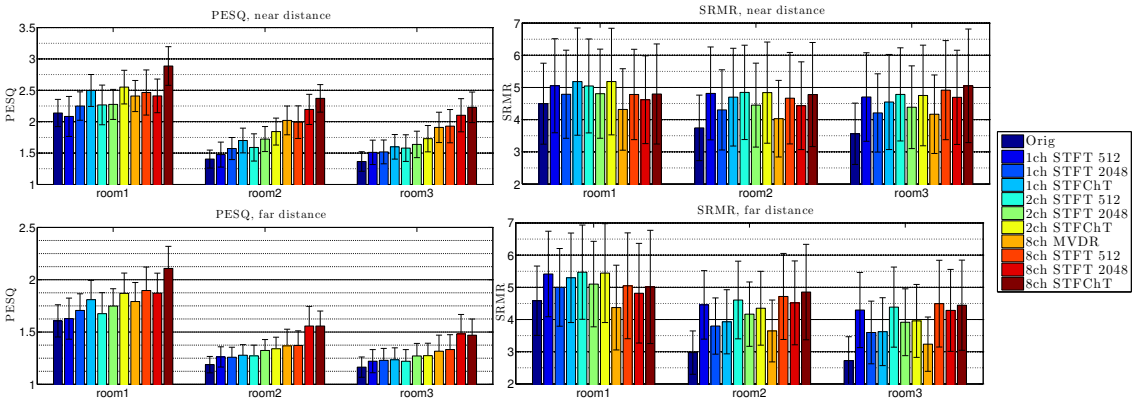


Figure 5.5: PESQ and SRMR results for SimData evaluation set. Upper plots are near distance condition, lower plots are far distance condition. Left plots are PESQ, right plots are SRMR.

Our results on REVERB evaluation data are shown in figures 5.5, 5.7, and 5.8. For the challenge, we submitted results using STFChT-based processing. We choose to display PESQ (Perceptual Evaluation of Speech Quality) [74] and SRMR (source-to-reverberation modulation energy ratio) [75] more prominently because the former is the ITU-T standard for voice quality testing [76] and the latter is both a measure of dereverberation and the

only non-intrusive measure that can be run on RealData (for which the clean speech is not available).

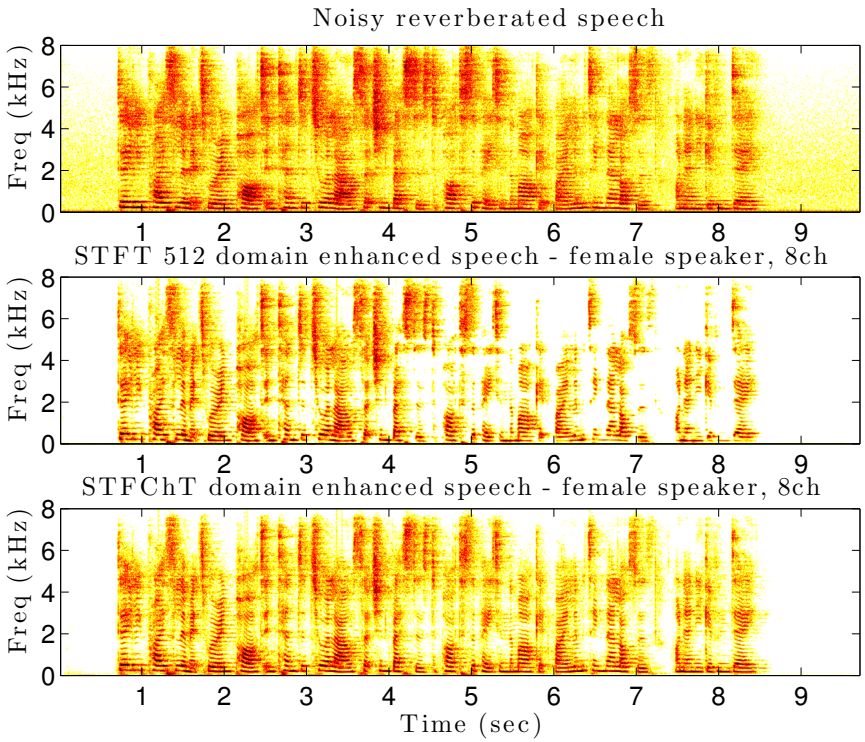


Figure 5.6: Spectrogram comparisons for one 8ch far-distance utterance, `c3bc020q`, from SimData evaluation set.

For SimData, STFChT-based enhancement always performs better in terms of PESQ than STFT-based enhancement using either a short (512-sample) window or a long (2048-sample) window, for the 8-, 2-, and 1-channel cases (except for 8-channel, far-distance data in room 3). Informal listening tests revealed an oversuppression of speech and some musical noise artifacts in STFT processing, while STFChT processing did not exhibit oversuppression or musical noise artifacts. The oversuppression of direct-path speech by STFT processing can be seen in the spectrogram comparisons shown in figure 5.6. In terms of SRMR, STFChT processing yields equivalent or slightly worse SRMR scores than long-window STFT processing for the 8-, 2-, and 1-channel cases (except for 8-channel, near-distance

data, where STFChT processing does slightly better). One issue with these SRMR comparisons, however, is that the variance of the SRMR scores is quite high. Thus, for SimData, STFChT processing achieves better perceptual audio quality while still achieving almost equivalent dereverberation compared to STFT processing.

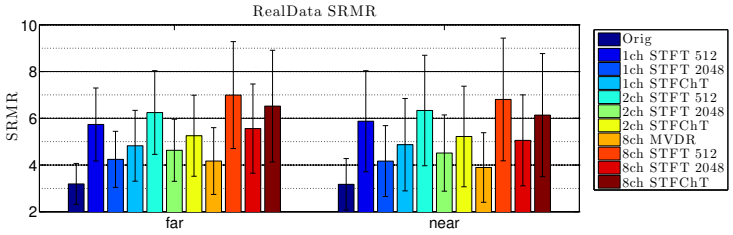


Figure 5.7: SRMR results for RealData evaluation set.

For RealData, we achieved SRMR improvements of over 3, as shown in figure 5.7. Short-window STFT processing achieved higher scores than STFChT processing (especially for 1- and 2-channel data), but informal listening tests revealed an oversuppression of speech and some musical noise artifacts in STFT processing, while little oversuppression and fewer artifacts were perceived in STFChT processing. Informal listening tests also indicated that the STFChT processing suppresses reverberation slightly less as compared to STFT processing, which concurs with lower SRMR scores for STFChT processing. Thus, though STFT processing achieves better dereverberation on RealData, better dereverberation performance seems to come at the cost of oversuppression of direct-path speech and addition of artifacts. STFChT processing, on the other hand, achieves slightly less dereverberation on RealData, but the enhanced speech does not seem to suffer from oversuppression or artifacts.

### *5.6 Summary and Future Work*

In this chapter we combined an optimal MMSE log-spectral amplitude estimator for joint dereverberation and noise reduction with a recently-developed adaptive time-frequency transform that is coherent with speech signals over longer durations. Our approach yielded improved results on the REVERB challenge dataset versus standard STFT processing. Processing in the STFChT domain resulted in less reverberation at the output without in-

**SimData summary**

| Ch. | Method | Comp. Time (s) | Mean CD | Median CD | SRMR | Mean LLR | Median LLR | Mean FWSegSNR | Median FWSegSNR | PESQ |
|---|---|---|---|---|---|---|---|---|---|---|
| | *Orig* | — | *3.97* | *3.68* | *3.68* | *0.57* | *0.51* | *3.62* | *5.39* | *1.48* |
| 8 | STFT 512/2048 | 7447.86 / 7622.38 | 3.56 / 3.18 | 3.23 / 2.83 | 4.77 / 4.56 | 0.61 / 0.43 | 0.50 / 0.38 | 8.06 / 6.79 | 8.47 / 9.31 | 1.83 / 1.94 |
| 8 | STFChT | 17132.60 | 2.97 | 2.49 | 4.82 | 0.43 | 0.37 | 9.21 | 10.63 | 2.10 |
| 2 | STFT 512/2048 | 1955.89 / 2022.04 | 3.80 / 3.57 | 3.42 / 3.22 | 4.86 / 4.47 | 0.65 / 0.49 | 0.55 / 0.44 | 7.26 / 5.46 | 7.93 / 7.86 | 1.60 / 1.66 |
| 2 | STFChT | 8248.49 | 3.33 | 2.83 | 4.75 | 0.51 | 0.45 | 7.68 | 9.19 | 1.77 |
| 1 | STFT 512/2048 | 1003.54 / 1074.93 | 3.87 / 3.84 | 3.48 / 3.51 | 4.79 / 4.28 | 0.68 / 0.54 | 0.58 / 0.47 | 6.72 / 4.65 | 7.62 / 6.71 | 1.53 / 1.59 |
| 1 | STFChT | 7454.59 | 3.57 | 3.07 | 4.55 | 0.57 | 0.49 | 7.07 | 8.60 | 1.69 |

SimData, far distance, room 1

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *Orig* | — | *2.67* | *2.38* | *4.58* | *0.38* | *0.35* | *6.68* | *9.24* | *1.61* |
| 8 | STFT 512/2048 | 7337.43 / 7477.34 | 3.16 / 2.28 | 2.88 / 2.05 | 5.05 / 4.82 | 0.52 / 0.37 | 0.43 / 0.34 | 8.50 / 9.16 | 8.73 / 10.57 | 1.90 / 1.87 |
| 8 | STFChT | 16730.48 | 2.47 | 2.04 | 5.01 | 0.34 | 0.30 | 10.16 | 11.26 | 2.11 |
| 2 | STFT 512/2048 | 1872.86 / 1936.2 | 3.32 / 2.41 | 2.98 / 2.14 | 5.47 / 5.10 | 0.50 / 0.38 | 0.41 / 0.35 | 8.06 / 8.30 | 8.48 / 10.14 | 1.68 / 1.75 |
| 2 | STFChT | 7175.92 | 2.66 | 2.18 | 5.44 | 0.35 | 0.31 | 8.99 | 10.14 | 1.87 |
| 1 | STFT 512/2048 | 975.37 / 1057.11 | 3.34 / 2.60 | 3.00 / 2.32 | 5.42 / 5.00 | 0.51 / 0.37 | 0.42 / 0.34 | 8.09 / 7.82 | 8.59 / 9.97 | 1.63 / 1.71 |
| 1 | STFChT | 6378.58 | 2.83 | 2.34 | 5.30 | 0.37 | 0.32 | 8.86 | 10.12 | 1.81 |

SimData, near distance, room 1

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *Orig* | — | *1.99* | *1.68* | *4.50* | *0.35* | *0.33* | *8.12* | *10.72* | *2.14* |
| 8 | STFT 512/2048 | 7230.86 / 7500.12 | 2.92 / 1.97 | 2.71 / 1.74 | 4.78 / 4.62 | 0.46 / 0.34 | 0.38 / 0.32 | 8.72 / 9.73 | 8.83 / 10.56 | 2.47 / 2.41 |
| 8 | STFChT | 16807.41 | 2.12 | 1.68 | 4.80 | 0.28 | 0.25 | 10.83 | 11.62 | 2.89 |
| 2 | STFT 512/2048 | 1842.15 / 1904.45 | 2.90 / 1.90 | 2.64 / 1.64 | 5.05 / 4.81 | 0.44 / 0.36 | 0.37 / 0.34 | 8.83 / 9.36 | 9.08 / 10.83 | 2.27 / 2.28 |
| 2 | STFChT | 7104.62 | 2.18 | 1.72 | 5.18 | 0.30 | 0.27 | 10.23 | 11.13 | 2.55 |
| 1 | STFT 512/2048 | 954.31 / 1032.19 | 3.02 / 2.11 | 2.75 / 1.83 | 5.05 / 4.79 | 0.48 / 0.36 | 0.41 / 0.34 | 8.66 / 8.83 | 8.88 / 10.66 | 2.08 / 2.25 |
| 1 | STFChT | 6389.66 | 2.29 | 1.84 | 5.18 | 0.31 | 0.28 | 10.07 | 11.01 | 2.50 |

SimData, far distance, room 2

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *Orig* | — | *5.21* | *5.04* | *2.97* | *0.75* | *0.63* | *1.04* | *1.77* | *1.19* |
| 8 | STFT 512/2048 | 7824.78 / 7954.58 | 4.31 / 4.25 | 3.85 / 3.87 | 4.72 / 4.52 | 0.72 / 0.47 | 0.60 / 0.40 | 7.43 / 4.98 | 8.29 / 8.28 | 1.37 / 1.56 |
| 8 | STFChT | 17652.06 | 3.82 | 3.29 | 4.85 | 0.58 | 0.49 | 7.84 | 9.71 | 1.56 |
| 2 | STFT 512/2048 | 2064.93 / 2134.76 | 4.59 / 4.75 | 4.05 / 4.46 | 4.61 / 4.17 | 0.78 / 0.58 | 0.64 / 0.50 | 6.24 / 3.33 | 7.49 / 5.72 | 1.27 / 1.32 |
| 2 | STFChT | 9119.6 | 4.29 | 3.76 | 4.35 | 0.71 | 0.61 | 5.93 | 7.74 | 1.34 |
| 1 | STFT 512/2048 | 1020.71 / 1075.23 | 4.59 / 4.98 | 4.08 / 4.75 | 4.46 / 3.80 | 0.82 / 0.68 | 0.69 / 0.57 | 5.26 / 2.20 | 6.70 / 3.71 | 1.26 / 1.26 |
| 1 | STFChT | 8177.94 | 4.53 | 4.01 | 3.93 | 0.79 | 0.68 | 5.01 | 6.68 | 1.28 |

SimData, near distance, room 2

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *Orig* | — | *4.63* | *4.24* | *3.74* | *0.49* | *0.40* | *3.35* | *5.52* | *1.40* |
| 8 | STFT 512/2048 | 7545.94 / 7782.53 | 3.31 / 3.33 | 3.03 / 2.84 | 4.67 / 4.43 | 0.51 / 0.28 | 0.41 / 0.20 | 9.82 / 7.68 | 9.93 / 11.63 | 1.99 / 2.20 |
| 8 | STFChT | 18103.93 | 2.78 | 2.32 | 4.78 | 0.33 | 0.26 | 11.54 | 13.18 | 2.37 |
| 2 | STFT 512/2048 | 1944.39 / 2010.15 | 3.69 / 4.07 | 3.41 / 3.55 | 4.85 / 4.45 | 0.60 / 0.36 | 0.51 / 0.28 | 8.69 / 5.80 | 8.94 / 9.53 | 1.59 / 1.72 |
| 2 | STFChT | 9007.05 | 3.30 | 2.84 | 4.84 | 0.46 | 0.38 | 9.42 | 11.26 | 1.84 |
| 1 | STFT 512/2048 | 988.66 / 1052.96 | 3.95 / 4.48 | 3.66 / 4.00 | 4.82 / 4.30 | 0.67 / 0.44 | 0.57 / 0.35 | 7.59 / 4.55 | 8.23 / 7.57 | 1.48 / 1.57 |
| 1 | STFChT | 8267.96 | 3.64 | 3.17 | 4.70 | 0.54 | 0.45 | 8.25 | 10.08 | 1.70 |

SimData, far distance, room 3

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *Orig* | — | *4.95* | *4.72* | *2.72* | *0.83* | *0.76* | *0.24* | *0.88* | *1.16* |
| 8 | STFT 512/2048 | 7526.24 / 7576.84 | 4.29 / 4.06 | 3.83 / 3.71 | 4.49 / 4.28 | 0.80 / 0.62 | 0.68 / 0.57 | 5.94 / 3.32 | 6.99 / 5.83 | 1.33 / 1.48 |
| 8 | STFChT | 16566.95 | 3.82 | 3.27 | 4.45 | 0.63 | 0.55 | 5.96 | 7.72 | 1.47 |
| 2 | STFT 512/2048 | 2037.28 / 2106.18 | 4.56 / 4.45 | 4.03 / 4.14 | 4.39 / 3.92 | 0.85 / 0.71 | 0.74 / 0.66 | 4.68 / 2.00 | 6.10 / 4.02 | 1.22 / 1.27 |
| 2 | STFChT | 8555.67 | 4.23 | 3.67 | 3.96 | 0.73 | 0.66 | 4.33 | 6.02 | 1.27 |
| 1 | STFT 512/2048 | 1055.52 / 1117.79 | 4.49 / 4.68 | 3.96 / 4.40 | 4.30 / 3.60 | 0.86 / 0.76 | 0.75 / 0.69 | 4.19 / 1.27 | 5.84 / 2.62 | 1.22 / 1.23 |
| 1 | STFChT | 7759.75 | 4.47 | 3.92 | 3.62 | 0.79 | 0.70 | 3.74 | 5.45 | 1.23 |

SimData, near distance, room 3

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *Orig* | — | *4.37* | *4.03* | *3.56* | *0.65* | *0.58* | *2.27* | *4.20* | *1.37* |
| 8 | STFT 512/2048 | 7221.89 / 7442.87 | 3.39 / 3.18 | 3.10 / 2.78 | 4.92 / 4.69 | 0.65 / 0.47 | 0.53 / 0.42 | 7.96 / 5.85 | 8.07 / 9.02 | 1.93 / 2.10 |
| 8 | STFChT | 16934.76 | 2.79 | 2.33 | 5.05 | 0.43 | 0.36 | 8.92 | 10.28 | 2.23 |
| 2 | STFT 512/2048 | 1973.75 / 2040.5 | 3.72 / 3.82 | 3.39 / 3.41 | 4.78 / 4.39 | 0.72 / 0.56 | 0.62 / 0.49 | 7.07 / 3.98 | 7.50 / 6.90 | 1.58 / 1.64 |
| 2 | STFChT | 8528.08 | 3.32 | 2.82 | 4.75 | 0.53 | 0.46 | 7.18 | 8.83 | 1.73 |
| 1 | STFT 512/2048 | 1026.68 / 1114.32 | 3.80 / 4.18 | 3.44 / 3.79 | 4.70 / 4.21 | 0.74 / 0.61 | 0.64 / 0.53 | 6.57 / 3.21 | 7.51 / 5.76 | 1.51 / 1.52 |
| 1 | STFChT | 7753.63 | 3.64 | 3.12 | 4.55 | 0.60 | 0.52 | 6.49 | 8.24 | 1.60 |

**RealData summary**

| Ch. | Method | Comp. Time (s) | SRMR |
|---|---|---|---|
| | *Orig* | — | *3.18* |
| 8 | STFT 512/2048 | 3080.52 / 3152.88 | 6.90 / 5.31 |
| 8 | STFChT | 5236.74 | 6.33 |
| 2 | STFT 512/2048 | 852.84 / 934.18 | 6.29 / 4.57 |
| 2 | STFChT | 3036.49 | 5.24 |
| 1 | STFT 512/2048 | 610.26 / 682.97 | 5.80 / 4.21 |
| 1 | STFChT | 2753.87 | 4.85 |

RealData, far distance

| Ch. | Method | Comp. Time (s) | SRMR |
|---|---|---|---|
| | *Orig* | — | *3.19* |
| 8 | STFT 512/2048 | 2908.92 / 2977.25 | 6.99 / 5.56 |
| 8 | STFChT | 4962.74 | 6.52 |
| 2 | STFT 512/2048 | 810.37 / 943.14 | 6.25 / 4.63 |
| 2 | STFChT | 2922.66 | 5.25 |
| 1 | STFT 512/2048 | 754.06 / 870.11 | 5.73 / 4.24 |
| 1 | STFChT | 2624.72 | 4.82 |

RealData, near distance

| Ch. | Method | Comp. Time (s) | SRMR |
|---|---|---|---|
| | *Orig* | — | *3.18* |
| 8 | STFT 512/2048 | 3252.12 / 3328.51 | 6.81 / 5.05 |
| 8 | STFChT | 5510.74 | 6.13 |
| 2 | STFT 512/2048 | 895.3 / 925.23 | 6.33 / 4.51 |
| 2 | STFChT | 3150.32 | 5.22 |
| 1 | STFT 512/2048 | 466.47 / 495.82 | 5.87 / 4.17 |
| 1 | STFChT | 2883.02 | 4.87 |

Figure 5.8: Results for SimData and RealData evaluation sets. STFT 512 is short-window STFT processing (conventional), STFT 2048 is long-window STFT processing (conventional, designed to match STFChT parameters), and STFChT is short-time fan-chirp transform-based processing (proposed). CD is cepstral distance, SRMR is speech-to-reverberation modulation energy ratio, LLR is log-likelihood ratio, FWSegSNR is frequency-weighted segmental signal-to-noise ratio, and PESQ is Perceptual Evaulation of Speech Quality.

troducing artifacts, which concurs with substantial increase in the PESQ scores, the ITU-T standard for voice quality. We also provided insight as to why enhancement performance improves using the STFChT domain. The improvement gained by STFChT-based processing is an interesting result, and warrants further investigation. Further exploration may be fruitful, as combination of the fan-chirp or other coherent transforms with other methods for dereverberation and/or noise reduction may yield improved results.

# Chapter 6
# CONCLUSIONS AND FUTURE WORK

In this thesis, we examined two simple subclasses of nonstationary random processes. In chapter two, we defined these processes, amplitude-modulated wide-sense stationary (AM-WSS) and filtered jittered pulse trains (F-JPT), and examined their second-order spectral statistics. We found that these processes exhibit substantial spectral correlation beyond what is usually measured when data is assumed stationary. In fact, we saw that when the fundamental frequency of the periodic components of these processes is constant, these processes exhibit the same spectral correlation structure as that of cyclostationary processes. In addition, the spectral increments of these processes exhibit correlation with their conjugates, which provides even more additional information beyond what is usually exploited in conventional processing. We also saw that the spectral covariance functions of these processes become smeared when the fundamental frequency of the periodic components of these processes varies over the analysis duration, which indicates the necessity of performing compensation of the time-varying fundamental frequency to restore the narrow spectral ridges.

To solve the problem of smeared spectral correlation caused by time-varying fundamental frequency, we proposed to measure the time-varying fundamental frequency and apply a time-warping to the signal that makes the fundamental frequency constant, and thus restore concentrated peaks to the spectrum of the signal. To do so, we reviewed an implementation of this method, the fan-chirp transform, in chapter 3. We provided measurements of the peakiness of the spectra of time-warped processes with time-varying fundamental frequency using normalized entropy, which showed the method was effective in restoring narrow, high-amplitude spectral peaks.

In chapter 4, we presented an improved detection scheme for AM-WSS signals with time-varying fundamental modulation frequency that uses an equivalent of time-warping

to increase the length of the analysis window and thus increase the detection statistic. We demonstrated the benefit of this method for improved detection of both synthetic and real-world ship noise in passive sonar.

In chapter 5, we combined the forward and inverse fan-chirp transform with an optimal enhancement algorithm that suppresses both noise and reverberation in speech signals. We presented superior results on the REVERB challenge dataset versus standard STFT processing, which assumes the fundamental frequency of speech does not vary over the analysis frame duration. By relaxing this assumption that fundamental frequency is constant over the analysis frame, estimating a linear model of the time-varying fundamental frequency, and compensating the frame data using time-warping, we were able to use longer analysis windows and thus achieved better enhancement performance.

Armed with the preliminary analysis of chapter 2 and the tools of chapter 3, and encouraged by the positive results in chapters 4 and 5, future work will investigate exploiting the additional information present in spectral correlations. Spectral correlation off the stationary manifold (which is the only information conventionally used) indicates the usefulness of combining frequency-shifted copies of a nonstationary signal, while the additional information present in the complementary spectral covariance function recommends the use of widely-linear estimators that are functions of both complex-valued data and the conjugate of complex-valued data [26]. Since the structure of the the spectral covariance functions for our signals of interest resembles that of cyclostationary signals, it would be interesting to attempt applying some of the algorithms [2]–[4], [6], [7], [77], [78] designed for cyclostationary signals to speech and sonar signals, especially since we now have the tools and understanding to apply a time-warping to data such that the signals are more cyclostationary. Furthermore, widely-linear versions of these algorithms [5], [27], [43] that exploit the information from both the Hermitian and complementary statistics could potentially improve performance even more.

# REFERENCES


[1] T. F Quatieri, *Principles of Discrete-Time Speech Processing*, English. Upper Saddle River, NJ: Prentice Hall, 2001.

[2] K. Abed-Meraim, Y. Xiang, J. Manton, and Y. Hua, "Blind source-separation using second-order cyclostationary statistics," *IEEE Transactions on Signal Processing*, vol. 49, no. 4, pp. 694–701, Apr. 2001.

[3] B. Agee, S. Schell, and W. Gardner, "Spectral self-coherence restoral: a new approach to blind adaptive signal extraction using antenna arrays," *Proceedings of the IEEE*, vol. 78, no. 4, pp. 753–767, 1990.

[4] Q. Wu and K. M. Wong, "Blind adaptive beamforming for cyclostationary signals," *IEEE Transactions on Signal Processing*, vol. 44, no. 11, pp. 2757–2767, Nov. 1996.

[5] W. Zhang and W. Liu, "Low-complexity blind beamforming based on cyclostationarity," in *Proc. European Signal Processing Conference (EUSIPCO)*, Aug. 2012, pp. 1–5.

[6] W. Gardner, "Cyclic wiener filtering: theory and method," *IEEE Transactions on Communications*, vol. 41, no. 1, pp. 151–163, 1993.

[7] W. Gardner and C. Spooner, "Detection and source location of weak cyclostationary signals: simplifications of the maximum-likelihood receiver," *IEEE Transactions on Communications*, vol. 41, no. 6, pp. 905–916, Jun. 1993.

[8] M. B. Priestley and T. Subba Rao, "A test for non-stationarity of time-series," *Journal of the Royal Statistical Society. Series B (Methodological)*, 140–149, 1969.

[9] P. Borgnat and P. Flandrin, "Stationarization via surrogates," en, *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2009, no. 01, P01001, Jan. 2009.

[10] P. Borgnat, P. Flandrin, P. Honeine, C. Richard, and J. Xiao, "Testing stationarity with surrogates: a time-frequency approach," *IEEE Transactions on Signal Processing*, vol. 58, no. 7, pp. 3459–3470, Jul. 2010.

[11] R. Von Sachs and M. H. Neumann, "A wavelet-based test for stationarity," *Journal of Time Series Analysis*, vol. 21, no. 5, 597–613, 2000.

[12] G. Nason, "A test for second-order stationarity and approximate confidence intervals for localized autocovariances for locally stationary time series," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 75, no. 5, 879–904, 2013.

[13] H. L. Gray and N. F. Zhang, "On a class of nonstationary processes," *Journal of Time Series Analysis*, vol. 9, no. 2, 133–154, 1988.

[14] H. L. Gray, C.-P. C. Vijverberg, and W. A. Woodward, "Nonstationary data analysis by time deformation," *Communications in Statistics—Theory and Methods*, vol. 34, no. 1, 163–192, 2005.

[15] H. Jiang, H. L. Gray, and W. A. Woodward, "Time-frequency analysis—-stationary processes," *Computational Statistics & Data Analysis*, vol. 51, no. 3, pp. 1997–2028, Dec. 2006.

[16] M. Xu, W. A. Woodward, and H. L. Gray, "Using time deformation to filter nonstationary time series with multiple time-frequency structures," *Journal of Probability and Statistics*, vol. 2013, Mar. 2013.

[17] A. Veeraraghavan, R. Chellappa, and A. Roy-Chowdhury, "The function space of an activity," in *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2006, pp. 959–968.

[18] J. Andén and S. Mallat, "Deep scattering spectrum," *arXiv:1304.6763*, Apr. 2013.

[19] W. Gardner, "Stationarizable random processes," *IEEE Transactions on Information Theory*, vol. 24, no. 1, pp. 8–22, 1978.

[20] E. D. Gladyshev, "Periodically correlated random sequences," *Soviet Math. Dokl.*, vol. 2, pp. 385–388, 1961.

[21] E. G. Gladyshev, "Periodically and almost-periodically correlated random processes with a continuous time parameter," *Theory of Probability & Its Applications*, vol. 8, no. 2, pp. 173–177, Jan. 1963.

[22] D. B. Percival and A. T. Walden, *Spectral Analysis for Physical Applications*. Cambridge University Press, Jun. 1993.

[23] P. J. Schreier and L. L. Scharf, *Statistical Signal Processing of Complex-Valued Data: The Theory of Improper and Noncircular Signals*, en. Cambridge University Press, Feb. 2010.

[24] M Loève, *Probability Theory*. New York: Springer-Verlag, 1977.

[25] W. M. Brown and R. B. Crane, "Conjugate linear filtering," *IEEE Transactions on Information Theory*, vol. 15, no. 4, pp. 462–465, 1969.

[26] T. Adali, P. Schreier, and L. Scharf, "Complex-valued signal processing: the proper way to deal with impropriety," *IEEE Transactions on Signal Processing*, vol. 59, no. 11, pp. 5101–5125, 2011.

[27] D. P. Mandic and V. S. L. Goh, *Complex valued nonlinear adaptive filters noncircularity, widely linear, and neural models*. Hoboken, N.J.: Wiley, 2009.

[28] B. Picinbono, "On circularity," *IEEE Transactions on Signal Processing*, vol. 42, no. 12, pp. 3473–3482, Dec. 1994.

[29] B. Picinbono and P. Chevalier, "Widely linear estimation with complex data," *IEEE Transactions on Signal Processing*, vol. 43, no. 8, pp. 2030 –2033, Aug. 1995.

[30] B. Picinbono and P. Bondon, "Second-order statistics of complex signals," *IEEE Transactions on Signal Processing*, vol. 45, no. 2, pp. 411–420, 1997.

[31] F. Neeser and J. Massey, "Proper complex random processes with applications to information theory," *IEEE Transactions on Information Theory*, vol. 39, no. 4, pp. 1293–1302, Jul. 1993.

[32] P. Clark, "Coherent demodulation of nonstationary random processes," PhD thesis, University of Washington, 2012.

[33] A. Napolitano, *Generalizations of Cyclostationary Signal Processing: Spectral Analysis and Applications*, en. John Wiley & Sons, Aug. 2012.

[34] L. Atlas and S. A. Shamma, "Joint acoustic and modulation frequency," *EURASIP Journal on Advances in Signal Processing*, vol. 2003, no. 7, p. 310 290, Jun. 2003.

[35] P. Clark, I. Kirsteins, and L. Atlas, "Existence and estimation of impropriety in real rhythmic signals," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 3713–3716.

[36] W. Gardner and L. Franks, "Characterization of cyclostationary random signal processes," *IEEE Transactions on Information Theory*, vol. 21, no. 1, pp. 4–14, 1975.

[37] H. Hurd, "Spectral coherence of nonstationary and transient stochastic processes," in *Proc. ASSP Workshop on Spectrum Estimation and Modeling*, 1988, pp. 387–390.

[38] J. Lourens and J. du Preez, "Passive sonar ML estimator for ship propeller speed," *IEEE Journal of Oceanic Engineering*, vol. 23, no. 4, pp. 448–453, 1998.

[39] P. Clark, I. Kirsteins, and L. Atlas, "Multiband analysis for colored amplitude-modulated ship noise," in *Proc. IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Dallas, TX, 2010, pp. 3970–3973.

[40] J. Schoentgen and R. De Guchteneere, "Predictable and random components of jitter," *Speech Communication*, vol. 21, no. 4, pp. 255–272, May 1997.

[41] D. Thomson and R. Chengalvarayan, "Use of periodicity and jitter as speech recognition features," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, May 1998, 21–24 vol.1.

[42] P. C. Bagshaw, S. M. Hiller, and M. A. Jack, "Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching.," International Speech Communication Association, 1993.

[43] S. C. Douglas, "Widely-linear recursive least-squares algorithm for adaptive beamforming," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, ser. ICASSP '09, Washington, DC, USA: IEEE Computer Society, 2009, 2041–2044.

[44] P. Cancela, E. López, and M. Rocamora, "Fan chirp transform for music representation," in *Proc. International Conference On Digital Audio Effects (DAFx)*, Graz, Austria, 2010, 1–8.

[45] M. Képesi and L. Weruaga, "Adaptive chirp-based time–frequency analysis of speech signals," *Speech Communication*, vol. 48, no. 5, pp. 474–492, May 2006.

[46] L. Weruaga and M. Képesi, "Speech analysis with the fast chirp transform," in *Proc. European Signal Processing Conference (EUSIPCO)*, Vienna, Austria, 2004, 1011–1014.

[47] T. M Cover and J. A. Thomas, *Elements of information theory*, English. Hoboken, N.J.: Wiley-Interscience, 2006.

[48] S. Wisdom, L. Atlas, and J. Pitton, "Extending coherence time for analysis of modulated random processes," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy, 2014, preprint.

[49] R. Tao, Y. Feng, and Y. Wang, "Theoretical and experimental study of a signal feature extraction algorithm for measuring propeller acceleration in a port surveillance system," *IET Radar, Sonar Navigation*, vol. 5, no. 2, pp. 172–181, 2011.

[50] H. Omer and B. Torrésani, "Estimation of frequency modulations on wideband signals; applications to audio signal analysis," arXiv e-print 1305.3095, May 2013.

[51] K. Kaewtip, L. N. Tan, and A. Alwan, "A pitch-based spectral enhancement technique for robust speech processing," in *Proc. Interspeech*, Lyon, France, 2013.

[52] S. M. Kay, *Detection Theory*, English. Upper Saddle River, NJ: Prentice Hall PTR, 1998.

[53] S. Wisdom, T. Powers, L. Atlas, and J. Pitton, "Enhancement of reverberant and noisy speech by extending its coherence," in *Proc. REVERB Challenge Workshop*, Florence, Italy, 2014, preprint.

[54] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, and R. Maas, "The REVERB challenge: a common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, 2013.

[55] E. A. P. Habets, "Speech dereverberation using statistical reverberation models," in *Speech Dereverberation*, Patrick A. Naylor and Nikolay D. Gaubitch, Eds., Springer, Jul. 2010.

[56] L. Weruaga and M. Képesi, "The fan-chirp transform for non-stationary harmonic signals," *Signal Processing*, vol. 87, no. 6, pp. 1504–1522, Jun. 2007.

[57] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function GSC and postfiltering," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 6, pp. 561–571, 2004.

[58] R. Maas, E. A. P. Habets, A. Sehr, and W. Kellermann, "On the application of reverberation suppression to robust speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, Japan, 2012, pp. 297–300.

[59] T. Nakatani, K. Kinoshita, and M. Miyoshi, "Harmonicity-based blind dereverberation for single-channel speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 80–95, 2007.

[60] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3, 187–207, 1999.

[61] J. Droppo and A. Acero, "A fine pitch model for speech.," in *Proc. Interspeech*, Antwerp, Belgium, 2007, 2757–2760.

[62] R. Dunn and T. Quatieri, "Sinewave analysis/synthesis based on the fan-chirp tranform," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, 2007, pp. 247–250.

[63] R. Dunn, T. Quatieri, and N. Malyska, "Sinewave parameter estimation using the fast fan-chirp transform," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, 2009, pp. 349–352.

[64] Y. Pantazis, O. Rosec, and Y. Stylianou, "Adaptive AM-FM signal decomposition with application to speech analysis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 290–300, 2011.

[65] G. Degottex and Y. Stylianou, "Analysis and synthesis of speech using an adaptive full-band harmonic model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9-10, 2085–2095, 2013.

[66] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[67] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[68] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Processing Letters*, vol. 9, no. 4, pp. 113–116, 2002.

[69] H. Löllmann, E. Yilmaz, M. Jeub, and P. Vary, "An improved algorithm for blind reverberation time estimation," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Tel Aviv, Israel, 2010, 1–4.

[70] P. C. Loizou, *Speech Enhancement: Theory and Practice*, en. CRC Press, Jun. 2007.

[71] H. L. Van Trees, *Optimum Array Processing. Part IV of Detection, Estimation, and Modulation Theory*, English. New York: Wiley-Interscience, 2002.

[72] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "Wsjcam0: a british english speech corpus for large vocabulary continuous speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 1995, 81–84.

[73] M. Lincoln, I. McCowan, J. Vepa, and H. Maganti, "The multi-channel wall street journal audio visual corpus (MC-WSJ-AV): specification and initial experiments," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, Nov. 2005, pp. 357–362.

[74] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, Salt Lake City, UT, 2001, 749–752.

[75] T. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, 2010.

[76] ITU-T P.862.2, *Wideband extension to rec. p.862 for the assessment of wideband telephone networks and speech codecs*, 2007.

[77] Y. Li and Z. Ding, "Blind channel identification based on second order cyclostationary statistics," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, Apr. 1993, 81–84 vol.4.

[78] L. Tong, G. Xu, and T. Kailath, "Blind identification and equalization based on second-order statistics: a time domain approach," *IEEE Transactions on Information Theory*, vol. 40, no. 2, pp. 340–349, Mar. 1994.

Appendix A

# DERIVATION OF SPECTRAL SECOND-ORDER STATISTICS OF AM-WSS PROCESSES

Recall from section 2.3 that the Fourier transform of a AM-WSS process is

$$X(f) = \int M(f-u)dZ(u) \tag{A.1}$$

## *A.1 Hermitian Spectral Covariance When $w(t)$ is White*

Here we prove (2.21). The Hermitian spectral covariance is

$$\begin{aligned} R_{XX}(f_1,f_2) &= E\left[X(f_1)X^*(f_2)\right] \\ &= E\left[\int M(f_1-u)dZ(u)\int M^*(f_2-v)dZ^*(v)\right] \\ &= E\left[\int\int M(f_1-u)M^*(f_2-v)dZ(u)dZ^*(v)\right] \\ &= \int\int M(f_1-u)M^*(f_2-v)E\left[dZ(u)dZ^*(v)\right] \\ &= \int\int M(f_1-u)M^*(f_2-v)\sigma_w^2\delta(u-v)dudv \\ &= \sigma_w^2\int M(f_1-u)M^*(f_2-u)du \end{aligned} \tag{A.2}$$

If $m(t)$ is real, $M(f) = M^*(-f)$, and the term $M^*(f_2-u)$ in (A.2) may be written as $M(u-f_2)$. Using a variable substitution $\mu \leftarrow u - f_2$, (A.2) becomes

$$\begin{aligned} R_{XX}(f_1,f_2) &= \sigma_w^2\int M(f_1-f_2-\mu)M(\mu)d\mu \\ &= \sigma_w^2 M(f_1-f_2) * M(f_1-f_2) \\ &= \sigma_w^2\int m^2(t)e^{-j2\pi t(f_1-f_2)}dt \end{aligned} \tag{A.3}$$

## *A.2 Hermitian Spectral Covariance When $w(t)$ is Arbitrary WSS*

Here we prove (2.24). We take $w(t)$ to be a real-valued WSS random process with spectral increments $dZ(f)$ and power spectral density $S_w(f)df = E[dZ(f_1)dZ^*(f_2)]$. The Hermitian

spectral covariance is

$$\begin{aligned} R_{XX}(f_1, f_2) &= E\left[X(f_1)X^*(f_2)\right] \\ &= E\left[\int M(f_1-u)dZ(u)\int M^*(f_2-v)dZ^*(v)\right] \\ &= E\left[\int\int M(f_1-u)M^*(f_2-v)dZ(u)dZ^*(v)\right] \\ &= \int\int M(f_1-u)M^*(f_2-v)E\left[dZ(u)dZ^*(v)\right] \\ &= \int M(f_1-u)M^*(f_2-u)S_w(u)du \end{aligned} \tag{A.4}$$

Further simplification is difficult because of the $S_w(u)$ within the integral, which acts as a kernel.

### *A.3 Complementary Spectral Covariance When $w(t)$ is White*

The proof for (2.22) follows similarly to section A.1. The spectral complementary covariance is

$$\begin{aligned} \widetilde{R}_{XX}(f_1, f_2) &= E\left[X(f_1)X(-f_2)\right] \\ &= E\left[\int M(f_1-u)dZ(u)\int M(f_2-v)dZ(v)\right] \\ &= E\left[\int\int M(f_1-u)M(f_2-v)dZ(u)dZ(v)\right] \\ &= \int\int M(f_1-u)M(f_2-v)E\left[dZ(u)dZ(v)\right] \\ &= \int\int M(f_1-u)M(f_2-v)\sigma_w^2\delta(u+v)dudv \\ &= \sigma_w^2\int M(f_1-u)M(f_2+u)du \end{aligned} \tag{A.5}$$

Using a variable substitution $\mu \leftarrow u + f_2$, (A.5) becomes

$$\begin{aligned} \widetilde{R}_{XX}(f_1, f_2) &= \sigma_w^2\int M(f_1+f_2-\mu)M(\mu)d\mu \\ &= \sigma_w^2 M(f_1+f_2) * M(f_1+f_2) \\ &= \sigma_w^2\int m^2(t)e^{-j2\pi t(f_1+f_2)}dt \end{aligned} \tag{A.6}$$

### *A.4 Complementary Spectral Covariance When $w(t)$ is Arbitrary WSS*

Here we prove (2.25). We take $w(t)$ to be a real-valued WSS random process with spectral increments $dZ(f)$ and power spectral density $S_w(f)df = E[dZ(f_1)dZ^*(f_2)]$. The complementary spectral covariance is

$$\begin{aligned}
\widetilde{R}_{XX}(f_1, f_2) &= E\left[X(f_1)X(f_2)\right] \\
&= E\left[\int M(f_1 - u)dZ(u) \int M(f_2 - v)dZ(v)\right] \\
&= E\left[\int\int M(f_1 - u)M(f_2 - v)dZ(u)dZ(v)\right] \\
&= \int\int M(f_1 - u)M(f_2 - v)E\left[dZ(u)dZ(v)\right]
\end{aligned}$$

Using the fact that $E[dZ(f_1)dZ(f_2)] = \begin{cases} S_w(f)df & \text{if } f_1 = -f_2 = f \\ 0 & \text{otherwise} \end{cases}$, the expression becomes a single integral and $v \leftarrow -u$, so

$$= \int M(f_1 - u)M(f_2 + u)S_w(u)du \tag{A.7}$$

Further simplification is difficult because of the $S_w(u)$ within the integral, which acts as a kernel.

Appendix B

# DERIVATION OF SPECTRAL SECOND-ORDER STATISTICS OF F-JPT

Recall from section 2.4 that a F-JPT random process $x(t)$ of duration $(2N+1)\mathcal{T}_0$ is

$$x(t) = h(t) * d(t) = h(t) * \sum_{n=-N}^{N} g(t - n\mathcal{T}_0 + k_n) \tag{B.1}$$

where $g(t)$ is a unit-energy rectangular pulse such that

$$g(t) = \begin{cases} \frac{1}{\sqrt{T_g}}, & \text{for } -\frac{T_g}{2} \le t \le \frac{T_g}{2} \\ 0 & \text{otherwise.} \end{cases} \tag{B.2}$$

The Fourier transform of a F-JPT random process $x(t)$ is

$$X(f) = H(f) \sum_{n=-N}^{N} G(f) e^{-j2\pi f(n\mathcal{T}_0 - k_n)}, \tag{B.3}$$

where $k_n$ are i.i.d. random variables uniformly distributed between $-D/2$ and $D/2$.

## *B.1 Hermitian Spectral Correlation*

The steps to derive the Hermitian spectral correlation $S_{XX}(f_1, f_2)$ of the F-JPT with an infinitely narrow pulse width are as follows. First, define $S_{XX}^{T_g}(f_1, f_2)$ to be the Hermitian spectral correlation of the F-JPT with a pulse width of $T_g$:

$$\begin{aligned} S_{XX}^{T_g}(f_1, f_2) \triangleq & E\left[ \left( H(f_1) \sum_{n_1=-N}^{N} G(f_1) e^{-j2\pi f_1(n_1\mathcal{T}_0 - k_{n_1})} \right) \right. \\ & \left. \cdot \left( H^*(f_2) \sum_{n_2=-N}^{N} G^*(f_2) e^{j2\pi f_2(n_2\mathcal{T}_0 - k_{n_2})} \right) \right] \\ = & H(f_1)H^*(f_2) \\ & \cdot \sum_{n_1=-N}^{N} \sum_{n_2=-N}^{N} G(f_1)G^*(f_2) e^{j2\pi\mathcal{T}_0(n_2 f_2 - n_1 f_1)} E\left[ e^{j2\pi f_1 k_{n_1}} e^{-j2\pi f_2 k_{n_2}} \right]. \end{aligned} \tag{B.4}$$

We are interested in the case when the pulse $g(t)$ is infinitely narrow. Thus, we take the limit as the pulse duration $T_g \to 0$. Note that

$$\lim_{T_g \to 0} G(f) = \lim_{T_g \to 0} \frac{\sin \pi T_g f}{\pi T_g f} = 1$$

so

$$\begin{aligned} S_{XX}(f_1, f_2) = & \lim_{T_g \to 0} S_{XX}^{T_g}(f_1, f_2) \\ = & H(f_1) H^*(f_2) \\ & \cdot \sum_{n_1=-N}^{N} \sum_{n_2=-N}^{N} \lim_{T_g \to 0} [G(f_1)] \lim_{T_g \to 0} [G^*(f_2)] \, e^{j2\pi \mathcal{T}_0 (n_2 f_2 - n_1 f_1)} E\left[ e^{j2\pi f_1 k_{n_1}} e^{-j2\pi f_2 k_{n_2}} \right] \\ = & H(f_1) H^*(f_2) \sum_{n_1=-N}^{N} \sum_{n_2=-N}^{N} e^{j2\pi \mathcal{T}_0 (n_2 f_2 - n_1 f_1)} E\left[ e^{j2\pi f_1 k_{n_1}} e^{-j2\pi f_2 k_{n_2}} \right]. \end{aligned} \tag{B.5}$$

At this point, we need to evaluate the expected value. We have two cases: $n_1 = n_2$ and $n_1 \neq n_2$. Recall that $k_{n_1} \perp k_{n_2}$ for $n_1 \neq n_2$, which means

$$\begin{aligned} & E[f(k_{n_1}) g(k_{n_2})] - E[f(k_{n_1})] \cdot E[g(k_{n_2})] = 0 && \text{for } n_1 \neq n_2 \\ \Rightarrow & E[f(k_{n_1}) g(k_{n_2})] = E[f(k_{n_1})] \cdot E[g(k_{n_2})] && \text{for } n_1 \neq n_2. \end{aligned}$$

When $n_1 = n_2 = n$,

$$E\left[ e^{j2\pi f_1 k_{n_1}} e^{-j2\pi f_2 k_{n_2}} \right] = E\left[ e^{-j2\pi k_n (f_2 - f_1)} \right].$$

To actually evaluate the expected value for the 2 cases, $\{1\}$ $n_1 = n_2$ and $\{2\}$ $n_1 \neq n_2$, use the definition of expected value, $E[g(x)] = \int_{-\infty}^{\infty} g(x) p_X(x) dx$, where $p_X(x)$ is the pdf of R.V. $x$.

Case $\{1\}$, $n_1 \neq n_2$:

$$\begin{aligned} E\left[ e^{j2\pi f_1 k_{n_1}} e^{-j2\pi f_2 k_{n_2}} \right] &= E\left[ e^{j2\pi f_1 k_{n_1}} \right] E\left[ e^{-j2\pi f_2 k_{n_2}} \right] \\ &= \left( \int_{-D/2}^{D/2} e^{j2\pi f_1 k_{n_1}} \frac{1}{D} dk_{n_1} \right) \left( \int_{-D/2}^{D/2} e^{-j2\pi f_2 k_{n_2}} \frac{1}{D} dk_{n_2} \right) \\ &= \left( \left[ \frac{1}{j2\pi D f_1} e^{j2\pi f_1 k_{n_1}} \right]_{-D/2}^{D/2} \right) \left( \left[ \frac{1}{-j2\pi D f_2} e^{-j2\pi f_2 k_{n_2}} \right]_{-D/2}^{D/2} \right) \\ &= \frac{\sin(\pi D f_1)}{\pi D f_1} \frac{\sin(\pi D f_2)}{\pi D f_2}. \end{aligned} \tag{B.6}$$

Case $\{2\}$, $n_1 = n_2 = n$:

$$\begin{aligned} E\left[e^{j2\pi f_1 k_{n_1}} e^{-j2\pi f_2 k_{n_2}}\right] &= E\left[e^{-j2\pi k_n (f_2 - f_1)}\right] \\ &= \int_{-D/2}^{D/2} e^{-j2\pi k_n (f_2 - f_1)} \frac{1}{D} dk_n \\ &= \left[\frac{1}{-j2\pi k_n (f_2 - f_1)} e^{-j2\pi k_n (f_2 - f_1)}\right]_{-D/2}^{D/2} \\ &= \frac{1}{-j2\pi k_n (f_2 - f_1)} \left(e^{-j2\pi \frac{D}{2}(f_2 - f_1)} - e^{j2\pi \frac{D}{2}(f_2 - f_1)}\right) \\ &= \frac{\sin\left(\pi D (f_2 - f_1)\right)}{\pi D (f_2 - f_1)}. \end{aligned} \tag{B.7}$$

The expected values split up the summations in (B.5) as

$$\begin{aligned} S_{XX}(f_1, f_2) =& H(f_1) H^*(f_2) \left( \sum_{n=-N}^{N} e^{j2\pi \mathcal{T}_0 n (f_2 - f_1)} E\left[e^{-j2\pi k_n (f_2 - f_1)}\right] \right. \\ &+ \sum_{n_1=-N}^{N} \sum_{n_2=-N}^{N} e^{j2\pi \mathcal{T}_0 (n_2 f_2 - n_1 f_1)} E\left[e^{j2\pi f_1 k_{n_1}}\right] E\left[e^{-j2\pi f_2 k_{n_2}}\right] \\ &\left. - \sum_{n=-N}^{N} e^{j2\pi \mathcal{T}_0 n (f_2 - f_1)} E\left[e^{j2\pi f_1 k_n}\right] E\left[e^{-j2\pi f_2 k_n}\right] \right). \end{aligned}$$

Now plug in (B.6) and (B.7) for the expected value expressions:

$$\begin{aligned} =& H(f_1) H^*(f_2) \left( \sum_{n=-N}^{N} e^{j2\pi \mathcal{T}_0 n (f_2 - f_1)} \frac{\sin\left(\pi D (f_2 - f_1)\right)}{\pi D (f_2 - f_1)} \right. \\ &+ \sum_{n_1=-N}^{N} \sum_{n_2=-N}^{N} e^{j2\pi \mathcal{T}_0 (n_2 f_2 - n_1 f_1)} \frac{\sin(\pi D f_1)}{\pi D f_1} \frac{\sin(\pi D f_2)}{\pi D f_2} \\ &\left. - \sum_{n=-N}^{N} e^{j2\pi \mathcal{T}_0 n (f_2 - f_1)} \frac{\sin(\pi D f_1)}{\pi D f_1} \frac{\sin(\pi D f_2)}{\pi D f_2} \right) \\ =& H(f_1) H^*(f_2) \left( \frac{\sin\left(\pi D (f_2 - f_1)\right)}{\pi D (f_2 - f_1)} \sum_{n=-N}^{N} e^{j2\pi \mathcal{T}_0 n (f_2 - f_1)} \right. \\ &+ \frac{\sin(\pi D f_1)}{\pi D f_1} \frac{\sin(\pi D f_2)}{\pi D f_2} \left( \sum_{n_2=-N}^{N} e^{j2\pi \mathcal{T}_0 (n_2 f_2)} \right) \left( \sum_{n_1=-N}^{N} e^{-j2\pi \mathcal{T}_0 (n_1 f_1)} \right) \\ &\left. - \frac{\sin(\pi D f_1)}{\pi D f_1} \frac{\sin(\pi D f_2)}{\pi D f_2} \sum_{n=-N}^{N} e^{j2\pi \mathcal{T}_0 n (f_2 - f_1)} \right). \end{aligned} \tag{B.8}$$

Now all that remains is to evaluate the summations in (B.8). Recall that

$$\sum_{n=-N}^{N} e^{jAnx} = \frac{\sin(\frac{A}{2}(2N+1)x)}{\sin(\frac{A}{2}x)}, \tag{B.9}$$

where $x$ is a continuous deterministic variable and $A$ is a constant. (B.9) is known as Dirichlet's kernel, which is a periodic function with period $2\pi/A$. Each period resembles a sinc function, and the peaks of each period occur at $n2\pi/A$, $-N < n < N$.

An interesting property of the Dirichlet kernel is that as $N \to \infty$, the sinc functions narrow until they become Dirac delta functions:

$$\lim_{N\to\infty} \sum_{n=-N}^{N} e^{jAnx} = \sum_{n=-\infty}^{\infty} \delta(x - n\frac{2\pi}{A}). \tag{B.10}$$

In our case, $A = 2\pi\mathcal{T}_0$ and $x = f$. Using the identity (B.9), (B.8) simplifies to

$$\begin{aligned} R_{XX}(f_1, f_2) = H(f_1)H^*(f_2) &\left( \frac{\sin(\pi D(f_2 - f_1))}{\pi D(f_2 - f_1)} - \frac{\sin(\pi D f_1)}{\pi D f_1} \frac{\sin(\pi D f_2)}{\pi D f_2} \right) \\ &\cdot \frac{\sin[\pi(2N+1)\mathcal{T}_0(f_2 - f_1)]}{\sin(\pi\mathcal{T}_0(f_2 - f_1)}. \end{aligned} \tag{B.11}$$

In this form, it is easy to see the effect of both the fundamental frequency $\mathcal{F}_\prime = 1/\mathcal{T}_0$ and amount of jitter $D$ of the glottal pulse excitation on the Hermitian correlation.

### *B.2 Complementary Spectral Correlation*

Having derived an expression for $S_{XX}(f_1, f_2)$, it is relatively easy to find an expression for the complementary spectral correlation, $\widetilde{S}_{XX}(f_1, f_2) = E[X(f_1)X(f_2)]$ for a finite-duration F-JPT process $x(t)$ with infinitely narrow pulse width. The only differences from the Hermitian spectral correlation $S_{XX}(f_1, f_2)$ are the lack of conjugation of $H(f_2)$ and the

sign of $f_2$, which is flipped because of the lack of conjugation on $X(f_2)$:

$$\begin{aligned}
\widetilde{S}_{XX}(f_1,f_2) &= E\left[\left(H(f_1)\sum_{n_1=-N}^{N} e^{-j2\pi f_1(n_1\mathcal{T}_0-k_{n_1})}\right)\left(H(f_2)\sum_{n_2=-N}^{N} e^{-j2\pi f_2(n_2\mathcal{T}_0-k_{n_2})}\right)\right] \\
&= H(f_1)H(f_2)\sum_{n_1=-N}^{N}\sum_{n_2=-N}^{N} e^{-j2\pi\mathcal{T}_0(n_2f_2+n_1f_1)}E\left[e^{j2\pi f_1k_{n_1}}e^{j2\pi f_2k_{n_2}}\right] \\
&= H(f_1)H(f_2)\left(\sum_{n=-N}^{N} e^{-j2\pi\mathcal{T}_0n(f_2+f_1)}E\left[e^{j2\pi k_n(f_2+f_1)}\right]\right. \\
&\quad + \sum_{n_1=-N}^{N}\sum_{n_2=-N}^{N} e^{-j2\pi\mathcal{T}_0(n_2f_2+n_1f_1)}E\left[e^{j2\pi f_1k_{n_1}}\right]E\left[e^{j2\pi f_2k_{n_2}}\right] \\
&\quad \left. - \sum_{n=-N}^{N} e^{-j2\pi\mathcal{T}_0n(f_2+f_1)}E\left[e^{j2\pi f_1k_n}\right]E\left[e^{j2\pi f_2k_n}\right]\right).
\end{aligned}$$

Using (B.6) and (B.7), replace expected values with sincs:

$$\begin{aligned}
&= H(f_1)H(f_2)\left(\sum_{n=-N}^{N} e^{-j2\pi\mathcal{T}_0n(f_2+f_1)}\frac{\sin\left(\pi D(f_2+f_1)\right)}{\pi D(f_2+f_1)}\right. \\
&\quad + \sum_{n_1=-N}^{N}\sum_{n_2=-N}^{N} e^{-j2\pi\mathcal{T}_0(n_2f_2+n_1f_1)}\frac{\sin(\pi Df_1)}{\pi Df_1}\frac{\sin(\pi Df_2)}{\pi Df_2} \\
&\quad \left. - \sum_{n=-N}^{N} e^{-j2\pi\mathcal{T}_0n(f_2+f_1)}\frac{\sin(\pi Df_1)}{\pi Df_1}\frac{\sin(\pi Df_2)}{\pi Df_2}\right),
\end{aligned}$$

and move the sincs out of the sums:

$$\begin{aligned}
&= H(f_1)H(f_2)\left(\frac{\sin\left(\pi D(f_2+f_1)\right)}{\pi D(f_2+f_1)}\sum_{n=-N}^{N} e^{-j2\pi\mathcal{T}_0n(f_2+f_1)}\right. \\
&\quad + \frac{\sin(\pi Df_1)}{\pi Df_1}\frac{\sin(\pi Df_2)}{\pi Df_2}\left(\sum_{n_2=-N}^{N} e^{-j2\pi\mathcal{T}_0n_2f_2}\right)\left(\sum_{n_1=-N}^{N} e^{-j2\pi\mathcal{T}_0n_1f_1}\right) \\
&\quad \left. - \frac{\sin(\pi Df_1)}{\pi Df_1}\frac{\sin(\pi Df_2)}{\pi Df_2}\sum_{n=-N}^{N} e^{-j2\pi\mathcal{T}_0n(f_2+f_1)}\right).
\end{aligned}$$

Finally, use (B.9) to simplify the summations:

$$\begin{aligned}\widetilde{S}_{XX}(f_1,f_2) =& H(f_1)H(f_2)\left(\frac{\sin\left(\pi D(f_2+f_1)\right)}{\pi D(f_2+f_1)}\frac{\sin\left[\pi(2N+1)\mathcal{T}_0(f_2+f_1)\right]}{\sin(\pi\mathcal{T}_0(f_2+f_1))}\right.\\ &+\frac{\sin(\pi Df_1)}{\pi Df_1}\frac{\sin(\pi Df_2)}{\pi Df_2}\left(\frac{\sin\left[\pi(2N+1)\mathcal{T}_0f_2\right]}{\sin(\pi\mathcal{T}_0f_2)}\right)\left(\frac{\sin\left[\pi(2N+1)\mathcal{T}_0f_1\right]}{\sin(\pi\mathcal{T}_0f_1)}\right)\\ &\left.-\frac{\sin(\pi Df_1)}{\pi Df_1}\frac{\sin(\pi Df_2)}{\pi Df_2}\frac{\sin\left[\pi(2N+1)\mathcal{T}_0(f_2+f_1)\right]}{\sin(\pi\mathcal{T}_0(f_2+f_1))}\right). \qquad \text{(B.12)}\end{aligned}$$

### *B.3 Derivation of Spectral Covariances*

The Hermitian spectral covariance is defined as

$$\begin{aligned}R_{XX}(f_1,f_2) &= E\left\{\left[X(f_1)-EX(f_1)\right]\left[X(f_2)-EX(f_2)\right]^*\right\}\\ &= S_{XX}(f_1,f_2)-E\left[X(f_1)\right]E\left[X(f_2)\right]^*. \qquad \text{(B.13)}\end{aligned}$$

and complementary spectral covariance is defined as

$$\begin{aligned}\widetilde{R}_{XX}(f_1,f_2) &= E\left\{\left[X(f_1)-EX(f_1)\right]\left[X(f_2)-EX(f_2)\right]\right\}\\ &= \widetilde{S}_{XX}(f_1,f_2)-E\left[X(f_1)\right]E\left[X(f_2)\right]. \qquad \text{(B.14)}\end{aligned}$$

Thus, we need to find an expression for $E\left[X(f)\right]$. Using expressions (B.6) and (B.9) from above,

$$\begin{aligned}E\left[X(f)\right] &= E\left[H(f)\sum_{n=-N}^{N}e^{-j2\pi f(n\mathcal{T}_0-k_n)}\right]\\ &= H(f)\sum_{n=-N}^{N}e^{-j2\pi fn\mathcal{T}_0}E\left[e^{j2\pi fk_n}\right]\\ &= H(f)\frac{\sin(\pi Df)}{\pi Df}\frac{\sin\left[\pi(2N+1)\mathcal{T}_0f\right]}{\sin(\pi\mathcal{T}_0f)} \qquad \text{(B.15)}\end{aligned}$$

Thus, plugging (B.12) and (B.15) into (B.13), we get

$$\begin{aligned}R_{XX}(f_1,f_2) =& H(f_1)H^*(f_2)\left[\frac{\sin\left(\pi D(f_2-f_1)\right)}{\pi D(f_2-f_1)}\frac{\sin\left[\pi(2N+1)\mathcal{T}_0(f_2-f_1)\right]}{\sin(\pi\mathcal{T}_0(f_2-f_1))}\right.\\ &+\frac{\sin(\pi Df_1)}{\pi Df_1}\frac{\sin(\pi Df_2)}{\pi Df_2}\left(\frac{\sin\left[\pi(2N+1)\mathcal{T}_0f_2\right]}{\sin(\pi\mathcal{T}_0f_2)}\right)\left(\frac{\sin\left[\pi(2N+1)\mathcal{T}_0f_1\right]}{\sin(\pi\mathcal{T}_0f_1)}\right)\\ &\left.-\frac{\sin(\pi Df_1)}{\pi Df_1}\frac{\sin(\pi Df_2)}{\pi Df_2}\frac{\sin\left[\pi(2N+1)\mathcal{T}_0(f_2-f_1)\right]}{\sin(\pi\mathcal{T}_0(f_2-f_1))}\right]\\ &-H(f_1)H^*(f_2)\frac{\sin(\pi Df_1)}{\pi Df_1}\frac{\sin(\pi Df_2)}{\pi Df_2}\left(\frac{\sin\left[\pi(2N+1)\mathcal{T}_0f_2\right]}{\sin(\pi\mathcal{T}_0f_2)}\right)\\ &\left(\frac{\sin\left[\pi(2N+1)\mathcal{T}_0f_1\right]}{\sin(\pi\mathcal{T}_0f_1)}\right)\end{aligned}$$

The 2nd and 4th terms cancel, yielding

$$\begin{aligned}=& H(f_1)H^*(f_2)\left(\frac{\sin\left(\pi D(f_2-f_1)\right)}{\pi D(f_2-f_1)}\frac{\sin\left[\pi(2N+1)\mathcal{T}_0(f_2-f_1)\right]}{\sin(\pi\mathcal{T}_0(f_2-f_1))}\right.\\ &\left.-\frac{\sin(\pi Df_1)}{\pi Df_1}\frac{\sin(\pi Df_2)}{\pi Df_2}\frac{\sin\left[\pi(2N+1)\mathcal{T}_0(f_2-f_1)\right]}{\sin(\pi\mathcal{T}_0(f_2-f_1))}\right)\end{aligned}$$

$$\begin{aligned}\Rightarrow R_{XX}(f_1,f_2) =& H(f_1)H^*(f_2)\left(\frac{\sin\left(\pi D(f_2-f_1)\right)}{\pi D(f_2-f_1)}-\frac{\sin(\pi Df_1)}{\pi Df_1}\frac{\sin(\pi Df_2)}{\pi Df_2}\right)\\ &\frac{\sin\left[\pi(2N+1)\mathcal{T}_0(f_2-f_1)\right]}{\sin(\pi\mathcal{T}_0(f_2-f_1))}\end{aligned}\tag{B.16}$$

The same procedure can be used to find $\widetilde{R}_{XX}(f_1,f_2)$:

$$\begin{aligned}\widetilde{R}_{XX}(f_1,f_2) =& H(f_1)H(f_2)\left(\frac{\sin\left(\pi D(f_2+f_1)\right)}{\pi D(f_2+f_1)}\frac{\sin\left[\pi(2N+1)\mathcal{T}_0(f_2+f_1)\right]}{\sin(\pi\mathcal{T}_0(f_2+f_1))}\right.\\ &\left.-\frac{\sin(\pi Df_1)}{\pi Df_1}\frac{\sin(\pi Df_2)}{\pi Df_2}\frac{\sin\left[\pi(2N+1)\mathcal{T}_0(f_2+f_1)\right]}{\sin(\pi\mathcal{T}_0(f_2+f_1))}\right)\end{aligned}$$

$$\begin{aligned}\Rightarrow \widetilde{R}_{XX}(f_1,f_2) =& H(f_1)H(f_2)\left(\frac{\sin\left(\pi D(f_2+f_1)\right)}{\pi D(f_2+f_1)}-\frac{\sin(\pi Df_1)}{\pi Df_1}\frac{\sin(\pi Df_2)}{\pi Df_2}\right)\\ &\frac{\sin\left[\pi(2N+1)\mathcal{T}_0(f_2+f_1)\right]}{\sin(\pi\mathcal{T}_0(f_2+f_1))}\end{aligned}\tag{B.17}$$

These quantities and their components are visualized in figures B.1, B.2, and B.3.

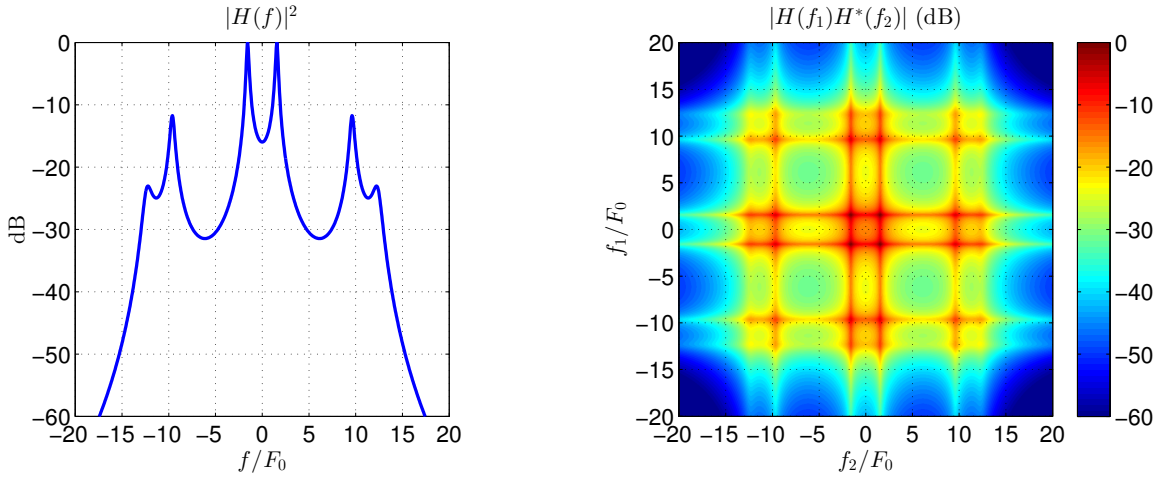


Figure B.1: Left: one-dimensional magnitude response of formant filter within region of interest. Right: two-dimensional magnitude response of formant filter within region of interest.

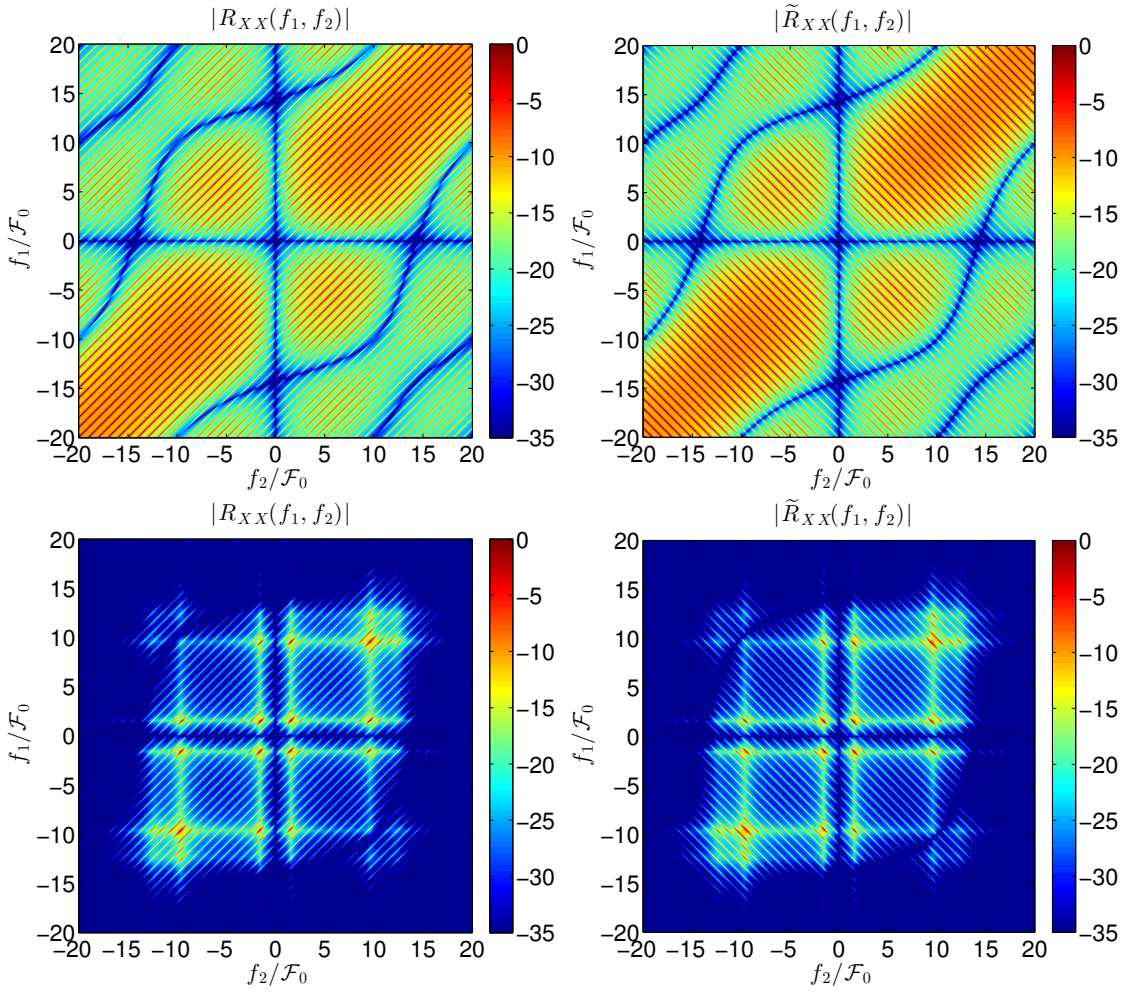


Figure B.2: Bifrequency spectral covariances for F-JPT when fundamental frequency is constant ($f_0(t) = \mathcal{F}_0 = 210$ Hz) and jitter is small (10% of the fundamental period $1/\mathcal{F}_0$). Left panels are Hermitian bifrequency covariances and right panels are complementary bifrequency covariances. Top panels correspond to no formant filter, bottom panels correspond to a typical 3-pole formant filter. Note that application of the formant filter is a simple multiplication of the two-dimensional filter response shown in the right panel of figure B.1.

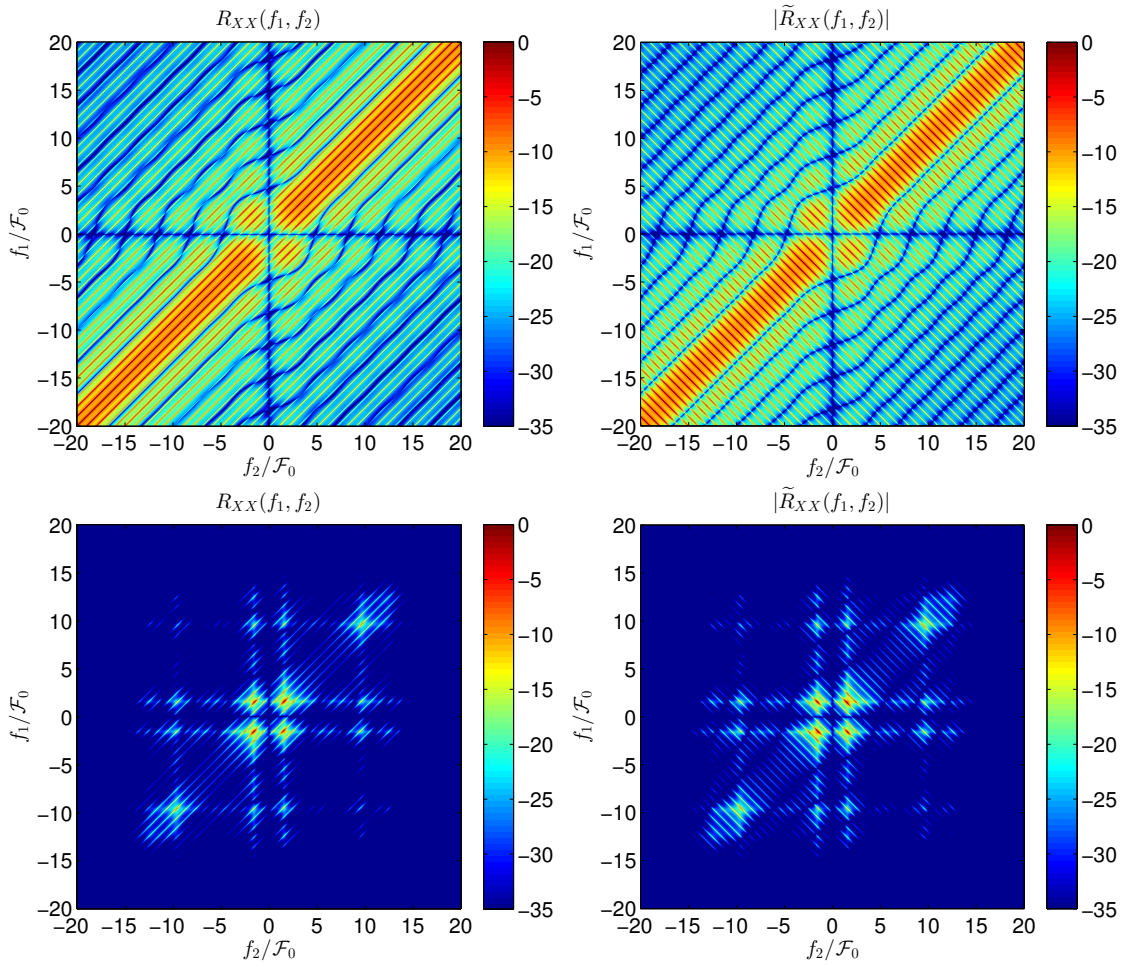


Figure B.3: Bifrequency spectral covariances for F-JPT when fundamental frequency is constant ($f_0(t) = \mathcal{F}_0 = 210$ Hz) and jitter is large (30% of the fundamental period $1/\mathcal{F}_0$). Left panels are Hermitian bifrequency covariances and right panels are complementary bifrequency covariances. Top panels correspond to no formant filter, bottom panels correspond to a typical 3-pole formant filter. Note that application of the formant filter is a simple multiplication of the two-dimensional filter response shown in the right panel of figure B.1.

# VITA

Scott Wisdom grew up in Colorado. An early interest in computers and music led him to emphasize signal processing and embedded systems during his undergraduate career. Today he maintains a wide variety of research interests, including signal processing, machine learning, statistics, optimization, and embedded systems, with a focus on audio and wireless applications. Scott is continuing on as a PhD student in the Department of Electrical Engineering at the University of Washington.