Very Long Baseline Interferometry and the VLBA
ASP Conference Series, Vol. 82, 1995
J. A. Zensus, P. J. Diamond, and P. J. Napier (eds.)

Printed June 13, 1995

CHAPTER 2

Theory of Correlation in VLBI

JONATHAN D. ROMNEY

NATIONAL RADIO ASTRONOMY OBSERVATORY

Abstract This lecture presents the general concepts of correlation and spatial coherence functions, as they apply to VLBI and VLBI correlator implementations. A uniform treatment, based in the spectral domain, is applied to both the conventional lag and the spectral or FX correlator architectures. Parallel presentations of both architectures are given, including their strengths and weaknesses, although more emphasis is placed on the FX correlator because it is less familiar.

Contents

2.1 Introduction

The correlator of a VLBI array is the subsystem in which the interferometers are actually formed. Although a conceptual interference pattern may be considered to exist at the time of observation, it is in the correlator that interference fringes are formed and the complex visibility or spatial coherence function—introduced in chapter 1 as the fundamental observable of VLBI—is measured. Some have referred to the correlator as the "lens" of a VLBI telescope for this reason.

This lecture will address the concepts and theory of VLBI correlation in general. The theoretical treatment, based more in the spectral domain than has been customary, is designed to lead naturally to both the conventional *lag correlator,* and the *spectral-domain* or *FX* architecture. Since the VLBA correlator represents the first implementation in VLBI of an FX correlator, the sections devoted to that topic will necessarily be somewhat specific to the VLBA system. A more detailed presentation on the VLBA correlator, including its particular capabilities, is given in chapter 7.

Modern VLBI correlators operate on digital data, sampled and recorded at a distance in both space and time from the event of correlation. For this reason, I will diverge from a common practice in many previous discussions of correlation theory, and omit the entire area of quantization and digitization of the data, and their impact on calibration. These topics are covered in section 5.7 on page 82 and section 8.4.3 on page 149.

Following the spectral-domain basis of the theoretical treatment, and in another departure from previous conventions, I also will assume throughout that the desired output of the correlator is spectroscopic—i.e., a set of visibility measurements on a closely-spaced grid in radio-frequency space. This is clearly the requisite measurement for investigations of spectral emission or absorption lines; chapter 10 shows that it is also the most appropriate form for constructing VLBI images of continuum sources and chapter 18 shows its usefulness for making high-precision astrometric or geodetic measurements.

2.2 Interfaces

The fundamental inputs to the correlator are the VLBI data tapes recorded at each station participating in the observations. Operating some time after completion of the observations, the correlator reproduces and resynchronizes the sampled signals recorded on the VLBI data tapes, and thus recreates in one room the situation that existed across the world as the observations occurred. At the VLBA correlator, the delay between observation and correlation currently ranges from a few weeks for intra-VLBA observations, to a few months for *global* observations involving non-VLBA telescopes. The range of turn-around times is limited only by the performance of package-delivery services at the minimum, and the lifetime of a particular VLBI data recording system at the maximum.

Besides the VLBI data tapes, important auxiliary inputs to the correlator are the station *logs,* chronologies recording all necessary details of the observations at each participating VLBI station. The correlator control system uses the logs to organize its work and to set its own instrumental parameters to match the observations. Logs which are incomplete or even missing are sometimes synthesized, at least partially, from observing schedules.

The correlator's basic output is a set of measurements of the complex visibility function, spanning a total of four dimensions: the spatial-frequency vector (u, v, w) and the observing frequency ν. In the conventional ordering imposed by the correlator implementation, these results consist of spectra (over ν), for each member of the set of baselines, at a sequence of observing times. The baselines comprise all pairwise combinations of the observing stations, including *zero-baseline,* single-dish measurements; observing time advances at intervals on the order of a second. Each baseline and time corresponds to a different (u, v, w) vector.

Important auxiliary data may accompany the fundamental output. One category of such data records the performance of the correlator system, including the tape playback and the internal computations which reproduce the geometric configuration of the observations. Other auxiliary information, used subsequently for calibration and editing of the results, has been produced or collected at the observing stations, and is simply gathered by the correlator and attached to the associated correlator results. For the VLBA, we have designed a format, based on the *FITS Binary Tables* standard, which includes these auxiliary data with the visibility measurements. This format provides a direct route into NRAO's AIPS image-processing software, as well as other data-reduction systems.

2.3 Background

At this point it's necessary to introduce some fundamental concepts and the formalism to be used throughout the lecture. The following subsections will have to summarize, in a very superficial fashion, some areas which are the subject of entire graduate-level courses, but I have given several references for further reading. And since I will be using mathematics principally as a heuristic tool, I will not pursue rigorous derivations.

2.3.1 Correlation and Coherence

The subjects of correlation and coherence can be approached from the point of view of optics (for which Born and Wolf 1980 is the definitive general reference), of communications engineering (where Blackman and Tukey 1958 is the classic work, although concerned only with what radio astronomers would call the *zero-baseline* case), or of stochastic processes (see, e.g., Parzen 1962). The terminology common in the latter field is closer to that of radio astronomy, and will be used in the following treatment.

We start with a quasi-monochromatic electromagnetic plane wave with an electric field component given by the real part of

$$E(t) = A(t)e^{+i2\pi\nu_0 t}\,, \tag{2.1}$$

where $A(t)$ is a complex, band-limited, covariance-stationary, ergodic stochastic process. $A(t)$ can be considered to modulate a carrier at frequency ν_0. The wave propagates *opposite* to the unit vector pointing at the source, i.e., along $\hat{\mathbf{k}} = -\hat{\mathbf{s}}$. Although $E(t)$ is an *analytic signal* as used in the usual complex-representation formalism, I will focus on $A(t)$, which is complex. Although the analytic-signal formalism thus will not be usable, I will nevertheless work with complex functions because they permit a clearer presentation. Generally, physical signal processes will be the real parts of these functions, and I will point out the few cases where this somewhat careless usage requires correction.

We will be interested principally in the spectrum of $A(t)$, and so we introduce at this point its *spectral representation*

$$A(t) = \int_{-\infty}^{+\infty} s(\nu)e^{+i2\pi\nu t}\,d\nu\,. \tag{2.2}$$

It must be cautioned that $s(\nu)$ is strictly a mathematical device, and in fact has no physical meaning and is not measurable. Mathematicians are reluctant even to give it a name. Only when we reach Equation 2.6–2.7 below do we obtain what is usually thought of as a *spectrum*. Nevertheless, we can still assert that $s(\nu)$ is zero outside a limited band $|\nu| \leq B$.

The interferometer measures what is called variously the cross-covariance, cross-correlation, or visibility function, defined as

$$\Gamma_{ij} = \left\langle E_i(t+\tau_{ij})E_j^*(t)\right\rangle\,, \tag{2.3}$$

where $\tau_{ij} = (\mathbf{r}_i - \mathbf{r}_j)\cdot\hat{\mathbf{k}}/c$ is the light-travel time between observation of the same plane wave at the two spatial positions labeled i and j. The superscript * indicates complex conjugation.

In terms of the band-limited quasi-monochromatic signal $A(t)$,

$$\Gamma_{ij} = \langle A(t+\tau_{ij})A^*(t)\rangle\, e^{+i2\pi\nu_0\tau_{ij}} = \gamma_{ij}(\tau)e^{+i2\pi\nu_0\tau_{ij}}\,, \tag{2.4}$$

where

$$\gamma_{ij}(\tau) = \langle A(t+\tau_{ij})A^*(t)\rangle \tag{2.5}$$

is the covariance function of the stochastic process $A(t)$. The angle brackets in each of Equations 2.3–2.5 denote an expectation or statistical average. Since $A(t)$ has been declared to be ergodic, these can be approximated adequately by averages with respect to t.

Since τ_{ij} is realized by receiving the wave at different locations on the Earth, it will change as diurnal rotation changes the angle between the baseline vector and source unit vector; this gives rise to the *fringe* signal of interferometry which we must measure. Here $\gamma_{ij}(\tau)$ represents the complex amplitude and phase which modulate the fringe signal given by the phasor in τ_{ij}.

Using Equation 2.2 we can develop a spectral representation for $\gamma_{ij}(\tau)$, as

$$\begin{aligned} \gamma_{ij}(\tau) &= \left\langle \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} s_i(\nu) s_j^*(\nu') e^{+i2\pi\left[(\nu-\nu')t+\nu\tau_{ij}\right]} \, d\nu \, d\nu' \right\rangle \\ &= \int_{-\infty}^{+\infty} S_{ij}(\nu) e^{+i2\pi\nu\tau_{ij}} \, d\nu \,, \end{aligned} \tag{2.6}$$

where

$$S_{ij}(\nu) = s_i(\nu) s_j^*(\nu) = \int_{-\infty}^{+\infty} \gamma_{ij}(\tau) e^{-i2\pi\nu\tau} \, d\tau \tag{2.7}$$

can now properly be called the *cross-power spectrum* of the process $A(t)$. A great deal of elegant mathematics has been devised to establish Equations 2.6–2.7 on a rigorous basis, even though the integrals in the Fourier transform relationship formally do not exist.

2.3.2 VLBI Observations

In this section I will outline the major steps that occur at a VLBI observing station on the way to producing the recorded tape which is shipped to the correlator. The quasi-monochromatic plane wave described by Equation 2.1 on the facing page is received at station i, at position $\mathbf{R}_i(t)$, with a delay

$$\tau_i = \mathbf{R}_i(t + \tau_i) \cdot \hat{\mathbf{k}}/c \tag{2.8}$$

relative to the time the wave passes the coordinate origin. Note that τ_i must be defined recursively, in terms of the position of the station at an earlier time when the wave was received there. (As defined here, τ_i is always *negative* for sources above a terrestrial horizon, since $\hat{\mathbf{k}}$ points *away* from the source.) The time dependence of τ_i is not shown explicitly.

The received field

$$E_i(t) = E(t + \tau_i) = A(t + \tau_i) e^{+i2\pi\nu_0(t+\tau_i)} \tag{2.9}$$

is converted to baseband by a single-sideband mixer with LO frequency ν_0:

$$V_i(t) = E_i(t) e^{-i2\pi\nu_0 t} = A(t + \tau_i) e^{+i2\pi\nu_0\tau_i} \,. \tag{2.10}$$

This equation is not physically correct as written, since the LO signal is actually real, and the sum-of-frequencies component is removed by bandpass filtering. Nevertheless, the result is valid. A complete treatment also must account for frequency and phase offsets among the LOs at different stations, which I must gloss over here.

The baseband signal is sampled at an interval Δt, through a non-linear function $f[\cdot]$:

$$R_i(n) = f\left[V_i(t = n\Delta t)\right] . \tag{2.11}$$

Sampling will impose both a loss of sensitivity and a modification of the original spectrum. Techniques for correcting these effects are described in section 8.4.3 on page 149 and section 11.3.4 on page 217. As I remarked at the outset, however, I will take the position here that the correlator's job is to measure the *sampled* signal.

All subsequent manipulations of the signal, through the recording system and into the correlator, operate on discrete, digital samples. However, most people find continuous functions and integrals more intuitive than series and summations, and I will proceed using the continuous forms whenever possible, which in fact will turn out to cover most of the material of this lecture. When necessary, I will resort to the discrete notation.

Finally, the digital samples are recorded on magnetic tape. The recording format includes embedded time tags which allow the time of each individual sample to be established. Again, there are always offsets (which can be kept reasonably small with modern technology) among the atomic time standards at the various stations, which would have to be included in a more complete treatment.

2.4 The Correlator Frontend

At the input to the correlator, the recorded samples on the tapes are reproduced, and the sample times re-established. It's sometimes possible to play back the tapes at speeds several times that used in recording, so that correlation can proceed faster than real time. Any such speedup must carefully be kept in mind in designing the correlator control software.

I have assigned to the correlator's *frontend* two other, essential functions, which compensate for the effects of the array geometry on the signals received. While it's not necessary (nor, is some cases, possible) to perform this compensation at the frontend of the correlator, it is convenient to introduce them conceptually at this point, because they follow naturally from the preceding section. At this point it's appropriate to discuss these functions in a completely station-based manner, although often they are thought of in a baseline sense. Subsequent sections will go into greater detail on their application in the two alternative correlator architectures, and will consider the baseline-based approach where necessary.

Implementation of these functions is an area where significant differences arise among specific correlator designs. Although a hierarchy of linear and/or polynomial approximations is almost always used, the number of coefficients and their update intervals vary widely.

2.4.1 Delay Compensation

Under the simplifying assumptions adopted earlier, compensation for the interferometer delay is extremely straightforward. Each reproduced signal $V_i(t)$ is delayed by δ_i:

$$P_i(t) = V_i(t - \delta_i) = A(t + \tau_i - \delta_i)e^{+i2\pi\nu_0\tau_i} . \tag{2.12}$$

(Remember, the actual signal is a set of discrete samples of $V_i(t)$, $P_i(t)$, etc.). Ideally, one would choose

$$\delta_i = \tau_i \tag{2.13}$$

to obtain the *white light fringe* (see below). Departures from this ideal are also discussed in subsequent sections.

The fact that VLBI data must be recorded makes it possible to implement arbitrarily large delays simply by offsetting the tapes during playback.

In practice, a digital buffer is necessary to accommodate minor variations in tape speed, so that at least a part of the delay can, and inevitably is, tracked by manipulating the buffer pointers.

Sampling before correlation also limits delay tracking at the correlator input to integral units of the sample interval. Techniques which can effect or approximate fractional-sample delay tracking under some circumstances will be discussed later.

2.4.2 Phase Compensation

A phase rotation must also be applied to the signal, in order to stop the fringe signal as described earlier. We denote the phase compensation function as $\theta_i(t)$. VLBI correlators inevitably apply this phase rotation as a *complex* mixer, for reasons explained later. The *real* output of this rotator is then the real part of

$$X_i(t) = P_i(t)e^{-i\theta_i(t)} = A(t + \tau_i - \delta_i)e^{+i2\pi\nu_0\tau_i - i\theta_i(t)} . \tag{2.14}$$

Again the ideal phase rotation will be shown later to be correct:

$$\theta_i = 2\pi\nu_0\tau_i . \tag{2.15}$$

2.5 The Lag Correlator

With this background, we can consider the conventional or lag correlator architecture quite straightforwardly. Generalize Equation 2.13 to provide a range of *lags,* additional delay increments whose spacing, for the observing bandwidth B limiting the quasi-monochromatic process $A(t)$, is given by $\Delta\tau = 1/2B$:

$$\delta_i(n) = \tau_i + n\Delta\tau \qquad n \in [-\tfrac{N}{2}, \tfrac{N}{2}-1] , \tag{2.16}$$

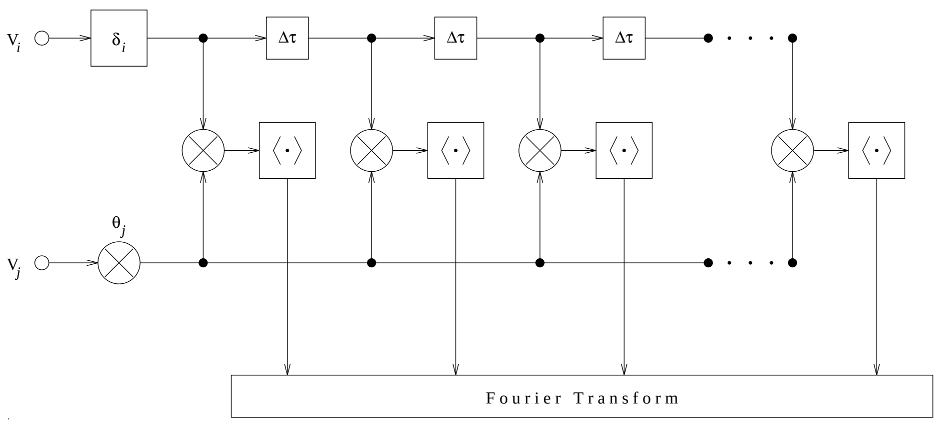

Figure 2.1: Lag correlator architecture

and, using Equations 2.16 and 2.15 in Equation 2.14, define a set of lagged signals

$$L_i(t,n) = A(t - n\Delta\tau)\,. \tag{2.17}$$

Then the correlation obtained by the lag correlator is

$$\begin{aligned} C_{ij}(n) &= \left\langle L_i(t,0)L_j^*(t,n)\right\rangle = \langle A(t)A^*(t-n\Delta\tau)\rangle \\ &= \langle A(t+n\Delta\tau)A^*(t)\rangle = \gamma_{ij}(n\Delta\tau)\,. \end{aligned} \tag{2.18}$$

Comparison with Equation 2.5 shows that we have thus obtained the complex amplitude and phase which modulate the fringe signal in Equation 2.4, and confirms that Equations 2.13 and 2.15 are correct. The N values of the cross-correlation function $C_{ij}(n)$ are now stationary, and can be integrated as long as desired, limited only by the range of residual fringe frequencies to be preserved.

Finally, to obtain the desired cross-power spectrum, it is then only necessary to apply the Fourier transform of Equation 2.7 on page 21. (This is a bit too glib; Equation 2.7 implies $\gamma_{ij}(\tau)$ is sampled over an infinite range of τ, when in fact only N lags are obtained; the effect of this truncation is explored below.)

A typical configuration of the electronic hardware which performs these operations, for a single baseline, is shown in figure 2.1. The box labeled δ_i represents the frontend delay compensation applied to input signal V_i. The individual one-lag delays marked $\Delta\tau$ yield a range of lagged signals. The conventional mixer symbol labeled θ_j applies the phase compensation to a different input signal V_j. The application of these corrections as baseline-based quantities to a single input stream is typical, for reasons explained immediately below. The correlation cells shown as mixer symbols multiply the lagged signals by the unlagged, phase-rotated signals, and an integration to provide the statistical average is shown by $\langle\cdot\rangle$ brackets to complete the implementation of equation 2.18. The final box is the Fourier transform corresponding to equation 2.7.

2.5.1 Fractional-Sample Delay

I pointed out previously that delay compensation at the correlator input necessarily is restricted to integral samples. We are now in a position to see the effects of this, and to consider some solutions. Instead of Equation 2.13, write

$$\delta_i = \tau_i + \varepsilon_i \,, \tag{2.19}$$

where ε_i represents a fractional-sample error due to truncation or rounding of δ_i to an integral multiple of the sample interval. Then in Equation 2.17 we will have

$$L_i(t,n) = A(t - n\Delta\tau - \varepsilon_i)\,, \tag{2.20}$$

and in Equation 2.18,

$$C_{ij}(n) = \gamma_{ij}\left(n\Delta\tau - (\varepsilon_i - \varepsilon_j)\right)\,. \tag{2.21}$$

On applying Equation 2.7, the result of this lag shift will be to change $S(\nu)$ to

$$S'(\nu) = e^{-i2\pi\nu(\varepsilon_i - \varepsilon_j)} S(\nu)\,. \tag{2.22}$$

Thus any fractional-sample error in delay compensation induces a *phase slope* across the observed band. Expressing the error in units of samples,

$$\varepsilon_i - \varepsilon_j = \frac{f}{2B}\,, \tag{2.23}$$

where B is the observing bandwidth,

$$S'(\nu) = e^{-i\pi f\nu/B} S(\nu)\,. \tag{2.24}$$

The effect induces a phase shift of πf at the band edge, and reduces the amplitude of a continuum signal by

$$\frac{1}{B}\int_0^B \cos \pi f \frac{\nu}{B}\, d\nu = \frac{\sin \pi f}{\pi f}\,. \tag{2.25}$$

Even if the error for each station is kept within $[-\frac{1}{2}, +\frac{1}{2}]$ sample by careful rounding, we see that independent, station-based delay compensation can allow combined errors extending to $f \in [-1, +1]$. Since the corresponding band-edge phase shift of $[-\pi, +\pi]$ and possible complete wash-out over a continuum band are unsatisfactory, lag correlators always require a fundamentally baseline-oriented delay compensation, where the rounding can ensure that $f \in [-\frac{1}{2}, +\frac{1}{2}]$. One common way to accomplish this is to use station-based delays with an extra, baseline-dependent *vernier delay* of $(0, \pm 1)$ bit. However implemented, this scheme reduces the band-edge phase shift to $[\frac{-\pi}{2}, \frac{\pm\pi}{2}]$ and the worst-case continuum amplitude loss to $\frac{2}{\pi}$.

A further improvement requires coordination of the delay and phase compensation, inserting an extra phase shift of $\frac{\pi}{2}$ into the phase rotator whenever a one-sample delay shift occurs. Instead of Equation 2.15 we have

$$\theta_i = 2\pi\nu_0\tau_i + \frac{\pi}{2}\frac{\delta_i}{1/2B} = 2\pi\left(\nu_0\tau_i + \frac{B}{2}\delta_i\right), \tag{2.26}$$

and eqs. 2.20–2.22 become

$$L_i(t,n) = A(t - n\Delta\tau - \varepsilon_i)e^{+i\pi B\delta_i}, \tag{2.27}$$

$$C_{ij}(n) = \gamma_{ij}\left(n\Delta\tau - (\varepsilon_i - \varepsilon_j)\right)e^{+i\pi B(\delta_i - \delta_j)}, \tag{2.28}$$

$$\begin{aligned} S''(\nu) &= e^{-i2\pi\left[\nu(\varepsilon_i-\varepsilon_j)+\frac{B}{2}(\delta_i-\delta_j)\right]}S(\nu) \\ &= e^{-i2\pi\left[(\nu-\frac{B}{2})(\varepsilon_i-\varepsilon_j)+\frac{B}{2}(\tau_i-\tau_j)\right]}S(\nu). \end{aligned} \tag{2.29}$$

In Equation 2.26, since τ_i and δ_i are nearly equal, we see that the effect of the extra phase steps is to shift the phase tracking from the edge to the center of the observed band. This is reflected in the second term in the phasor of Equation 2.29; in the first, we obtain the desired change where the phase slope now pivots about $\nu = \frac{B}{2}$. For baseline-based delay tracking with $f \in [-\frac{1}{2}, +\frac{1}{2}]$, the band-edge phase shift is reduced to $\pm\frac{\pi}{2}$ (now at both edges), and the continuum amplitude loss to

$$\frac{1}{B}\int_{-\frac{B}{2}}^{+\frac{B}{2}} \cos \pi f \frac{\nu}{B}\, d\nu = \frac{\sin\frac{\pi f}{2}}{\frac{\pi f}{2}}. \tag{2.30}$$

Finally, it is possible in some circumstances to correct the phase slope of Equation 2.22 directly, after integration. This requires that the phasor $e^{+i2\pi\nu(\varepsilon_i-\varepsilon_j)}$ remain essentially constant during an integration interval. This condition in turn may or may not be realizable, depending on the number of baselines and spectral resolution (which, for a given output data rate limit, determine the integration interval), and on the baseline lengths, observing frequency, and bandwidth (which determine how fast the phase slope varies).

2.5.2 Phase Rotator Implementation

A lag correlator must form a large number of lag products for each of many baselines and so the efficiency of each lag is a critical design factor. This has limited lag correlator designs to at most four-level by four-level products. Thus with two- or four-level signal inputs, only primitive phase-rotation functions can be used. One very popular choice has been the three-level rotator (Clark, Weimer and Weinreb 1972) which passes a signal unchanged for $\frac{3}{8}$ cycle, then blanks the correlator for the next $\frac{1}{8}$ cycle, and then inverts its input for $\frac{3}{8}$ cycle, *etc.*

Since the function is nearly a square wave, it is not surprising that a significant fraction of its power appears at frequencies other than the desired fundamental. Harmonic components will shift the input signals by the wrong frequency; the correlation of these shifted components will not be stationary during an integration and this fraction of the input signal power will be lost. Consequently, lag correlators using this rotator function reduce the amplitude of the output fringe signal by a factor 0.96, which must be corrected in the calibration process.

Use of such a fringe rotation function has a further consequence for lag correlators. It is not possible to do the straightforward station-based fringe rotation implied by Equation 2.15, for chance coincidences of different harmonics of two different fringe rotation functions will produce unexpected spurious correlations. Thus, as was true of the delay, phase rotation must be implemented on a baseline basis in a lag correlator.

2.5.3 Spectral Response

The limited range of lags, N in Equation 2.16, has an important consequence in the profile of spectral lines observed with the VLBI instrument. The effect is fairly obvious, but should be developed here for comparison with the alternative FX correlator architecture, where the corresponding situation is less well known.

We write $N\Delta\tau = T$, and represent the truncation of $\gamma_{ij}(\tau)$ as

$$\gamma'_{ij}(\tau) = \gamma_{ij}(\tau) \cdot \sqcap(\frac{\tau}{T})\,. \tag{2.31}$$

The *unit rectangle function* $\sqcap(x)$, an extremely useful device introduced in Bracewell's classic work (1965) on Fourier transform theory, is unity for $x \in [-\frac{1}{2}, \frac{1}{2}]$ and zero elsewhere. Using $\gamma'_{ij}(\tau)$ in Equation 2.7, we can then apply the convolution theorem of the Fourier transform, and write immediately

$$S'(\nu) = S(\nu) * T\,\mathrm{sinc}(T\nu)\,, \tag{2.32}$$

where the symbol $*$ indicates the convolution operation and the *sinc function* $\mathrm{sinc}(s) = \sin \pi s/\pi s$ is the Fourier transform of $\sqcap(x)$.

2.6 The FX Correlator

At the simple level adopted at the beginning of the preceding section for the lag correlator, the derivation of the spectral-domain or FX correlator is almost too trivial to write down. If only as a framework for the description of the hardware stages, however, I will do so anyway. Starting again with the compensated input signal given by Equation 2.14, for delay and phase compensation specified by Equations 2.13 and 2.15, we have directly

$$X_i(t) = A(t)\,. \tag{2.33}$$

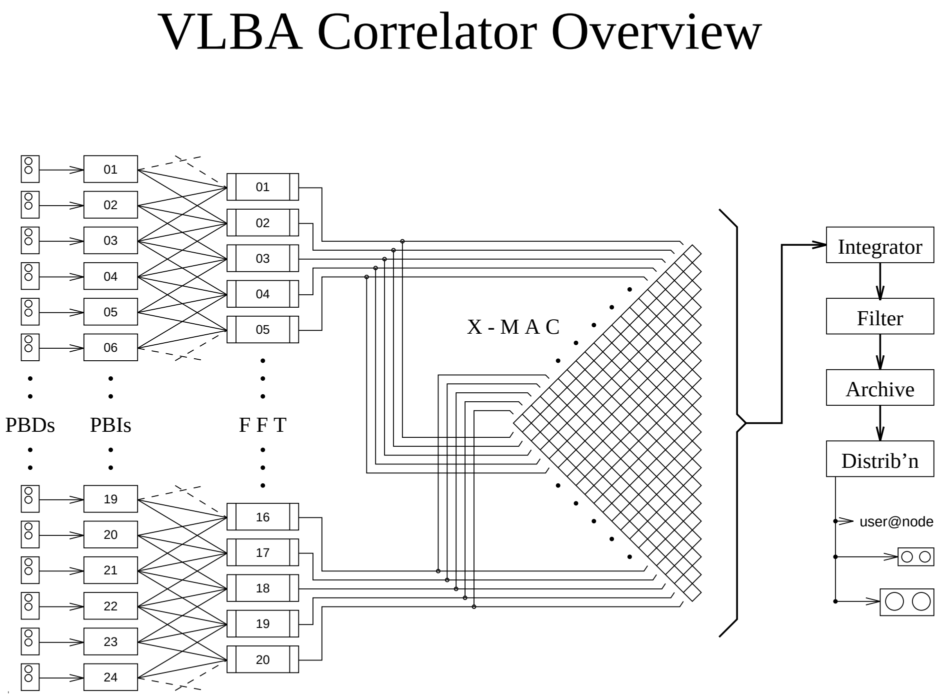

Figure 2.2: FX architecture: the VLBA correlator

The FX correlator immediately Fourier transforms this signal—the "F" of FX—to give

$$s_i(\nu) = \int_{-\infty}^{+\infty} X_i(t) e^{-i2\pi\nu t}\, dt\,. \tag{2.34}$$

The FFT algorithm makes the entire FX concept practical, but is not a fundamental requirement. Next, these *station spectra* are *cross-multiplied*—the "X"—yielding

$$S_{ij}(\nu) = s_i(\nu) s_j^*(\nu)\,. \tag{2.35}$$

The cross-power spectrum in Equation 2.35 can then be integrated. Again, the integration time is limited only by the desired range of residual fringe frequencies.

The name "FX" was originated by Chikada (Chikada et al. 1987), who also built the first such correlator, to indicate this reversal of the order of operations compared to the conventional lag correlator. The VLBA correlator group, while considering one design of each type, adopted the term "XF" which is sometimes used for the latter architecture.

An overview of the VLBA correlator, representing the FX architecture, is shown in figure 2.2. The 24 VLBI tape playback drives (PBDs) appear explicitly;

they and the *playback interfaces* (PBIs) are only of interest here as implementing the integral-sample delay compensation. Phase compensation is performed in the unlabeled small boxes at the inputs to the 20 FFT processors, which themselves perform the "F" operation of the correlator. The small boxes at the FFT outputs will be discussed in the next subsection. In the cross-multiplier/accumulator (X-MAC) a single square corresponds to the "X" operation of the one-baseline correlator described here, and represents both Equation 2.35 and a subsequent partial integration, which in turn is extended to longer times in the Integrator. Other components of the figure are not germane to the present discussion.

2.6.1 Fractional-Sample Delay

The effect of integral-sample delay compensation at the correlator's front end is fundamentally the same for FX and lag correlators, but significantly different in the practical application. Again we use Equation 2.19 to represent the integral-sample compensation with fractional-sample errors, and replace Equation 2.33 with

$$X_i(t) = A(t - \varepsilon_i)\,. \tag{2.36}$$

And similarly to the lag case, the FX correlator output is seen to be modified to

$$S'_{ij}(\nu) = e^{-i2\pi\nu(\varepsilon_i - \varepsilon_j)}S(\nu)\,. \tag{2.37}$$

Although the same phase slope arises in both correlator architectures, in the FX scheme we can correct the slope at the Fourier transform output, before cross-multiplication and before integration. This operation is indicated in figure 2.2 on the facing page by the small unlabeled boxes at the FFT outputs. Applying the correction at this point is equivalent to interpolating the fractional-sample delay at the correlator input. It allows the FX correlator to operate in all cases with no fractional-sample loss at all.

2.6.2 Phase Rotator Implementation

Although the first stage of an FX correlator's Fourier transform could be optimized to accept the one- or two-bit samples reproduced from the recorded tapes, in practical designs all stages are as similar as possible. Thus, the input typically can be presented at the same moderate precision as is needed at the Fourier transform's output. This allows the phase rotation to be done at the input using a multi-level rotator function, at essentially no extra cost.

Signals passed through such a rotator lose no significant power to harmonics. Thus, beyond just retaining the fringe power lost in a three-level rotator, it becomes practical to implement station-based phase rotation without spurious correlations—and without the cost of a large number of baseline-based rotation elements.

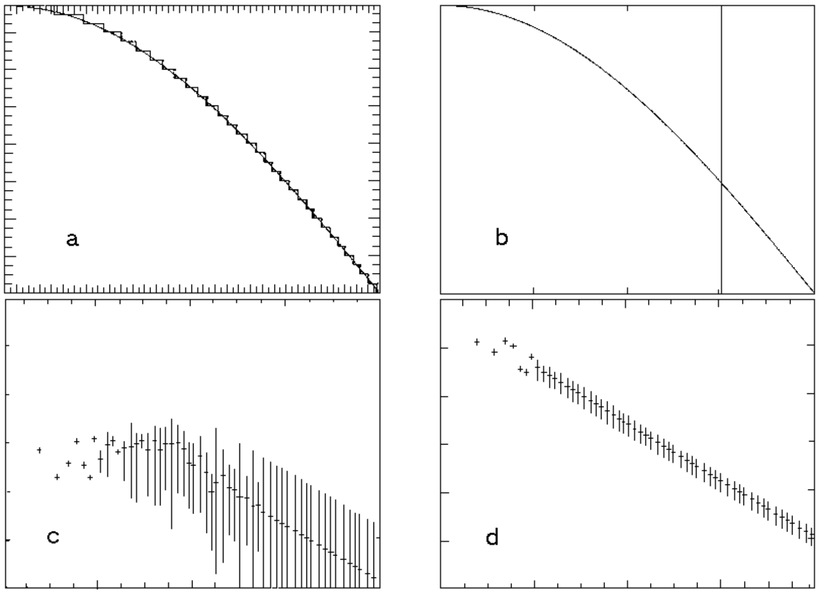

Figure 2.3: Multi-level rotation

Figure 2.3*a* shows the multi-level rotator function used in the VLBA correlator, or actually just the real part of the first quadrant; all the rest is determined by symmetry. For comparison, figure 2.3*b* shows the same plot for the three-level rotator. Figure 2.3*c* and figure 2.3*d* are the corresponding spectra, normalized to the fundamental in the upper left corner, on logarithmic axes. Since the density of harmonics increases exponentially, beyond the first four octaves a range-and-average symbol is used. One can see that the multi-level function adopted for the VLBA correlator suppresses the third and higher harmonics through the first seven octaves, leaving all harmonics at least some 25 dB below the fundamental.

2.6.3 Spectral Response

Reversing the order of F and X operations has an interesting effect on the spectral profile. A practical FX correlator transforms a limited number, N, of samples in its F component—just as the lag correlator can form only a limited number of lags.

Proceeding as for the lag correlator, we write $N\Delta t = T$ (which is now an extent in time, rather than lag), and in place of Equation 2.31,

$$X_i'(t) = X_i(t) \cdot \sqcap\left(\frac{t}{T}\right) . \tag{2.38}$$

Table 2.1: FX correlator response

Correlator:	Lag	FX
Response:	$\mathrm{sinc}(\cdot)$	$\mathrm{sinc}^2(\cdot)$
Full Width at 50%	1.21	0.88
Full Width at 10%	1.82	1.48
First Sidelobe	-0.22	+0.05
Half-Channel Ripple	0.64	0.41

The corresponding convolution in the spectral domain,

$$s'_i(\nu) = s_i(\nu) * T\,\mathrm{sinc}(T\nu)\,, \tag{2.39}$$

is applied to *each* of the station spectra, *before* cross-multiplication. Consequently, the cross-power spectrum is convolved twice by the $\mathrm{sinc}(\cdot)$ function:

$$S'_{ij}(\nu) = S(\nu) * T^2\,\mathrm{sinc}^2(T\nu)\,. \tag{2.40}$$

The $\mathrm{sinc}^2(\cdot)$ spectral response of the FX correlator has both advantages and disadvantages compared to the $\mathrm{sinc}(\cdot)$ response of the lag correlator. It is narrower, and has lower sidelobes, which generally are preferable characteristics for spectroscopy. On the other hand, the narrow profile causes a greater *ripple* as a narrow line is scanned across the discrete spectrum. These characteristics are compared in table 2.1.

2.6.4 Segmentation

In addition to the narrower spectral response, truncation of the input time series has a further effect on the performance of the FX correlator. This can be made evident by applying an inverse Fourier transform to Equation 2.40 to derive the effective cross-correlation function measured by the correlator:

$$\gamma'_{ij}(\tau) = \gamma_{ij}(\tau) \cdot T \wedge(\tau/T)\,. \tag{2.41}$$

The *unit triangle function* $\wedge(x) = \sqcap(x) * \sqcap(x)$ is the self-convolution of $\sqcap(x)$, having unit amplitude at $x = 0$ and tapering at unit positive and negative slope to zero at $x = (-1, 1)$. Note that the width of $\wedge(x)$ is twice that of $\sqcap(x)$, while their integrals are equal.

Equation 2.41 reveals two interesting characteristics of the FX correlator's response in the lag domain. First, transforming data samples in segments of length N yields a range of $2N$ lags covering $\tau \in (-N\Delta t, +N\Delta t)$. More importantly, however, these lags are heavily tapered by the triangle function; there are fewer pairs of samples within the set of N which can form large lags than small.

Put another way, an FX correlator makes the same number of correlation measurements as a lag correlator with N lags, but they are distributed differently. The FX correlator reaches twice the lag range, but makes few measurements at the extremes of the range.

The tapered distribution of independent cross-correlation measurements implies an oppositely-shaped distribution of noise in the measurements. The effect of this noise on a particular observation depends on the spectral characteristics of the source observed. A continuum spectrum has little power beyond the first few lags away from zero anyway, so the triangle tapering has little effect, while a narrow spectral line has significant power even at large lags. Before considering this effect of segmentation more quantitatively, we should consider an enhancement which can be applied when sufficient processing capacity is available: overlapping segments.

Operating with overlap factor f, an FX correlator advances its input data stream by only N/f samples for each N-point transform—requiring f times the processing capacity, of course. Overlapping does not change the triangle taper on the effective cross-correlation, nor the spectral response. It merely generates f times the number of correlation measurements. Some of these measurements are new: at large lags, $|\tau| \geq \frac{f-1}{f}N$, overlapping segments contain pairs of samples not otherwise combined. For smaller $|\tau|$, however, previously measured correlations are just repeated in the overlapping segments.

Thus we define a measurement weight

$$W(\tau) = \begin{cases} 1, & |\tau| \leq \frac{f-1}{f}T \\ f(1 - \frac{|\tau|}{T}), & \frac{f-1}{f}T \leq |\tau| \leq T \\ 0, & |\tau| \geq T\,, \end{cases} \tag{2.42}$$

with $T = N\Delta t$ as previously. This function, a trapezoid with ever steeper ends as f increases (and degenerating to a triangle for $f = 1$), gives the fraction of all possible correlations which are measured at each lag τ. Then, for cross-correlation amplitude given by $C(\tau)$, we can adopt this relative measure of signal-to-noise ratio:

$$R = \frac{\int_{-T}^{+T} W(\tau) C(\tau)\, d\tau}{\int_{-T}^{+T} C(\tau)\, d\tau}\,. \tag{2.43}$$

R is normalized to unity for the non-overlapped case $f = 1$ which yields the same number of correlation measurements as a lag correlator.

Finally, we can evaluate Equation 2.43 in two limiting cases. For a continuum source, $C(0) = C_0$ and vanishes elsewhere, and

$$R_{\text{continuum}} = 1, \qquad \text{for any } f\,. \tag{2.44}$$

And for a monochromatic line, $C(\tau) = C_0$ for all τ;

$$R_{\text{line}} = \int_{-T}^{+T} W(\tau)\, d\tau = \frac{2f-1}{f}\,. \qquad (2.45)$$

2.6.5 Miscellaneous FX Topics

Concluding my discussion of the FX correlator, I'd like to mention a few other aspects in which it differs significantly from the lag correlator. I'll start with a negative feature: it doesn't cope well with episodically invalid data. In a lag correlator, it's possible, at least in principle, to have a data-validity signal accompany each sample, and then inhibit correlation, independently at each lag, when either sample is invalid. In practice, this requires enough additional signal paths and normalization counters that it may never be done—but in an FX correlator, it can't be done at all. Any practical FX design must use the FFT algorithm, which requires equi-spaced data samples. There is no way to omit an invalid sample; the best that can be done is to count the errors for each FFT data segment, and invalidate the entire FFT output if the count exceeds some threshold. For typical errors in reproducing the recorded signals, this is quite satisfactory, in fact, since the errors often occur in bursts long enough to cover many FFTs.

Next, I'll turn to two features which can only be implemented in an FX architecture. An input time-domain window can be applied at the same point as the phase rotation. Smoothing in the spectral domain, as commonly done in post-processing, typically applies only functions with a few points close to the peak, to minimize the computational load. With the input time-domain window, however, any real function can be used.

A pulsar gate can be applied at the FFT output, in the spectral domain. At that point, all delay and phase compensation has been performed, and the signal resolved into narrow spectral bins. Under those circumstances, an array-wide pulsar gate function will depend only on frequency and pulsar phase. Such a pulsar gate thus can perform a form of de-dispersion impossible in a lag correlator, where the time-domain samples span an entire baseband channel.

Finally, we come to the dangerous *hybrid modes.* It's possible to configure an FX correlator so as to transform, say, four contiguous baseband channels at relatively low spectral resolution, and cross-multiply the spectra against that from a single channel of four times the bandwidth, resolved into four times as many spectral bins. The individual bins from both sides thus match up, with the same frequency spacing. Such modes are useful in dealing with incompatibilities of disparate signal-processing and recording systems. I've labeled them "dangerous" because, once people understand the concept, they expect any conceivable version to be achievable. In practice, each such case must be planned as part of the hardware design effort.

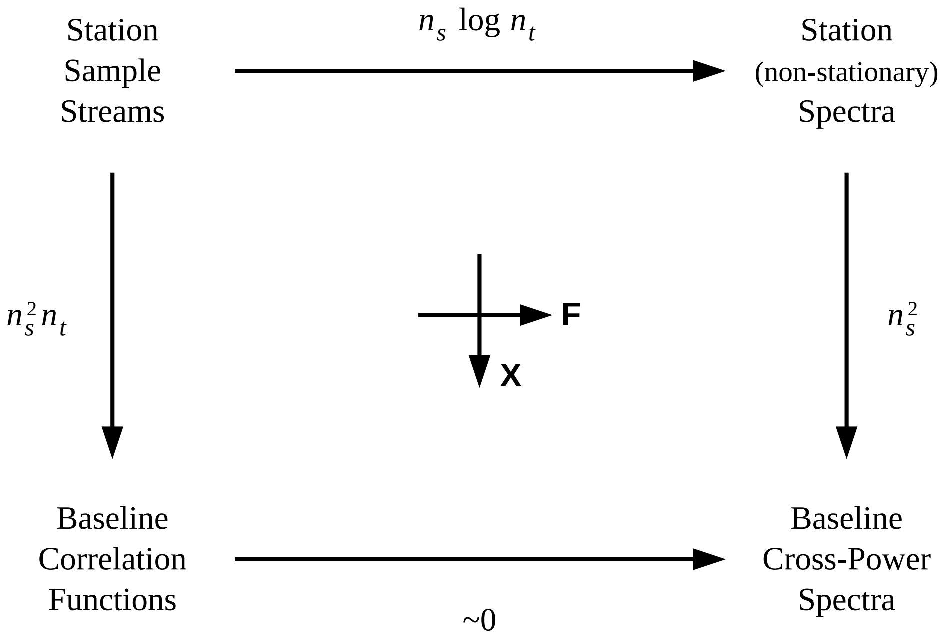

Figure 2.4: VLBA correlator data domains

2.7 FX and Lag Correlator Intercomparison

This final section will compare the two alternative VLBI correlator architectures from several quite different points of view. I'll begin with some very general considerations, then turn to a very particular basis for comparison—cost, and conclude by summarizing the other advantages and disadvantages of each.

Figure 2.4 illustrates the four basic data domains involved in VLBI correlation, and presents the lag and FX architectures graphically as two different routes from the common input, station sample streams, to the common result, baseline cross-power spectra. Two general conclusions should be apparent from this figure, as well as from the parallel derivations of the two correlators I presented earlier. First, the lag and FX correlators represent opposite sides of the convolution theorem of the Fourier transform: the transform of the convolution (correlation) equals the product of the transforms. And second, the FX architecture minimizes the number of operations required to perform the task, by exploiting the efficient FFT algorithm, and by organizing as much processing as possible on a station basis.

The figure shows, in terms of the system dimensions—n_s stations, and n_t samples transformed or lags formed—the operations needed along each of the four paths, and leads naturally into a more detailed discussion of the relative costs of the two alternatives.

2.7.1 Cost

During the VLBA correlator group's early consideration of the FX architecture, two different approaches to cost comparison were developed, based on simple operation counts (Romney 1986), and on estimates of the required number of logic gates (Romney 1987). Both results are summarized here, for a correlator supporting n_s station inputs, each with n_c baseband channels (of sample rate r_0), and with n_t samples per segment transformed or lag correlations accumulated. The computations are based on $n_s^2/2$ baselines, including the real, single-dish *self-spectra.*

The simplest approach to counting operations is just to count multiplication operations. It can be shown that to a very good approximation, one addition accompanies each multiplication in both cases.

The lag correlator forms n_t lags per baseline and channel, each at rate r_0. As described earlier, phase rotation can only be performed on one input signal, so that each lag has one complex and one real input, requiring 2 multiplies. Thus the aggregate multiplier rate for cross-correlation is

$$r_L = r_0 n_s^2 n_c n_t \,. \tag{2.46}$$

The contribution of the subsequent FFT operation can be neglected, since these lags can be accumulated for some time beforehand, so r_L is then the total multiply rate for the lag correlator system

For simplicity, assume the FX correlator's FFTs are implemented using a straightforward radix-2 Cooley-Tukey algorithm. Then the transform consists of $\frac{1}{2} n_t \log_2 n_t$ *butterfly* stages, of 4 multiplies each, per station and channel. Transforms are executed at a rate r_0/n_t, so the aggregate multiplier rate in the FX correlator is

$$r_F = 2 r_0 n_s n_c log_2 n_t \,. \tag{2.47}$$

The cross-multiplier must process $n_t/2$ points of each station/channel spectrum (because only one side of the spectrum contains signal power), at the same rate r_0/n_t. Since the spectra are complex, 4 multiplies are required per point. The aggregate cross-multiplication rate is then

$$r_X = r_0 n_s^2 n_c \,. \tag{2.48}$$

The ratio of multiplies for the two architectures,

$$R_{\frac{Lag}{FX}} = \frac{r_L}{r_F + r_X} = \frac{n_s n_t}{2 log_2 n_t + n_s} \,, \tag{2.49}$$

shows that the FX architecture requires significantly fewer operations when *either* of n_s or n_t is large.

The gate-count approach was developed to circumvent one very serious oversimplification of the operation counts: the multiplies are of different complexities, with the lag correlator requiring only a few bits. Some serious preliminary work is required before a gate-count comparison can be attempted. Fairly thorough designs under both architectures, using the same generation of micro-electronic technology (preferably as current as possible), must be developed.

From these one can extract the basic gate-count units: g_L gates per lag per baseline, g_F gates per FFT stage, and g_X gates per baseline cross-multiplied. It's important to ensure that the operations included in these basic gate-count are as functionally comparable as possible.

Then the lag correlator's total gate count (using the same system parameters as previously) is

$$G_L = n_c \frac{n_s^2}{2} n_t \cdot g_L \,, \tag{2.50}$$

and the corresponding total gate count for the FX correlator

$$G_{FX} = n_c n_s \lceil \log_4 n_t \rceil \cdot g_F + n_c \frac{n_s^2}{2} \cdot g_X \,. \tag{2.51}$$

Note that a radix-4-plus-2 FFT implementation is now assumed; the logarithm is to base 4, and the integer ceiling operation $\lceil \cdot \rceil$ accounts for the case where a radix-2 stage is required for n_t an odd power of 2. I will not quote the numerical results from the VLBA correlator's study, done in 1987, because they are completely obsolete by this time. The comparison showed, however, that an FX correlator would require fewer total gates, for any number of stations, for even quite modest spectral resolution.

2.7.2 Summary of Other Advantages

The FX architecture allows a number of desirable features to be implemented which are not possible in a lag correlator. Some, such as truly station-based phase rotation and exact fractional-sample delay compensation, eliminate sensitivity losses as well as permitting more efficient implementation of essential functions. Others make unique, specialized observations possible (the spectral pulsar gate), or reduce the need for post-correlation processing (the time-domain window, and indeed the station-based phase rotation belongs in this category too). Its $\text{sinc}^2(\cdot)$ spectral response may be superior, although it must be said there is as yet no experience with this. Although subject to segmentation loss, overlapped processing allows this to be restored, and to better sensitivity than an equivalent lag correlator.

The lag correlator is more tolerant of substandard reproduction of the recorded VLBI signals. And it actually obtains the cross-correlation function, which may be useful for certain cases of the correction for the non-linear sampling imposed on the observed signals.

References

Blackman, R. B., & Tukey, J. W. 1958. *The Measurement of Power Spectra*. New York: Dover.

Born, M., & Wolf, E. 1980. *Principles of Optics* (Sixth ed.). Oxford: Pergamon Press.

Bracewell, R. N. 1965. *The Fourier Transform and its Applications*. New York: McGraw-Hill.

Chikada, Y., Ishiguro, M., Hirabayashi, H., M., Morimoto, K., Morita, T., Kanzawa, H., Iwashita, Nakazima, K., Ishikawa, S., Takahashi, T., Handa, K., Kasuga, T., Okumura, S., Miyazawa, T., Nakazuru, T., Miura, K., & Nagasawa, S. 1987. "A 6x320 MHz 1024-channel FFT cross-spectrum analyzer for radio astronomy". *IEEE Proc.* **75**, 1203.

Clark, B. G., Weimer, R., & Weinreb, S. 1972. "The Mark II VLB system". Internal Report 118, National Radio Astronomy Observatory.

Parzen, E. 1962. *Stochastic Processes*. San Francisco: Holden-Day.

Romney, J. D. 1986. "Introduction to the special-domain (FX) correlator". VLBA Correlator Memorandum 60, National Radio Astronomy Observatory.

Romney, J. D. 1987. "Spectral-vs. lag-domain corrleation: Comparison of fundamental hardware requirements". VLBA Correlator Memorandum 80, National Radio Astronomy Observatory.