TECHNICAL RESEARCH REPORT

Robust Spectro-Temporal Reverse Correlation for the Auditory System: Optimizing Stimulus Design

by D.J. Klein, D.A. Depireux, J.Z. Simon, S.A. Shamma

**Center for Auditory
and Acoustic Research**

Robust spectro-temporal reverse correlation for the auditory system: Optimizing stimulus design

D. J. Klein   D. A. Depireux   J. Z. Simon
S. A. Shamma*
Institute for Systems Research
University of Maryland
College Park, MD 20742
Tel: 1-301-405-6842
E-mail: sas@isr.umd.edu
FAX: 1-301-314-9920

Abstract

The *spectro-temporal receptive field* (STRF) is a functional descriptor of the linear processing of time-varying acoustic spectra by the auditory system. By cross-correlating sustained neuronal activity with the 'dynamic spectrum' of a spectro-temporally rich stimulus ensemble, one obtains an estimate of the STRF. In this paper, the relationship between the spectro-temporal structure of any given stimulus and the quality of the STRF estimate is explored, and exploited. Invoking the Fourier theorem, arbitrary dynamic spectra are described as sums of basic sinusoidal components, i.e., 'moving ripples.' Accurate estimation is found to be especially reliant on the prominence of components whose spectral and temporal characteristics are of relevance to the auditory locus under study, and is sensitive to the phase relationships between components with identical temporal signatures. These and other observations have guided the development and use of stimuli with deterministic dynamic spectra composed of the superposition of many 'temporally orthogonal' moving ripples having a restricted, relevant range of spectral scales and temporal rates. The method, termed *sum-of-ripples*, is similar in spirit to the 'white-noise approach,' but enjoys the same practical advantages — which equate to faster and more accurate estimation — attributable to the time-domain sum-of-sinusoids method previously employed in vision research. Application of the method is exemplified with both modeled data and experimental data from ferret primary auditory cortex (AI).

Key Words: reverse correlation, moving ripples, sum-of-sinusoids, spectro-temporal, receptive field, auditory cortex

---

*To whom correspondence should be addressed

1 Introduction

Two primary identifying features of a sound are its spectral content, and its temporal behavior. In the physiological investigation of the auditory system, it has often been assumed, by the sole use of descriptors such as spectral tuning and modulation rate tuning, that each of these qualities are processed independently. However, it is becoming more widely recognized that this is not, in general, a well-advised assumption. As evidenced by most any time-frequency representation (Cohen, 1995) (e.g., the spectrogram), the particular time-dependency of the spectrum, i.e., the *dynamic spectrum*,[1] seems to set a sound's character. Thus, it seems likely that the particular conjunction of a sound's spectral and temporal features, and not simply their separate existence, is ultimately of interest to a hearing system.

Such information is readily available to the mammalian auditory system, from the very transduction process. There, the cochlea transduces the impinging sound wave into a frequency-ordered (tonotopic) pattern of activity on the auditory nerve (AN) (Shamma, 1985; Ruggero, 1992). Following the example of the cochlea, and moving to a domain where the spectral and temporal aspects of sounds are represented jointly — the *spectro-temporal domain* — one expects to be able to more effectively characterize those acoustic patterns that afferent neurons are most responsive to. The resulting description of the auditory system's input-to-output transformation — the *spectro-temporal receptive field* (STRF) — is one of both dimensions intertwined, and is potentially more complete than that provided by the two marginal descriptions (Aertsen and Johannesma, 1980; Aertsen and Johannesma, 1981b; Eggermont et al., 1981; Hermes et al., 1981; Cohen, 1995).

STRF-like descriptions of auditory processing have been used in a number of studies. All are endowed with a common, linear functionality, of the general form

$$r(t) = \int \int STRF(\tau, f) \cdot S(t - \tau, f) \, d\tau \, df. \tag{1}$$

The interpretation of (1) is fairly straight-forward. At any particular instant t, a neuron's response r is given by the correlation of the STRF with stimulus' dynamic spectrum S. As the spectrum continues to evolve, the STRF acts as a filter, producing the strongest responses to spectro-temporal features that most resemble its own structure. In doing so, the STRF can be thought of both as a time-dependent spectral weighting function (a.k.a. *receptive field*) and as a frequency-dependent dynamical filter. Figure 1 should assist in the visualization of these concepts.

Though the functionality of the STRF is generally agreed upon, it has been measured with a variety of methods. These differ in the type of stimuli used (e.g., white noise (Hermes et al., 1981; Eggermont et al., 1983b; Epping and Eggermont, 1985; Eggermont and Smith, 1990; Backoff and Clopton, 1991; Clopton and Backoff, 1991; Kim and Young, 1994; Nelken et al., 1997; Carney and Friedman, 1998), natural vocalizations (Aertsen and Johannesma, 1981a; Yeshurun et al., 1985; Schafer et al., 1992; Theunissen et al., 1998), moving ripples (Kowalski et al., 1996a; Kowalski et al., 1996b; Depireux et al., 1998b; Escabí et al., 1998), tone pulses (Aertsen and Johannesma, 1981a; Epping and Eggermont, 1985; deCharms et al., 1998; Kvale et al., 1998; Theunissen et al., 1998)), the representation of the dynamic spectrum (e.g., Wigner

---

[1] The term 'dynamic spectrum' is to be used throughout in a general sense. It is meant to subsume all specific time-frequency representations.
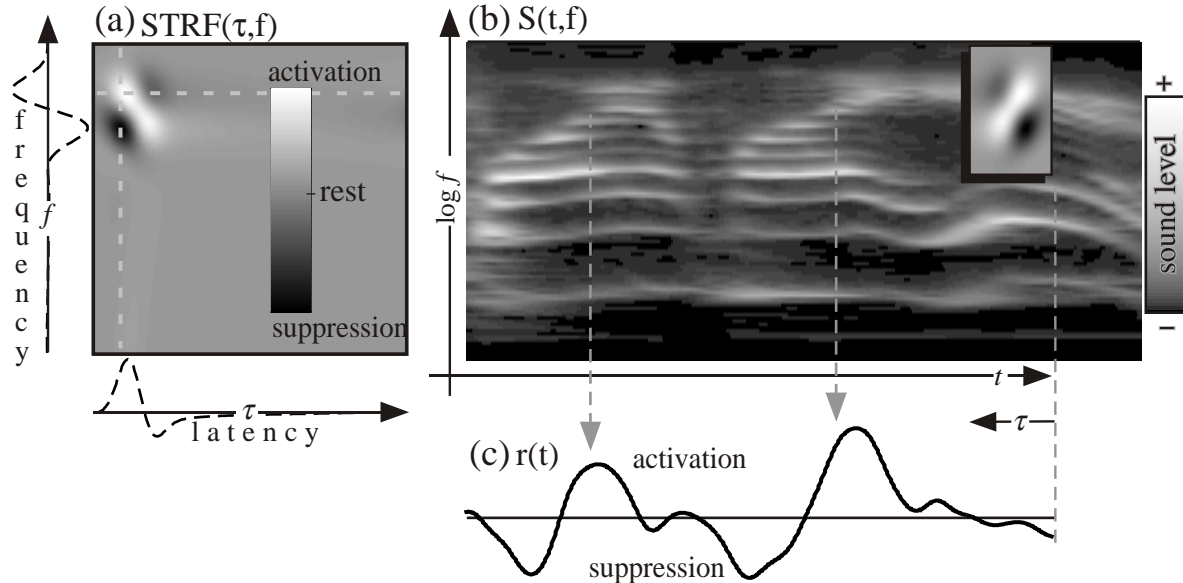
Figure 1: The STRF (a) reflects many aspects of auditory system function — such as spectral tuning, latency, memory, refractoriness, frequency-modulation direction selectivity, and modulation rate tuning — in one compact form. A cross-section of the STRF at a particular latency and frequency yields a characterization of the spectral and dynamical processing of the system which, in general, depends on the latency and frequency chosen. The specific interdependence of these features conspires to produce enhanced responsiveness to certain spectro-temporal patterns, as exemplified in (b) and (c). The stimulus in (b) is the dynamic spectrum of a speech segment ("Why am I here?") produced with a cochlea-like filter bank and plotted on a log-frequency scale. The flipped, overlaid STRF depicts the correlation operation performed at each time t to produce the response in (c), as given by Eq. (1). Characterizing this neuron with a single peak excitatory frequency would not be sufficient; pronounced activation (arrows) requires the presence of specific spectro-temporal patterns resembling the (flipped) STRF.

distribution (Eggermont and Smith, 1990; Kim and Young, 1994; Nelken et al., 1997), Rihaczek distribution (Hermes et al., 1981; Epping and Eggermont, 1985; Eggermont and Smith, 1990; Backoff and Clopton, 1991; Clopton and Backoff, 1991), short-time Fourier transform, (Yeshurun et al., 1985; Schafer et al., 1992), filter bank output (Aertsen and Johannesma, 1981a; Eggermont et al., 1983b; Carney and Friedman, 1998), spectro-temporal envelope (Kowalski et al., 1996a; Kowalski et al., 1996b; deCharms et al., 1998; Depireux et al., 1998b; Escabí et al., 1998; Kvale et al., 1998)), and the analysis method (e.g., reverse correlation (Aertsen and Johannesma, 1981a; Hermes et al., 1981; Eggermont et al., 1983b; Epping and Eggermont, 1985; Eggermont and Smith, 1990; Backoff and Clopton, 1991; Clopton and Backoff, 1991; Schafer et al., 1992; Kim and Young, 1994; Nelken et al., 1997; Carney and Friedman, 1998; deCharms et al., 1998; Escabí et al., 1998; Kvale et al., 1998; Theunissen et al., 1998), Laguerre polynomial correlation (Yeshurun et al., 1985), sinusoidal steady-state analysis (Kowalski et al., 1996a; Kowalski et al., 1996b; Depireux et al., 1998b)).

A majority of STRF measurements have been made by stimulating with Gaussian white noise (GWN) and performing a kind of input-output correlation called *spectro-temporal reverse correlation.* The method is similar to 'classical' reverse correlation (de Boer, 1967; de Boer and de Jongh, 1978), with which the portions of a stimulus waveform preceding the occurrence of a neuron's action potentials are averaged. With spectro-temporal reverse correlation, rather, a

representation of the stimulus' dynamic spectrum is averaged instead, as illustrated in Figure 2. The motivation in both cases is typically of a stochastic nature — to preserve only those stimulus patterns consistently causing a neuron to spike while eventually averaging out other, randomly occurring patterns.
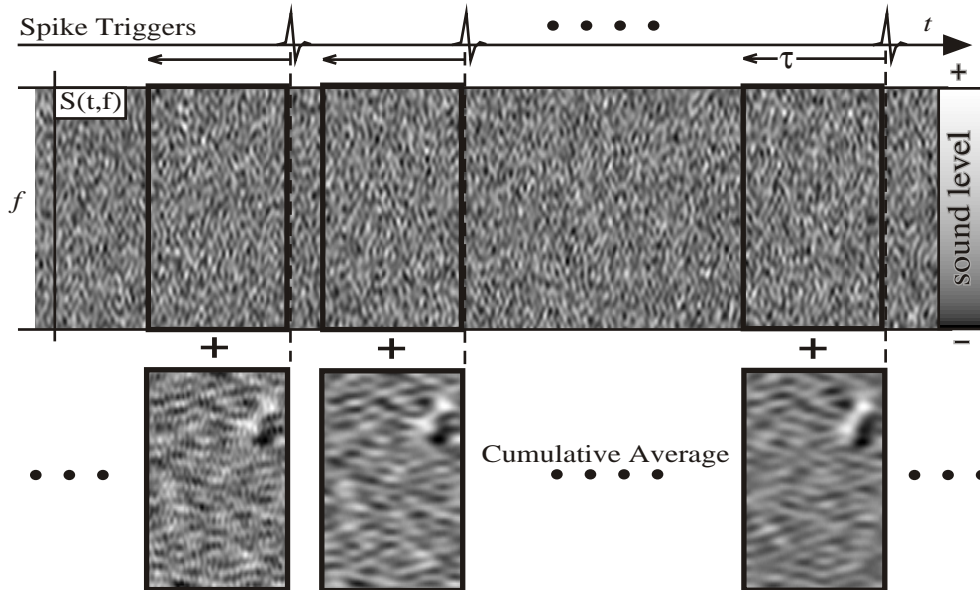

Figure 2: For spiking systems, spectro-temporal reverse correlation can be viewed as a spike-triggered average. Using white-noise stimulation, the average stimulus preceding a spike resembles the STRF after a sufficient number of spikes are recorded.

This general methodology, through which the functionality of a neuron is explored by correlating its responses with various functionals of a GWN stimulus, is referred to as the *white-noise approach.* The white-noise approach has intuitive appeal, but it is far from *ad hoc*; in fact, it is closely related to the *cross-correlation method* (Lee and Schetzen, 1965) for developing a Wiener-series model of a system. The Wiener series, and the closely related Volterra series, have been used extensively to model auditory (and other sensory) system function, particularly in its peripheral aspects (Marmarelis and Marmarelis, 1978; Eggermont, 1993). These studies, and the scrutiny they've received, provide valuable background for studies of the STRF.

Despite the promise of the representation, it is fair to say that the ability of STRF estimates gleaned under white-noise stimulation to quantitatively describe the auditory system has been limited. The estimates are often excessively noisy, computationally burdensome, and above all unable to predict responses to novel stimuli. Paradoxically, the difficulties are more severe for more central auditory areas, where the STRF description is presumed to be most valuable. In fact, to our knowledge there are no accounts of a successful application of the white-noise method in any auditory locus higher than the mid-brain, and in mammals no higher than the lateral-superior olive (LSO). In the face of this, J. Eggermont speculated in his 1993 review of the subject (Eggermont, 1993), "...it is expected that Wiener or Volterra-like characterization methods could largely fail for central auditory areas such as auditory cortex."

There is growing evidence, however, that this forecast may have been premature. Actually, Wiener and Volterra-*like* methods have in time proved to be successful; but *like* is the operative

suffix, for the extreme generality of GWN has had to be compromised in favor of alternative stimuli. These 'improved' stimuli differ from GWN in that they are defined by their dynamic spectra, which is often designed with the guidance of the known functionality of the auditory area under study. While still spectro-temporally rich, the stimuli have a more specific structure than that of GWN, they are more effective in eliciting responses, and they have yielded striking results from mammalian primary auditory cortex (AI) and inferior colliculus (IC) (Kowalski et al., 1996a; Kowalski et al., 1996b; Depireux et al., 1998b; Escabí et al., 1998; Kvale et al., 1998; deCharms et al., 1998; Shamma et al., 1998). Although thorough quantitative analyses of much of the data has not yet surfaced, some of these studies have yielded STRF estimates which are well capable of predicting neurons' responses to novel stimuli (Kowalski et al., 1996b; Depireux et al., 1998b; Shamma et al., 1998).

The approach specifically advocated and developed here, termed *sum-of-ripples*, might be thought of as an extension of the time-domain sum-of-sinusoids method (Victor and Knight, 1979; Victor and Shapley, 1980) into the spectro-temporal domain. This approach uses *moving ripples* (Kowalski et al., 1996a; Depireux et al., 1998b), broad-band sounds that are modulated sinusoidally both in spectrum and in time, as basic stimulus building blocks. Exploiting the linearity of the STRF functional (1), and invoking the formalism of two-dimensional Fourier analysis, the use of single moving ripples (Kowalski et al., 1996a; Kowalski et al., 1996b; Depireux et al., 1998b) and white noise defines the opposite ends of a continuum of possible methods. As such, the Fourier-series description of dynamic spectra, and of the STRF, allows for a general reevaluation of the spectro-temporal reverse correlation method itself, through which basic structural conditions that arbitrary stimuli must meet, in order to allow for an accurate STRF estimate, are easily obtained.

These conditions are ultimately used to guide the synthesis of stimuli, by summing together specific 'temporally orthogonal' combinations of ripples, that are both general in their exploratory power and tailored to a particular auditory locus. It is shown that cross-correlating the neural response with these 'ripple combinations' can quickly build an accurate estimate of the STRF. This has been found to be true both in principle, and in practice; the method has been applied in ferret AI and IC (Shamma et al., 1998; Depireux et al., 1998a). Some of the AI results will be considered here for illustrative purposes. However, an exhaustive analysis of the physiology will be treated elsewhere.

The organization of this paper is as follows. First, the STRF is further examined both empirically and theoretically, as it has evolved, with the aid of the Volterra and Wiener functional expansions. In this context, some challenges posed by the measurement of the STRF are discussed. In section 3, the Fourier transforms of arbitrary dynamic spectra and STRFs are defined and elaborated. This allows the problem of spectro-temporal reverse correlation to be considered within a most general framework. This subsequently motivates the sum-of-ripples construction, explained in section 4. Finally, the results are summarized, and some additional views concerning the link between the sum-of-ripples method and other methods are offered.

2 Background

2.1 The functional expansions of Volterra and Wiener

The use of functional expansions for non-linear systems modeling is essentially due to the pioneering work of Volterra (Volterra, 1930). As summarized by M. Korenberg (Korenberg and Hunter, 1996), "A functional, F, transforms values of the input defined over the input domain (e.g., time) into a value of the output defined at a single point of the output domain (e.g., at a fixed instant)." Typically, then, the distinction between functions and functionals echoes that, for example, between systems without and with memory.

For a large class of systems, it is possible to expand the functional relationship, F, between an input, s, and an output, r, into a sum of n elementary functionals:

$$r = F\left[s\right] = K_0\left[s\right] + K_1\left[s\right] + \cdots + K_n\left[s\right] \tag{2}$$

where n, the *order* of the system, is conceivably infinite. Deciding on the form of the K_i's involves a fundamental compromise between notions of separate functionality, favored in theoretical studies, and, on the other hand, separately measurable properties, necessitated by the experimental approach. This compromise is typified by the relationship between the Volterra and Wiener functional expansions.

The Volterra series expansion (Volterra, 1930; Korenberg and Hunter, 1996) prescribes for the K_i's homogeneous polynomial functionals of order i. In this aspect it is analogous to the Taylor series expansion of functions. If, for example, the input is merely a function of time, the Volterra functionals take the form of progressively higher-order temporal convolutions with progressively higher-order auto-products of the input process,

$$K_i\left[s(t)\right] = \int \cdots \int v_i(\tau_1, \cdots, \tau_i) \cdot s(t - \tau_1) \cdots s(t - \tau_i)\, d\tau_1 \cdots d\tau_i. \tag{3}$$

so that each term in the Volterra series describes, through the *Volterra kernels* $v_i(\tau_1 \cdots \tau_i)$ (which are weighting functions analogous to the Taylor series coefficients), how the output of the system at any particular instant depends on a particular order of the input.

The first-order Volterra functional is the most familiar,

$$K_1\left[s(t)\right] = \int v_1(\tau) \cdot s(t - \tau)\, d\tau \tag{4}$$

as it is the standard time-domain description of a linear time-invariant system with impulse response $v_1(\tau)$. The higher-order functionals follow from a straight-forward generalization of this equation.

There are, however, practical difficulties in the direct measurement of the Volterra kernels, especially for a system of unknown order (as is the auditory system). Fortunately, these difficulties can be largely circumvented if the K_i's are designed to be mutually orthogonal with respect to a particular input function, in which case they, unlike the Volterra functionals, can each be measured separately.

N. Wiener (Wiener, 1958) detailed a particularly useful series of functionals that are closely related to Volterra's but which are orthogonal with respect to a GWN input. As a by-product, the 'Wiener functionals' are inhomogeneous, i.e., they no longer fully describe the response to

a single order i of the input process but also include contributions from all higher orders of the same parity ($i + 2$, $i + 4$, $\cdots$), and furthermore, they can depend on power level of the input (Eggermont, 1993; Korenberg and Hunter, 1996). Thus, the Wiener and Volterra kernels are generally different. Holistically, though, the two descriptions are equivalent, and the Wiener series can be converted to the stimulus-invariant Volterra series, provided that all of the Wiener kernels have been identified. This is often done, because the Volterra series is in general more amenable to interpretation (Aertsen and Johannesma, 1981b; Boyd et al., 1983).

It is important to note the conditions by which this type of characterization is valid. According to Wiener (Wiener, 1958), "...we are considering non-linear networks of a certain deadbeat character." The characterization of the brain as "deadbeat" may be alarming, but in formalizing Wiener's slang one finds that the auditory system actually fulfills many, though not strictly all, of the requirements. For example, the response must only depend on a finite extent of the input domain, or at least exhibit fading dependence (Korenberg and Hunter, 1996). This seems to be met by the current knowledge of the auditory system. However, the characteristics of the system must not change with time, a condition not so obviously met, since it excludes adaptive processing. Also, due to practical concerns, the system is required to have a sufficiently low order expansion, i.e., it must satisfy a 'continuity requirement,' so that small changes at its input result in small changes at its output. While this requirement is not strictly met by a 'spiking' system, it can be satisfied by substituting the smoother spike probability, or the spike rate, as the response (Johnson, 1980).

2.2   Kernel estimation by reverse correlation

Two decades after the Wiener series was developed, Y. Lee and M. Schetzen (Lee and Schetzen, 1965) published a simple and influential algorithm, now known as the cross-correlation method, for estimating the first- and higher-order Wiener kernels of a system by applying GWN to the input and computing a series of first- and higher-order input-output cross-correlations. One of the many merits of their work was that it allowed for a solid relation to be drawn between the Wiener series and the experimental practice of reverse correlation (de Boer, 1967) being pioneered contemporaneously in physiological studies of cat auditory nerve fibers.

It was subsequently recognized that the 'reverse-correlation function' is basically identical to the first-order cross-correlation function and, hence, the first-order Wiener kernel of the system (Eggermont et al., 1983c). Therefore, the reverse-correlation function does not in general solely reflect linear processing, but instead should be considered the best (in a mean-square error sense) linear fit to the observed input-output transformation (Palm and Popel, 1985; Eggermont, 1993). In any case, the reverse-correlation function has been useful in describing cochlear transduction at middle to low acoustic frequencies (de Boer and de Jongh, 1978; Carney and Yin, 1988).

However, if the response of a neuron is not precisely synchronized, i.e., *phase locked*, to fine details of the stimulus waveform, a first-order stimulus-response cross-correlation is not productive, i.e., the first-order Wiener kernel is negligible (Eggermont, 1993). In the mammalian AN, phase locking is limited to neurons tuned to frequencies below about 4-6 kHz (Ruggero, 1992; Kim and Young, 1994). This limit is progressively lower for higher auditory loci, and by AI phase locking to broad-band noise is rarely observable. Thus, for a large fraction of neurons, particularly in more central areas, the classical reverse correlation function has proved to be

useless (Hermes et al., 1981; Clopton and Backoff, 1991; Kim and Young, 1994), and models of their behavior have had to be shifted to higher-order functionals of the stimulus. Fortunately, the reverse correlation methodology can be extended for this purpose; now, various higher-order functionals of the stimulus are to be correlated with the response.

2.3   The role of the STRF in a functional expansion

In particular, the second-order Volterra-Wiener functional,

$$
\begin{aligned}
K_2\left[s(t)\right] &= \int \int v_2(\tau_1, \tau_2) \cdot s(t - \tau_1) s(t - \tau_2) \, d\tau_1 \, d\tau_2 \\
&= \int \int \hat{v}_2(\tau, \sigma) \cdot s(t - \tau - \frac{\sigma}{2}) s(t - \tau + \frac{\sigma}{2}) \, d\tau \, d\sigma
\end{aligned}
\tag{5}
$$

here rewritten with $\tau = \frac{1}{2}(\tau_1 + \tau_2)$ and $\sigma = \tau_2 - \tau_1$ for convenience, has proven to be a primary descriptor of auditory neurons that do not phase lock (Eggermont, 1993; Temchin et al., 1995; Yamada et al., 1997; van Dijk et al., 1997; Yamada and Lewis, 1999). Fortunately, it has an interpretation by which its superseding importance in higher auditory loci is intuitive.

Like K_1, K_2 essentially describes a linear system; but instead of the raw stimulus waveform, its input receives the time-dependent (deterministic) auto-correlation of the stimulus, $s(t - \frac{\sigma}{2}) s(t + \frac{\sigma}{2})$. Although the auto-correlation may be difficult to interpret, it is closely related, through a single Fourier transform, to a large class of time-frequency representations via the *Wigner distribution* (Eggermont, 1993; Cohen, 1995),

$$
W(t, f) = \int s^*(t - \frac{\sigma}{2}) s(t + \frac{\sigma}{2}) \cdot exp(-j 2\pi\sigma f) \, d\sigma,
\tag{6}
$$

where * denotes complex conjugation (which could be omitted since all stimuli considered here are real) and $j = \sqrt{-1}$. The Wigner distribution may be thought of as a generalized spectrogram. Such representations of dynamic spectra are strongly reminiscent of the activity at the output of the peripheral auditory system (Shamma, 1985). Thus, K_2 is expected to have a special applicability for describing the processing being performed in the afferent auditory pathway (Aertsen and Johannesma, 1981b; Hermes et al., 1981).

Indeed, using (6), and defining

$$
STRF^{K_2}(\tau, f) \triangleq \int \hat{v}_2(\tau, \sigma) \cdot exp(-j 2\pi\sigma f) \, d\sigma,
\tag{7}
$$

K_2 can be rewritten:

$$
K_2\left[s(t)\right] = \int \int STRF^{K_2}(\tau, f) \cdot W(t - \tau, f) \, d\tau \, df.
\tag{8}
$$

where it is noted that, for a real valued stimulus, both $STRF^{K_2}$ and W real and are symmetric around $f = 0$.

Therefore, the second-order Volterra functional corresponds to a linear system that processes the Wigner time-frequency representation of the stimulus, with an $STRF^{K_2}$ that is the Fourier transform (across $\tau_2 - \tau_1$) of the second-order Volterra kernel. This provides the link between K_2 and the generalized STRF functional (1); since other time-frequency representations S can

be considered linearly filtered versions of the Wigner distribution (Cohen, 1995), it can be shown (see Appendix A) that STRF^{K_2} is, in general, a filtered version of the STRF, and *vice versa*.

Given the relative importance of K_2 over K_1 in central loci, it is tempting to disregard the stimulus waveform completely and instead treat the dynamic spectrum as the effective stimulus (Eggermont et al., 1983b; Yeshurun et al., 1985; Nelken et al., 1997)), though this is only strictly valid if *all* odd-order functionals, with respect to the waveform, can also be neglected (Aertsen and Johannesma, 1981b). Characteristically, the dynamic spectrum is sectioned into multiple, tonotopically arrayed inputs, representing energy fluctuations within discrete frequency bands (Eggermont et al., 1983b; Yeshurun et al., 1985; Schafer et al., 1992). The STRF, instead of being associated with the second-order functional, is then readily associated with the *first-order* Volterra functional of a multiple-input system, composed of the concatenation of the system's (linear) impulse responses to each of these inputs.

It is important to remember that the spectro-temporal reverse correlation (with an appropriate time-frequency representation) yields the single Fourier transform of the second-order Wiener kernel (Eggermont et al., 1983c; Eggermont, 1993), which is generally different from the Volterra kernel. Nevertheless, because the highest two Volterra and Wiener functionals are always identical, this distinction is not relevant as long as the employed system description doesn't extend, in parity, more than one order beyond the STRF. As such, the STRF is more conveniently treated in the Volterra sense (Aertsen and Johannesma, 1981b); there exists for every neuron a stimulus-invariant STRF, and any attempt to measure it results in an STRF *estimate* which may contain various errors due the particular stimuli used (e.g., white noise) and the measurement technique (e.g., cross-correlation).

2.4   Some problems with white-noise stimulation

The white-noise approach to STRF estimation initially delivered promising results; after sufficient spike-triggered averaging, there were consistent indications of specific regions in the spectro-temporal domain of elevated or diminished intensity preceding the occurrence of spikes (Hermes et al., 1981; Eggermont et al., 1983b). Unfortunately, after some additional use of the method, a number of consistent problems were also apparent. In addition to troubles with noisiness (Hermes et al., 1981; Eggermont and Smith, 1990; Backoff and Clopton, 1991; Kim and Young, 1994) and weighty computational requirements (Eggermont et al., 1983c; Eggermont and Smith, 1990; Clopton and Backoff, 1991; Kim and Young, 1994), it was reported that STRF estimates were only in weak agreement with measurements made with other, more established methods (e.g., tones) (Backoff and Clopton, 1991; Clopton and Backoff, 1991; Kim and Young, 1994). Most importantly, the few attempts to use the measured STRFs to predict neurons' responses to stimuli substantially different from GWN failed (Eggermont et al., 1983a; Nelken et al., 1997). Consequently, it has been concluded that the STRF holds only limited, qualitative value (Eggermont et al., 1983a; Eggermont, 1993; Nelken et al., 1997).

However, it seems likely that some of the trouble was not with the STRF *per se*, but instead stemmed from the choice of stimulus, or, more specifically, the finite-length pseudo-random noise substituted for the physically unrealizable GWN. The primary problems with this stimulus are three-fold, concerning its statistical inadequacies, its weak response-driving capability, and its difference from natural sounds.

The cross-correlation method strictly requires GWN due to its statistical properties, but finite-length noise sequences can deviate considerably from this white ideal, resulting in considerable estimation error, especially for higher-order kernels (Swerup, 1978; Eggermont, 1993). Efforts are often made to improve some of the statistical properties of the waveform (e.g., using 'inverse-repeat' stimuli (Swerup, 1978), or the 'transformation method' (Eggermont, 1993)). However, the statistical adequacy of the *dynamic spectrum*, paramount for accurate estimation of the STRF, is not assured by these efforts (Eggermont et al., 1983b).

Furthermore, in more central loci, copious estimation error is brought about by the weak responses evoked by stationary broad-band noise. Fulfilling the 'continuity requirement,' described above, necessitates using some 'average response' measure, such as the peri-stimulus time histogram (PSTH) (Johnson, 1980). However, weak responses (e.g., low spike probabilities) are overshadowed by the considerable variance associated with the PSTH estimate (Johnson, 1980), not to mention the extraneous variability associated with biological neural systems themselves (Azouz and Gray, 1999). In overcoming this, either a prolonged stimulus must be used, which greatly increases both the measurement and analysis duration, or many repetitions of a shorter stimulus must be used, in which case its statistical properties are eroded further (Eggermont et al., 1983b; Clopton and Backoff, 1991).

Finally, the reverse-correlation estimate is only strictly valid with respect to the particular stimulus used for its derivation; even in the mild presence of measurement noise and system non-linearities, as the 'distance' between a test stimulus and the stimulus used to identify the system increases, the error in a truncated functional series characterization can be expected to become substantial (Johnson, 1980; Palm and Popel, 1985). Again, this is particularly problematic for the most central areas, because the increasingly specific spectro-temporal patterns that neurons are responsive to are increasingly improbably generated by stationary random noise. Thus, the inability of STRF estimates gleaned under GWN stimulation to generalize to the 'distant' natural stimuli is understandable, and suggests that the use of stimuli with natural properties may be more productive for central loci (Palm and Popel, 1985; Yeshurun et al., 1985; Nelken et al., 1997).

2.5   Alternatives to white noise

Due to the problems commonly associated with the white-noise approach, a great deal of work has been devoted towards developing improved stimuli and analyses for Volterra and Wiener kernel estimation. For example, if there is little control over the stimuli, alternatives to cross-correlation, such as Korenberg's fast exact orthogonalization method (Korenberg and Hunter, 1996) and Marmarelis' improved Laguerre polynomial expansion method (Marmarelis, 1993), have been developed to allow for the 'best' possible kernel estimates for arbitrary stimulation. Such improvements generally come, however, at the expense of computational complexity (Korenberg and Hunter, 1996).

If control over the stimuli is afforded, it seems to be more beneficial to focus on improving its structure. Using elementary knowledge about the system, the extreme generality of GWN can be greatly reduced, resulting in stimulation of greater over-all relevance (Sutter, 1992). A simple example, taken from classical reverse correlation, is the reduction of the stimulus bandwidth to match the expected input bandwidth of the system. Such improvements are accomplished most efficiently with deterministic, rather than stochastic, stimuli, e.g., binary

sequences (Sutter, 1992) and sums of sinusoids (Victor and Knight, 1979; Victor and Shapley, 1980) (for whom, functional expansions well approximating the Wiener and Volterra series have been formulated (Victor, 1991; Sutter, 1992)). Consequently, much of the work can be put in prior to an experiment, allowing for accurate results to be obtained *en route*, with simple and 'ultra-fast' correlation-based algorithms (Victor, 1979; Sutter, 1992).

In order to improve the quality of STRF estimation beyond that which is possible with the white-noise approach, such considerations should be applied not to the waveform but to the dynamic spectra of stimuli. However, without a general enough description of dynamic spectra and its relationship to the reverse-correlation estimate, it is not clear how the modifications are best made. In the next section, a quite general formulation of the problem is elaborated, taking advantage of the linearity of the STRF functional, and invoking the Fourier theorem.

3 Fourier analysis in the spectro-temporal domain

In the previous sections, the spectro-temporal domain was established as a plausible input domain for central auditory neurons, both from an intuitive and from a rigorous standpoint. Some problems associated with the use of white-noise stimulation for the measurement of the input-output relationships of such neurons were discussed. Finally, it was recognized that the implementation of improvements wants for a universal descriptive framework for dynamic spectra. That is the aim of this section.

Regardless of the specific nature of the stimulus, the common goal in the measurement of the STRF is to characterize the *linear* processing of dynamic spectra. Evaluation of the STRF functional may exploit the principle of superposition obeyed by all linear systems; namely, the response to any stimulus is the sum of the responses to its constituent parts. One may thus use the STRF to form an alternative 'transfer function' description, in which the responses to a basic set of stimuli are made explicit. If this basic set can, in combination, be used to describe all stimuli of practical interest, then determining the system's response simply involves determining the stimulus' composition in terms of these basic parts.

A common means of uniquely breaking a stimulus into parts is provided by the Fourier series (Papoulis, 1962), with which *any* stimulus can be approximated to any level of precision with a sum of sinusoidal components of various amplitudes, frequencies, and phases. Performing the Fourier decomposition of dynamic spectra thus engenders a particularly valuable 'ripple transfer function' and, subsequently, allows the form of the stimulus-response cross-correlation function to be derived for arbitrary stimulation.

3.1 The spectro-temporal domain

The spectro-temporal domain is the input domain of the STRF functional (1). It is also the axes on which dynamic spectra, and the STRF, are to be defined. In accordance with typical notions of auditory system function, we will consider a logarithmic frequency axis,

$$x = log_2 \frac{f}{f_0},$$

where x denotes the number of octaves above $f_0 > 0$, the lowest frequency considered to be relevant to the system.[2] The x axis is thought of as a spatial axis, corresponding to the auditory sensory epithelium, i.e., the tonotopic axis (Pickles, 1988).

A Volterra-type model restricts the dependence of the system to a limited extent of the input domain. In this case, the system is expected to have finite memory and finite frequency-tuning bandwidth, within $0 \leq x \leq X, \quad 0 \leq t \leq T$ where X is the stimulus bandwidth (with units of octaves) and T is the stimulus duration. In other words, it is expected that the STRF is zero outside of these bounds. Since all stimuli are, for all practical purposes, of finite duration and finite bandwidth, we are always considering a finite region of the spectro-temporal domain.

3.2   Ripples and ripple decomposion

Perhaps the most radical (and pertinent) departure from the white-noise approach for STRF estimation is that of *dynamic ripple analysis* (Kowalski et al., 1996a; Kowalski et al., 1996b; Depireux et al., 1998b), with which STRFs of neurons in ferret AI were measured via stimulation with broad-band sounds having dynamic spectra of the general form

$$S(t, x) = a \cos \left\{ 2\pi (wt + \Omega x) + \psi \right\}. \tag{9}$$

These functions describe for each frequency location x, a sinusoidal modulation of the level around some mean (left unspecified above), at a rate of $|w|$ cycles per second. The relative phases of the modulations at different x's produce a sinusoidal spectral profile with a periodicity of Ω cycles per octave (c/o), which, over time, drifts across the spectral axis, with a velocity determined by the magnitude of w, and a direction determined by the polarity of the product of w and Ω. These sounds are nick-named *moving ripples*. The parameter w is sometimes called the *ripple velocity* or *rate* and Ω is the *ripple frequency*, *ripple peak density*, or *spectral scale*. The dynamic spectra of a several moving ripples, produced with different combinations of w and Ω, are illustrated in Figure 3.

Moving ripples are useful because they form the basis for the Fourier decomposition of the spectro-temporal domain, i.e., over the finite extent $t \in [0, T]$ and $x \in [0, X]$, the real function S is completely and uniquely specified by the Fourier series (Papoulis, 1962):

$$S(t, x) = \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} a_{k,l} \exp \left\{ j[2\pi(w_k t + \Omega_l x) + \psi_{k,l}] \right\}. \tag{10}$$

where

$$w_k = \frac{k}{T}, \quad \Omega_l = \frac{l}{X}. \tag{11}$$

The terms in this sum come in 'complex-conjugate' pairs, such that $a_{k,l} = a_{-k,-l}$, $w_k = -w_{-k}$, $\Omega_l = -\Omega_{-l}$, and $\psi_{k,l} = -\psi_{-k,-l}$. Given that $\cos(\alpha) = \frac{1}{2} \left[\exp(j\alpha) + \exp(-j\alpha) \right]$, each such pair corresponds to a single moving ripple (9). For convenience, the 'DC' amplitude and phase, $a_{0,0}$ and $\psi_{0,0}$, corresponding to the mean sound level, are set to zero (without loss of generality) for the subsequent development.

Since the discrete set of velocities and densities (w_k, Ω_l) of the stimulus components used in (10) are fixed, dynamic spectra are uniquely described by the amplitudes $a_{k,l}$ and phases $\psi_{k,l}$

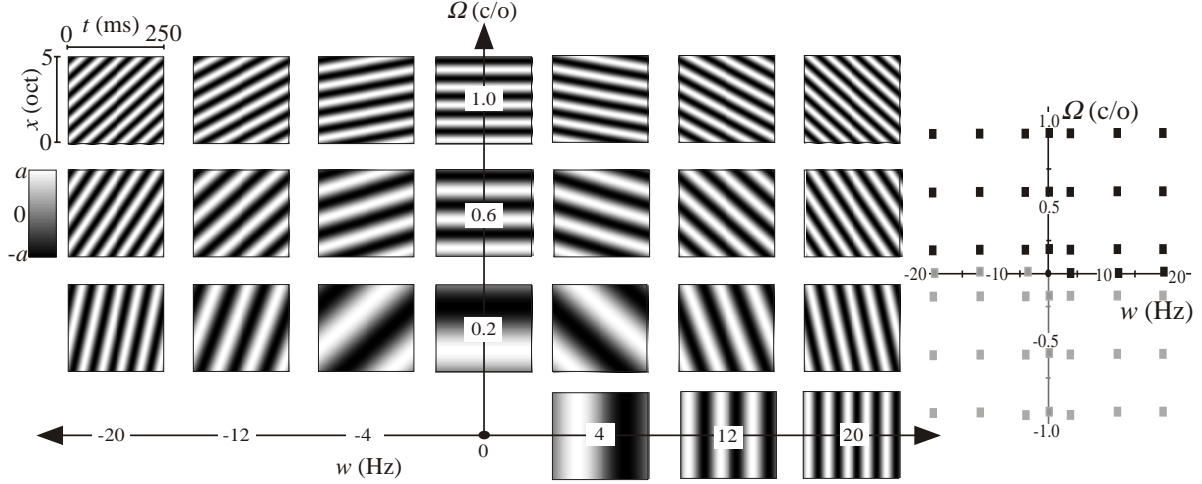[2]Negative frequencies are ignored here, since they provide no additional information.

Figure 3: The dynamic spectra of various moving ripples are illustrated (left), along with their locations in the ripple domain (right). The points in the upper two quadrants (in black) of the ripple domain are sufficient for characterizing a ripple; points in the lower two quadrants (in gray) are given by the complex conjugates of those in the upper two. Quadrant one, where the product of w and Ω is positive, corresponds to 'downward-moving' ripples. Quadrant two, where it is negative, corresponds to 'upward-moving' ripples.

of these components. As such, they form a representation equivalent to S, and can, in fact, be computed from S via the double Fourier transform,

$$\widetilde{S}(w,\Omega) = \frac{1}{TX}\int_0^T\int_0^X S(t,x)\cdot\exp\left\{-j2\pi(wt+\Omega x)\right\}\,dx\,dt$$
$$= \sum_k\sum_l a_{k,l}\exp\left\{j\cdot\psi_{k,l}\right\}\cdot\delta(w-w_k,\Omega-\Omega_l), \tag{12}$$

which produces the complex function $\widetilde{S}$ referred to as the *ripple spectrum* of the stimulus. Here, $\delta(\cdot,\cdot)=1$ when its arguments are zero, and otherwise equals zero. Thus, $\widetilde{S}$ is only nonzero at the points $\widetilde{S}(w_k,\Omega_l)=a_{k,l}\exp\left\{j\cdot\psi_{k,l}\right\}$. Apparently, $a_{k,l}$ and $\psi_{k,l}$ correspond to the magnitude and phase of the ripple spectrum, respectively, at these points. The (w,Ω) axes on which they are displayed are called the *ripple domain*. These two complementary views of dynamic spectra, in the spectro-temporal and ripple domains, is exemplified in Figure 4(a).

A useful descriptor of S is its *total power* P:

$$P \triangleq \frac{1}{TX}\int_0^T\int_0^X S^2(t,x)\,dx\,dt$$
$$= \sum_k\sum_l (a_{k,l})^2. \tag{13}$$

Intuitively, dynamic spectra with higher total power (or just 'power,' for brevity) spend more time further away from the mean level. As can be verified using (10) in the definition, P corresponds to the sum of the squares of the ripple-component amplitudes.
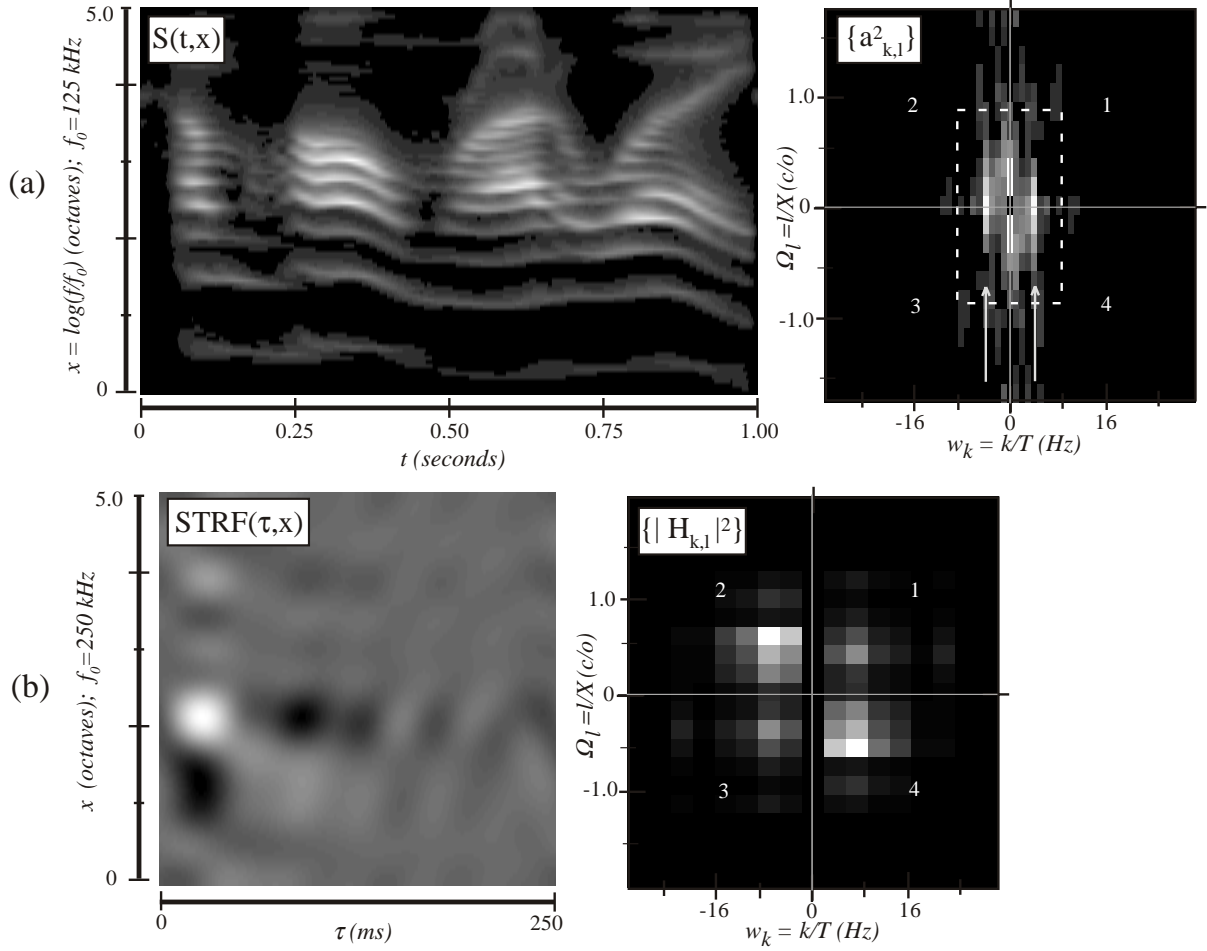
Figure 4: (a) The dynamic spectrum S (left) of a speech segment ("Come home right away.") produced by a cochlea-like filter bank, and the magnitude squared of its ripple spectrum (right), obtained from the double Fourier transform of S, are shown. Note (arrows) the prominent 4 Hz peaks in the ripple spectrum, corresponding to the speech tempo over this one-second interval. A majority of the power of the ripple spectrum (about 65%) is restricted to low ripple densities (< 1 c/o) and low modulation rates (< 8 Hz) indicated by the dashed box. (b) An STRF estimate obtained (using the sum-of-ripples method) from ferret AI, and the amplitude squared of the corresponding ripple transfer function are shown. See the text for a description.

3.3 The ripple transfer function

It is well known that a linear time-invariant single-input single-output system is fully characterized in the time-domain by its impulse response (the first-order Volterra kernel; see (4)), and equivalently in the frequency domain by its transfer function, i.e., its frequency response. Similarly, the STRF, which in (1) basically acts as an impulse response (Depireux et al., 1998b), can also be described, in the ripple domain, by a *ripple transfer function*, H:

$$
\begin{aligned}
H(w, \Omega) &\triangleq \frac{1}{TX} \int_0^T \int_0^X STRF(\tau, x) \cdot \exp\left\{-j2\pi(w\tau - \Omega x)\right\} \, dx \, d\tau \\
&= \sum_k \sum_l H_{k,l} \cdot \delta(w - w_k, \Omega - \Omega_l),
\end{aligned}
\tag{14}
$$

where H has been defined (for reasons to follow) as the ripple spectrum of the STRF with the spectral axis flipped. Following (12), the magnitude and phase of $H(w_k, \Omega_l) = H_{k,l}$ then yields the unique set of ripple component amplitudes $b_{k,-l}$ and phases $\theta_{k,-l}$, that describe the flipped STRF, i.e.,

$$H_{k,l} \triangleq b_{k,-l} \exp\{j \cdot \theta_{k,-l}\}, \tag{15}$$

where, akin to (10),

$$\begin{aligned} STRF(\tau, x) &= \sum_k \sum_l b_{k,l} \exp\{j[2\pi(w_k\tau + \Omega_l x) + \theta_{k,l}]\} \\ &= \sum_k \sum_l H_{k,-l} \exp\{j2\pi(w_k\tau + \Omega_l x)\}. \end{aligned} \tag{16}$$

An example of an STRF measured from ferret AI and the squared magnitude of the corresponding ripple transfer function are illustrated in Figure 4(b).

As defined, H can also be called the 'ripple response' because it details the system's responses to individual moving ripples. This is shown by inserting (10) and (16) into the STRF functional (1), to obtain a general form for the response:

$$r(t) = \sum_k \sum_l H_{k,l} a_{k,l} \exp\{j(2\pi w_k t + \psi_{k,l})\} \tag{17}$$

Thus, the response to *any* stimulus consists of the sum of the responses to each of the individual stimulus ripple components (w_k, Ω_l). The response to each component is sinusoidal, with a frequency w_k, and an amplitude scaled and phase shifted, relative to the stimulus, by the magnitude and phase of $H(w_k, \Omega_l)$.

Besides describing the structure of the STRF, H provides a useful complementary view, mediated by the properties of the Fourier transform, of the functionality of the STRF; common neuronal descriptors such as excitatory and inhibitory tuning, spectral and modulation tuning bandwidth, latency, and memory can all be derived from a the ripple transfer function (Kowalski et al., 1996a; Depireux et al., 1998b). Moreover, seemingly complex features of the STRF can translate to simple functionality as described by H. For example, the transfer function of Figure 4(b) is band-pass in w, because this neuron was only responsive to a certain range of temporal modulation frequencies. This is also reflected (and perhaps determined) by the temporal pattern of excitatory and suppressive influences on the neuron, evidenced by the STRF. Furthermore, the slanted orientation of the STRF indicates that this neuron responded most strongly to rising frequencies. This is corroborated by the the magnitude of H, which is strongest in the second and fourth quadrants, corresponding to 'upward-moving' ripples.

3.4 Dynamic ripple analysis

Because the STRF can easily determined from H by (16), and H details a neuron's responses to moving ripples, the STRF can be estimated by presenting various moving ripple stimuli, one at a time, and measuring the amplitude and phase of the responses (at the appropriate frequencies w_k). This procedure is very much akin to conventional sinusoidal linear-system analysis; if the right combination of sinusoids are presented to the system, the important features of the

ripple transfer function, and hence the STRF, can be characterized. Subsequently, responses to arbitrary dynamic spectra can be predicted via (17). This is the crux of the dynamic ripple analysis method (Kowalski et al., 1996a; Kowalski et al., 1996b; Depireux et al., 1998b).

An important advantage of this method is that, for every stimulus-response pair, all of the stimulus power is concentrated at a single ripple (w_k, Ω_l) and all of the (linear) response power is concentrated at a single frequency w_k; hence, each point on the ripple transfer function $H_{k,l}$ is measured with maximal signal power. However, the chief disadvantage of this method is the time required to present all of the stimuli necessary to build a complete characterization of the transfer function. Typically, only two perpendicular cross-sections are measured within each quadrant of H — one in which Ω is varied while w is fixed and *vice versa*. The remainder of the quadrant is then estimated by the cross-product of these two sections. In doing so, it is presumed that the actual transfer function is *quadrant separable* (Kowalski et al., 1996a; Depireux et al., 1998b).

3.5 Spectro-temporal reverse correlation

Preliminary studies have suggested that the quadrant separability assumption is reasonable for neurons in ferret AI (Kowalski et al., 1996a; Depireux et al., 1998a; Depireux et al., 1998b). However, it is possible that spectro-temporal processing in other auditory centers is not well described by quadrant separable STRFs. A more general approach to STRF estimation, which can be used to avoid such *a priori* assumptions about the structure of the STRF, is offered by the spectro-temporal reverse correlation method. The method was previously cast within a stochastic framework. Now, we have the tools to reevaluate it in the ripple domain.

First, it is noted that the measured response, R, may contain, in addition to the 'linear portion' r, produced by the STRF functional, another portion e, due to non-linear and random aspects of the system transformation, which are not described by the STRF:

$$R(t) = r(t) + e(t). \tag{18}$$

In the following, we will refer to the 'ideal linear case' as the case in which the response is completely specified by the STRF, i.e., $e(t) = 0$.

The spectro-temporal reverse-correlation function C, is obtained by cross-correlating the dynamic spectrum of the stimulus with the response:

$$C(\tau, x) \triangleq \frac{1}{T} \int_0^T S(t - \tau, x) \cdot R(t) \, dt. \tag{19}$$

Since C is often used, without modification, as the STRF estimate, it is of interest to explore the conditions by which it resembles the STRF.

Substituting (18) into (19), another useful expression for C is obtained:

$$C(\tau, x) = c(\tau, x) + \epsilon(\tau, x), \tag{20}$$

where the following definitions have been made:

$$c(\tau, x) \triangleq \frac{1}{T} \int_0^T S(t - \tau, x) \cdot r(t) \, dt, \tag{21}$$

$$\epsilon(\tau, x) \triangleq \frac{1}{T} \int_0^T S(t - \tau, x) \cdot e(t) \, dt. \tag{22}$$

Ideally, then, $\epsilon(\tau, x) = 0$. The remaining, linear part of the cross-correlation function c, is derived using (10), (17), and (21):

$$
\begin{aligned}
c(\tau, x) &= \frac{1}{T} \sum_k \sum_l \sum_{k'} \sum_{l'} a_{k,l} a_{k',l'} H_{k',l'} exp\left\{ j2\pi(-w_k t + \Omega_l x) \right\} \\
&\quad \cdot exp\left\{ j(\psi_{k,l} + \psi_{k',l'}) \right\} \int_0^T \exp\left\{ j2\pi(w_k + w_{k'})t \right\} \, dt \\
&= \sum_k \sum_l \sum_{l'} a_{k,l} a_{-k,l'} H_{-k,l'} exp\left\{ j[2\pi(w_{-k}\tau + \Omega_l x) + \psi_{k,l} + \psi_{-k,l'}] \right\}. \tag{23}
\end{aligned}
$$

This function contains the sum of the cross-correlations between each stimulus ripple component (w_k, Ω_l) and each response component $(w_{k'})$. Those terms for which $k' \neq -k$ (i.e., $w_{k'} \neq -w_k$) integrate to zero over the stimulus duration, i.e., they are temporally orthogonal. The remaining terms (for which $k' = -k$) are naturally divided into two groups: the 'self terms' c_s, for which $l' = -l$, and the 'cross terms' $c_\times$, for which $l' \neq -l$:

$$c(\tau, x) = c_s(\tau, x) + c_\times(\tau, x). \tag{24}$$

Each self term results from the cross-correlation between a particular stimulus ripple component and the corresponding response component evoked by it via the STRF functional. In comparing the self terms with the form of the STRF (16), it is immediately evident that they consist of STRF components weighted by the squared amplitudes of the stimulus components, i.e.,

$$
\begin{aligned}
c_s(\tau, x) &= \sum_k \sum_l (a_{k,l})^2 H_{-k,-l} exp\left\{ j2\pi(w_{-k}t + \Omega_l x) \right\} \\
&= \sum_k \sum_l (a_{k,-l})^2 H_{k,-l} exp\left\{ j2\pi(w_k t + \Omega_l x) \right\}. \tag{25}
\end{aligned}
$$

Note, e.g., by considering the magnitude of these terms $(a_{k,-l})^2 b_{k_l} = (a_{-k,l})^2 b_{k,l}$, that the recovered STRF components actually correspond to the stimulus components with the opposite modulation direction. This is essentially because, as defined, the time axis of the stimulus indicates progression whereas the time axis of the STRF indicates precedence. Nevertheless, it should be apparent that if the $a_{k,l}$ are relatively constant wherever $H_{k,l}$ is of significant magnitude, the self terms should resemble the STRF, aside from an over-all scale factor.

However, the cross terms bear no special resemblance to the STRF; although they do depend on the form of the STRF, the cross terms also depend on the phases of the stimulus components and are 'smeared' over all Ω:

$$c_\times(\tau, x) = \sum_k \sum_l \sum_{l' \neq -l} a_{k,-l} a_{k,l'} H_{k,l'} exp\left\{ j[2\pi(w_k t + \Omega_l x) - \psi_{k,-l} + \psi_{k,l'}] \right\}. \tag{26}$$

Cross terms arise when there are multiple stimulus components with the same modulation frequency $|w|$. As illustrated in Figure 5, although they have different ripple densities, such
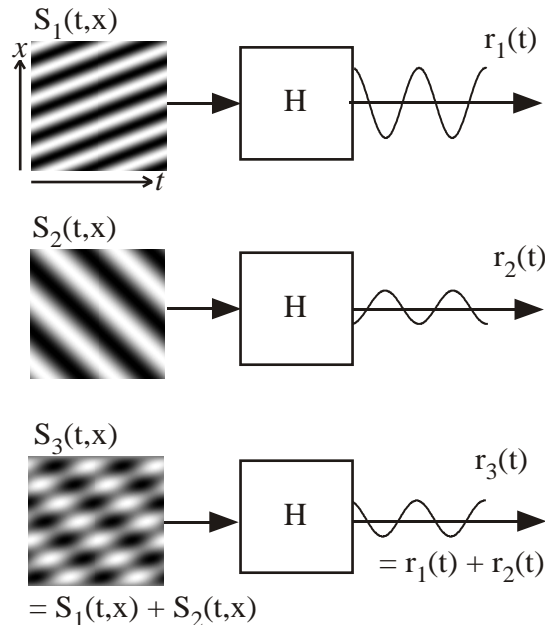
Figure 5: The responses r_1 and r_2 of this STRF-system H, to the ripple stimuli S_1 and S_2 with identical modulation rates $|w|$, are sinusoids with identical frequencies. Thus, the response r_3 to the sum of these stimuli S_3 is also a sinusoid of the same frequency. Consequently, one cannot unambiguously derive the form of the ripple transfer function H at the points corresponding to S_1 and S_2 from the response to S_3.

components evoke identical (i.e., overlapping) response frequencies, which sum to a single observed response component. This will, in turn, be correlated with every stimulus component having that same modulation frequency. The portion of the response not due to a particular stimulus component, but which is temporally correlated with it, constitutes a cross term.

Unfortunately, although the self terms and cross terms have been conceptually separated, they are not separate *per se* because they have overlapping ripple spectra. Moreover, the cross terms are in general of the same order of magnitude as the self terms. Thus, for arbitrary stimulation, there is no guarantee that C will closely resemble the STRF, even if the system is completely linear and S is spectro-temporally 'rich'. This is especially problematic for brief stimuli, when T is comparable to the memory of the system. In this case, the cross terms, which are relatively unstructured and diffuse, will be entirely manifest over the same range of latencies that the STRF components are expected to be.

4 The sum-of-ripples approach

Above, it was shown how, by use of the Fourier series, the dynamic spectrum S of any given stimulus can be described as a sum of ripples. It was found that, in general, non-idealities of the dynamic spectrum, manifest by significant cross terms $c_\times$ and irregular ripple-component amplitudes, will cause the spectro-temporal cross correlation function,

$$C(\tau, x) = c_s(\tau, x) + c_\times(\tau, x) + \epsilon(\tau, x), \tag{27}$$

to be significantly different from the STRF, even in the ideal linear case, i.e., $\epsilon(\tau, x) = 0$. The non-ideal stimulus structure might be corrected for *a posteriori* (Eggermont et al., 1983b; Theunissen et al., 1998), but the correction procedure can be difficult and time consuming, in general requiring the computation and delicate manipulation of large multi-dimensional correlation matrices. The considerable error associated with these procedures in turn propagates to the STRF estimate (Eggermont et al., 1983b).

An alternative and considerably simpler approach is to create stimuli for which the correction procedure is trivialized. Such is the credo of the 'sum-of-ripples approach', where deterministic dynamic spectra are conscientiously designed using a finite sum of ripples. Considering the results so far, the design process is primarily focused on three issues: equalization of ripple-component amplitudes, minimization of the cross-term power, and maximization of the self-term power. Given experimental constraints, this leads to the design of short-duration stimuli with structure enriched with those spectral and temporal qualities germane to the auditory locus under study, and which consist only of ripples that are mutually temporally orthogonal.

4.1 Stimulus synthesis

The general expression used for the stimulus synthesis,

$$S(t, x) = \sum_{i=1}^{N} a_{k_i, l_i} cos \left\{ 2\pi(w_{k_i} t + \Omega_{l_i} x) + \psi_{k_i, l_i} \right\}, \tag{28}$$

details the design of the dynamic spectrum with the sum of N *distinct* moving ripples. As before, the ripple densities and velocities available for selection depend on the duration, T, and bandwidth, X, of the stimuli, through the relations given in (11). To preserve the notation of (10), the particular ripples chosen are parameterized by the list of indices $\mathbf{k} = [k_1, k_2, \cdots, k_N] \in (-\infty, \infty)$ and $\mathbf{l} = [l_1, l_2, \cdots, l_N] \in [0, \infty)$, corresponding to the points (w_{k_i}, Ω_{l_i}) located in the upper two quadrants of the ripple domain. In exception are the points along the left side of the w axis ($w < 0, \Omega = 0$) which, being already specified by the complex conjugates of the points on the right, must be excluded from this list. Consequently, in the stimulus analysis expression (10), there will be $2N$ terms for which $a_{k,l}$ is non-zero. These correspond directly to the N points above plus an additional N points at the complex-conjugate locations $(-k_i, -l_i)$ in the lower two quadrants (which includes the left side of the w axis). In practice, S also has a mean sound level $a_{0,0}$ which is set to a reasonable, intermediate value.

To facilitate the recovery of the STRF from the self terms, the stimuli are constructed with constant-amplitude ripples, i.e., $a_{k_i, l_i} = a$ for all $1 \leq i \leq N$, so that (25) reduces to

$$c_s(\tau, x) = a^2 \cdot STRF(\tau, x; -\mathbf{k}, \mathbf{l}). \tag{29}$$

Thus, the deconvolution procedure generally required to recover the STRF from c_s reduces to a division by a known scalar. The recovered $STRF(\tau, x; -\mathbf{k}, \mathbf{l})$ corresponds exactly to the STRF, 'reconstructed' with those moving ripples present in the stimulus, as parameterized by the indices $\mathbf{k}$ and $\mathbf{l}$.

It is desired for the power of the self terms, and thus a, to be maximized, so that they will have a robust presence in C. For a given N, this is equivalent to maximizing the stimulus power $P = 2Na^2$. However, there are constraints on P, because the dynamic range of S is limited, i.e.,

the sound level cannot be modulated beyond certain extreme values (e.g., below zero and above damaging levels). Therefore, the phases of the stimulus components are typically randomized, or otherwise chosen to reduce the 'peakiness' of the dynamic spectrum. Intuitively, this allows one to pack more power over a limited dynamic range (Boyd et al., 1983).

The range of ripples used to build the stimulus (i.e., **k** and **l**) should be relevant to the auditory locus being studied, i.e., they should lie within the expected non-zero extent of the ripple transfer function. For example, in mammal AI it has been found that a great majority of neurons respond only to temporal modulations from about 4 to 40 Hz and to spectral modulations within 0 to 2 c/o (Langner, 1992; Schreiner and Calhoun, 1995; Shamma et al., 1995; Kowalski et al., 1996a; Depireux et al., 1998b). A stimulus tailored for AI, which contains two-hundred constant-amplitude random-phase ripples arranged within these bounds, is shown in Figure 6(a). In contrast, a stimulus constructed for use in the inferior colliculus (IC) is shown in Figure 6(b). The main difference between the two stimuli is the inclusion of much higher modulation rates in the IC stimulus, because of the sensitivity to higher rates known to exist in this locus (Langner, 1992).
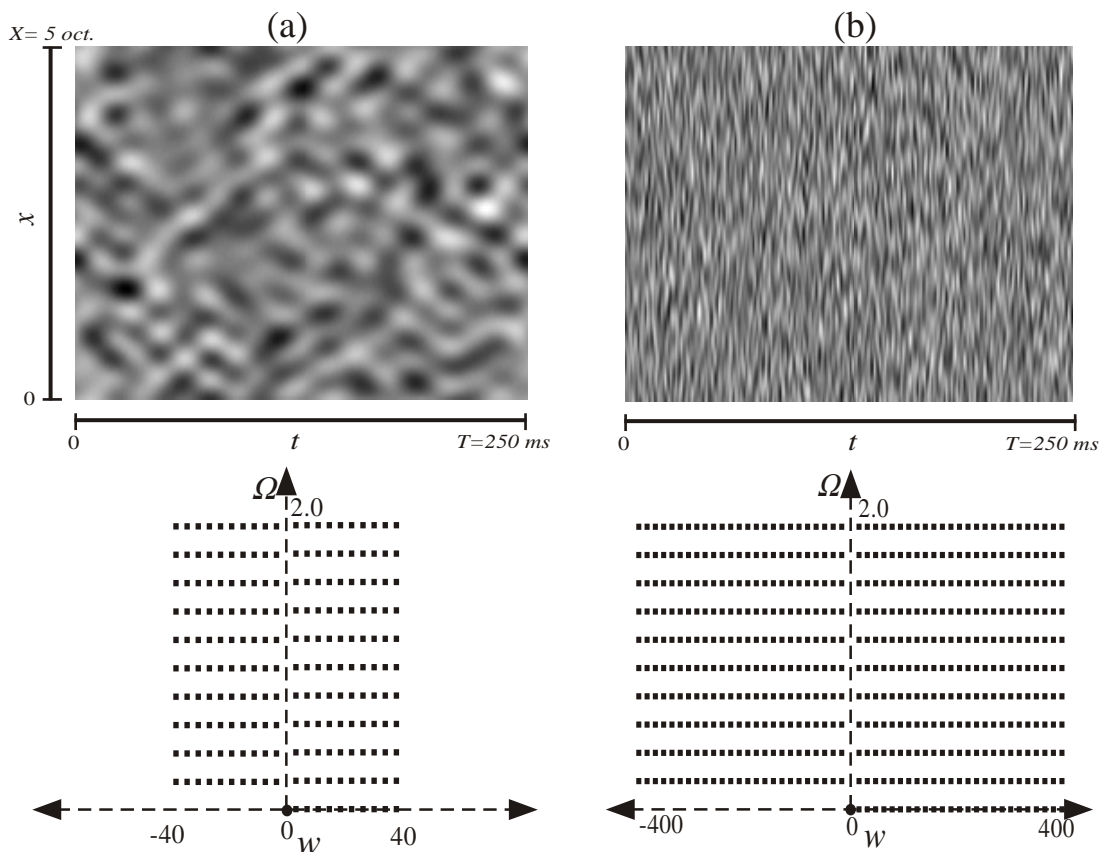

Figure 6: Shown here are examples of sum-of-ripples stimuli used for STRF estimation (phase-averaging method; see Section 4) in ferret AI (a) and IC (b). Each stimulus is composed of a range of constant-amplitude, random-phase ripples presumed, from prior experiments, to be relevant to the corresponding auditory locus. Above, the spectro-temporal envelopes (Depireux et al., 1998b) are shown. Below, the ripple content of each is indicated.

While it is obviously paramount for the stimulus to *contain* relevant spectral and temporal modulations, the main reason to *restrict* the ripple spectrum as such is that the inclusion of additional ripples doesn't improve the reconstruction of the STRF; it only serves to decrease a, and thus the self-term power. This has been sketched in Figure 7. In the limit, as the number of ripples N extends to infinity, the stimulus becomes spectro-temporally 'white.' However, since P is limited, the ripple amplitudes must all decrease by $\sqrt{N}$ in the process.
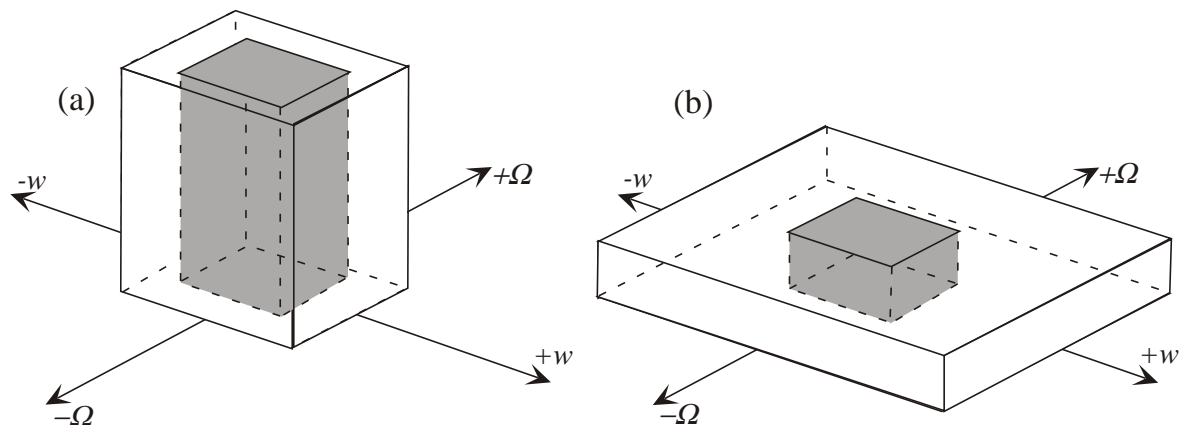

Figure 7: The ripple spectra of two hypothetical stimuli are shown. The vertical axis indicates the amplitude squared of the ripple components at each point (w, Ω). The total power $P = 2Na^2$ thus corresponds to the total volume of the box. The shaded volume indicates that portion 'seen' by a neuron that is only responsive to a central region of w's and Ω's. Although the two stimuli have identical total power, stimulus (b), with power spread outside the relevant range of the ripple domain (e.g., white noise or tones), will on average evoke much weaker responses.

4.2 Cross-term removal by phase averaging

Dynamic spectra such as that shown in Figure 6 are efficacious in that they can be used to measure the entire STRF at once. However, in general such stimuli present a problem in that they contain multiple components with common ripple velocities, resulting in the generation of cross terms in C. One way to remove the cross terms, detailed in this section, is by 'phase averaging'. This method has been used for STRF estimation in ferret AI (Shamma et al., 1998). Although it has since been supplanted by the preferred 'TORC method,' detailed in the next section, the phase-averaging method provides a valuable view of how accurate STRF estimation is, in principle, possible with stochastic and ergodic stimulation, for whom time averaging is equivalent to phase averaging (Wiener, 1958; Victor and Knight, 1979).

The phase-averaging method takes advantage of the fact that the phases of the cross terms (26) depend on the phases of the stimulus components $\mathbf{\Psi} = [\psi_{k_1,l_1}, \psi_{k_2,l_2}, \cdots, \psi_{k_N,l_N}]$, and so their expected value over all sets of stimulus phases is zero, i.e.,

$$E_\psi \left\{ c_\times(\tau, x; \mathbf{\Psi}) \right\} = \frac{1}{(2\pi)^N} \int\limits_0^{2\pi} c_\times(\tau, x; \mathbf{\Psi}) \, d\mathbf{\Psi} = 0, \tag{30}$$

as long as the phases ψ_{k_i,l_i} each vary uniformly over an interval of 2π radians.

Using (27), (29), and (30), the expected value of C is then

$$
\begin{aligned}
E_\psi\left\{C(\tau,x)\right\} &= c_s(\tau,x) + E_\psi\left\{\epsilon(\tau,x)\right\} \\
&= a^2 \cdot STRF(\tau,x;-\mathbf{k},\mathbf{l}) + E_\psi\left\{\epsilon(\tau,x)\right\}
\end{aligned}
\tag{31}
$$

since the self terms do not depend on $\boldsymbol{\Psi}$. In the ideal linear case, then, the expected value of C is identical to the self terms, which bear a scalar relationship to the STRF components. In practice, however, this expected value cannot be met exactly, but must be approximated either as an average over a sufficiently long time, or over a sufficiently large number of random-phase stimuli. Below, a multiple-stimulus phase-averaging procedure is detailed.

A total of M stimuli are used for the phase average. The ith stimulus, $S_i(t,x;\boldsymbol{\Psi}_i)$, is constructed with the same set of constant-amplitude ripples but with a new, random set of phases $\boldsymbol{\Psi}_i$ drawn from a uniform distribution. Eq. (31) is then approximated by averaging C over M stimulus-response pairs, denoted by $\langle C \rangle_M$:

$$
\langle C(\tau,x)\rangle_M = \frac{1}{M}\sum_{i=1}^{M} C_i(\tau,x;\boldsymbol{\Psi}_i),
\tag{32}
$$

where C_i is the cross-correlation of the ith stimulus-response pair. Finally, the phase-average STRF estimate is obtained by dividing (32) by $a^2 = P/2N$:

$$
\begin{aligned}
STRF_{est}(\tau,x) &= \frac{1}{a^2}\langle C(\tau,x)\rangle_M \\
&= STRF(\tau,x;-\mathbf{k},\mathbf{l}) + \frac{2N}{P}\left\{\langle c_\times(\tau,x)\rangle_M + \langle\epsilon(\tau,x)\rangle_M\right\}.
\end{aligned}
\tag{33}
$$

The random-phase-averaged cross terms $\langle c_\times(\tau,x)\rangle_M$ scale in magnitude by approximately $1/\sqrt{M}$. Thus, the error power, P_E,

$$
P_E \triangleq \frac{1}{TX}\int_0^T\int_0^X \left\{STRF_{est}(\tau,x) - STRF(\tau,x;-\mathbf{k},\mathbf{l})\right\}^2 dx\,dt
\tag{34}
$$

will be reduced roughly by $1/M$. Simulation results of this method are shown in Figure 8(a), and results obtained from ferret AI can be also seen in Figures 8(b) and 10(b).

4.3 Temporally-orthogonal ripple combinations

Because the STRF functional dictates a two-dimensional to one-dimensional input-output domain transformation, separate (i.e., orthogonal) stimulus components can evoke overlapping response components (as in Figure 5). This produces an initial ambiguity in the STRF estimate, manifest by the cross terms in (23). Above, it was shown that these cross terms can be removed by phase-averaging but, like the white-noise approach, an infinite amount of time is required to achieve error-free STRF estimation. Thus, with the phase-average method, an acceptable error level may not always be practically achievable, even if the system were linear and noiseless.
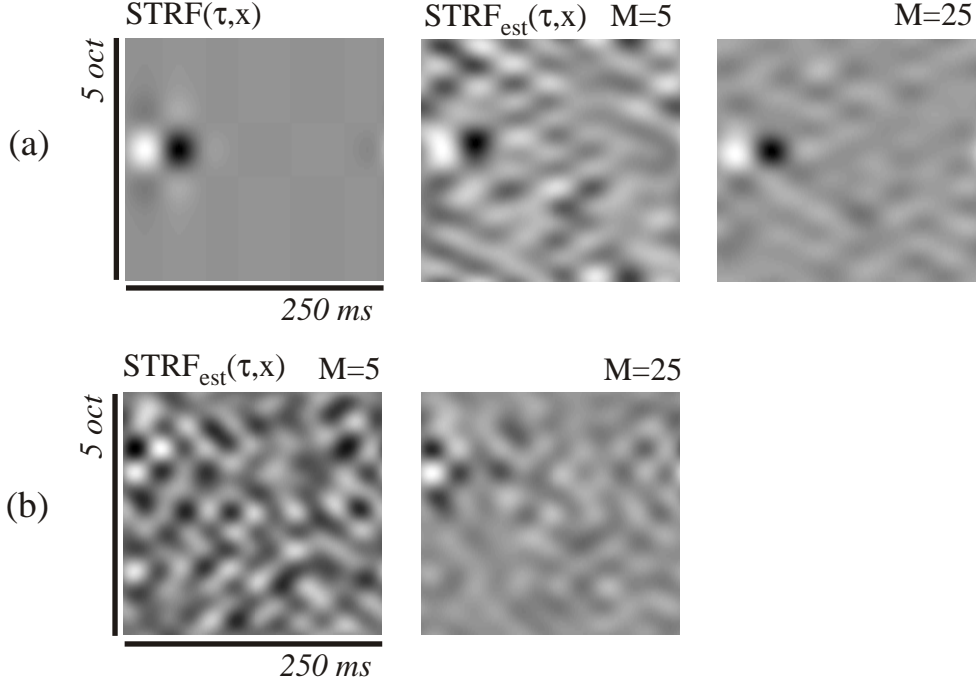
Figure 8: (a) Simulation results of the phase-averaging method using ideal linear data. Shown are the estimates after and 5 and 25 stimuli, as noted. (b) Experimental results from ferret AI. For all stimuli, $N = 96$, $X = 5$ oct., $T = 250$ ms., $|w| \leq 24$ Hz. and $|\Omega| \leq 1.6$ c/o. For the experimental results, delivered acoustic waveforms were constructed from their predetermined spectro-temporal envelopes, as detailed in (Depireux et al., 1998b). PSTH's (with 1 ms. bins) from single neurons, constructed from 100 periods T of each stimulus, were used as responses. For additional results, see also Figure 10(b).

In order to reap the full benefits that can be attributed to the sum-of-ripples approach, the stimulus components can be chosen more carefully so that they all evoke separate response frequencies. This is accomplished by only superimposing moving ripples whose velocities all differ in absolute value, i.e.,

$$|w_k| \neq |w_{k'}|, \tag{35}$$

for all $k \neq k'$. Through (11), this is equivalent to the restriction $|k_u| \neq |k_v|$ for all $u \neq v$ for $\mathbf{k}$ used in the stimulus synthesis (28).

Such ripples are called 'temporally orthogonal' since their temporal correlation is zero, i.e.,

$$\int_0^T \{[a_{k,l} \cos(2\pi w_k t + 2\pi \Omega_l x + \psi_{k,l})] \cdot [a_{k',l'} \cos(2\pi w_{k'} t + 2\pi \Omega_{l'} x + \psi_{k',l'})]\} \, dt = 0 \tag{36}$$

for any $|w_k| \neq |w_{k'}|$. A stimulus composed of two or more such ripples is referred to as a *temporally-orthogonal ripple combination* (TORC).

Since there is not any component overlap in the STRF-based response to a TORC, each response component is temporally orthogonal to every stimulus component *except* for the one responsible for evoking it. Therefore, the cross terms are identically zero, and C reduces directly to

$$C(\tau, x) = a^2 \cdot STRF(\tau, x; -\mathbf{k}, \mathbf{l}) + \epsilon(\tau, x), \tag{37}$$

without the need for any averaging. As such, the STRF components at $(-\mathbf{k}, \mathbf{l})$ are recovered, by simply dividing C by a^2, with an accuracy which immediately surpasses the capabilities of the phase-averaging method.

Along with the temporal orthogonality restriction comes some additional challenges for dynamic-spectral design, because there is less flexibility for the positioning of stimulus components within the significant extent of H. Below, two stimulus configurations, which have proved to be useful in experiments and simulations, are detailed.

With TORC method I, a set of N ripples, deemed adequate to reconstruct the STRF, is equally subdivided into a group of M TORCs. One such stimulus set is shown in Figure 9(a). In this particular design, the individual stimuli span different rows in the ripple domain; each TORC was built using a single ripple density and a range of ripple velocities. Consequently, each stimulus-response pair is used to measure a single row of H. This design is interesting because it can be used to directly investigate how the dynamical processing of the system changes at different levels of spectral-peak density.
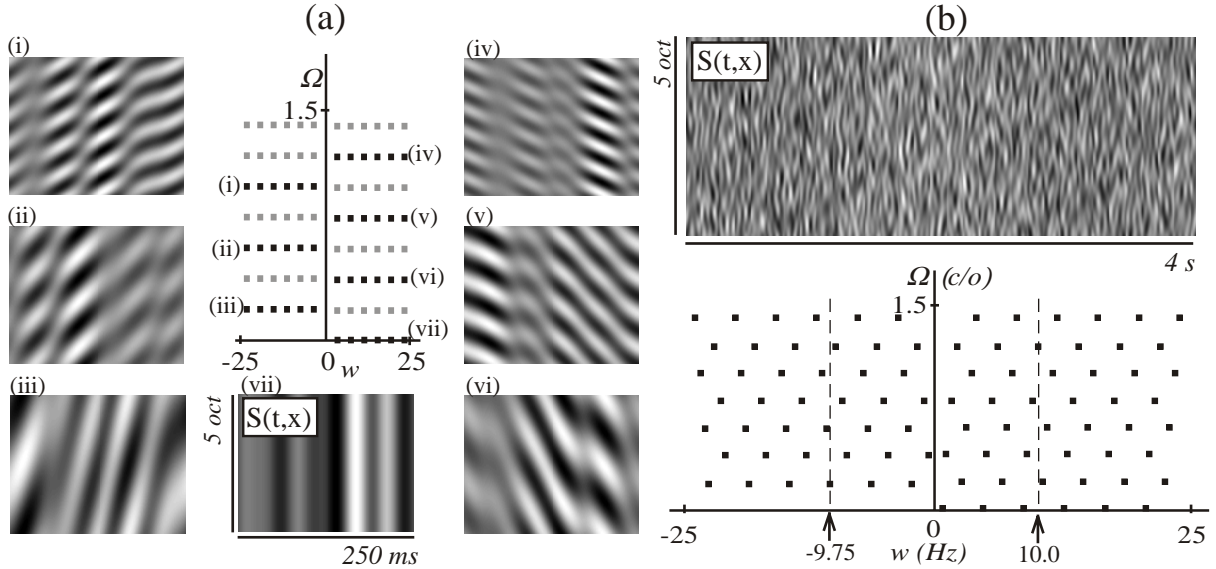

Figure 9: (a) TORC method I. The indicated ripple content (black and gray points) refers to that of the entire stimulus ensemble. The dynamic spectrum of several individual stimuli (i–vi), are shown. The ripple content of each is indicated by the black rows of points similarly labeled. (b) TORC method II. The stimulus shown has roughly the same ripple content and duration as that of the entire ensemble in (a). Because of its longer duration, the 0.25 Hz spacing allows a more clever positioning of ripple components. Note that none of the components have the same modulation rate $|w|$.

If some stimulus power can be sacrificed, one can alternatively use TORC method II: a single TORC with a longer duration. This allows for finer resolution in the w direction (through (11)), and thus greater flexibility for the positioning of ripple components. This approach is illustrated by the stimulus shown in Figure 9(b), which covers a portion of the ripple domain relevant to AI while fulfilling the temporal orthogonality condition. The stimulus is sixteen times as long as the stimuli in 9(a). However, it still spans only four seconds, corresponding to the 0.25 Hz spacing between ripple components.

These two TORC design strategies can be merged into a single expression for the STRF estimate:

$$
\begin{aligned}
STRF_{est}(\tau, x) &= \frac{2(N/M)}{P} \sum_{i=1}^{M} C_i(\tau, x) \\
&= STRF(\tau, x; -\mathbf{k}, \mathbf{l}) + \frac{2N}{P} \langle \epsilon(\tau, x) \rangle_M .
\end{aligned}
\tag{38}
$$

Here, the C_i are added, not averaged, and then scaled by $1/a^2$. The indices $\mathbf{k}$ and $\mathbf{l}$ now refer to the ripple content of the entire set of M stimuli, and each stimulus (indexed by i) contains an equal-sized portion N/M of a total of N ripples.

Finally, P_E in this case is

$$
P_E = \left(\frac{2N}{P} \right)^2 \int_0^T \int_0^X \{ \langle \epsilon(\tau, x) \rangle_M \}^2 \, dx \, dt.
\tag{39}
$$

If $e(t) = 0$, $P_E = 0$. Thus, perfect recovery of the STRF is now achievable in the ideal case.

In Figure 10, STRF estimates produced by the TORC methods are illustrated, and compared to the phase-averaging method. Both for ideal, simulated data (a) and for actual data from ferret AI (b), the TORC estimates are apparently superior in all cases (see also Figure 8).

4.4   Non-idealities

It has been shown that, if a neuron's response is completely determined by the STRF functional, the TORC method achieves perfect STRF estimation by avoiding the generation of cross terms in C. However, as the experimental results suggest (Figure 10(b)), in reality the remaining error power P_E (39), is different from zero. The source of this estimation error is e (18), the portion of the response not accounted for by the STRF. Thus, like e, it may include a random element due to noise, and a deterministic element due to system non-linearities.

For spiking systems, the PSTH is likely to be a primary source of estimation error. As evident in Figure 11, this error is manifest over a broad range of frequencies. With general stimulation, it will, in turn, be erroneously and randomly correlated with a broad range of stimulus components, thus creating an unstructured variability in the STRF estimate. Fortunately, the error, having resulted in (22) from a cross-correlation of e with S, is restricted to the same portion of the ripple domain as the stimulus ensemble. Thus, sum-of-ripples estimates can be considered filtered versions of estimates obtained with general stimulation, such that only those ripple components expected to be strongly manifest in the STRF appear in the estimate. Consequently, much of the noise, especially at high-frequencies, that has been typical of published STRF estimates and has necessitated the use of smoothing filters and thresholds, can be avoided.

Furthermore, as seen in Figure 11, the quality of the PSTH approximation generally improves as more spikes are used in the histogram. To benefit this cause, the sum-of-ripples stimuli are trivially made periodic without compromising their structural idealities. Since the temporal frequencies w_k used to design the stimulus are commensurate with period T, periodicity is achieved simply by extending the total stimulus duration include as many repetitions of the stimulus as necessary.
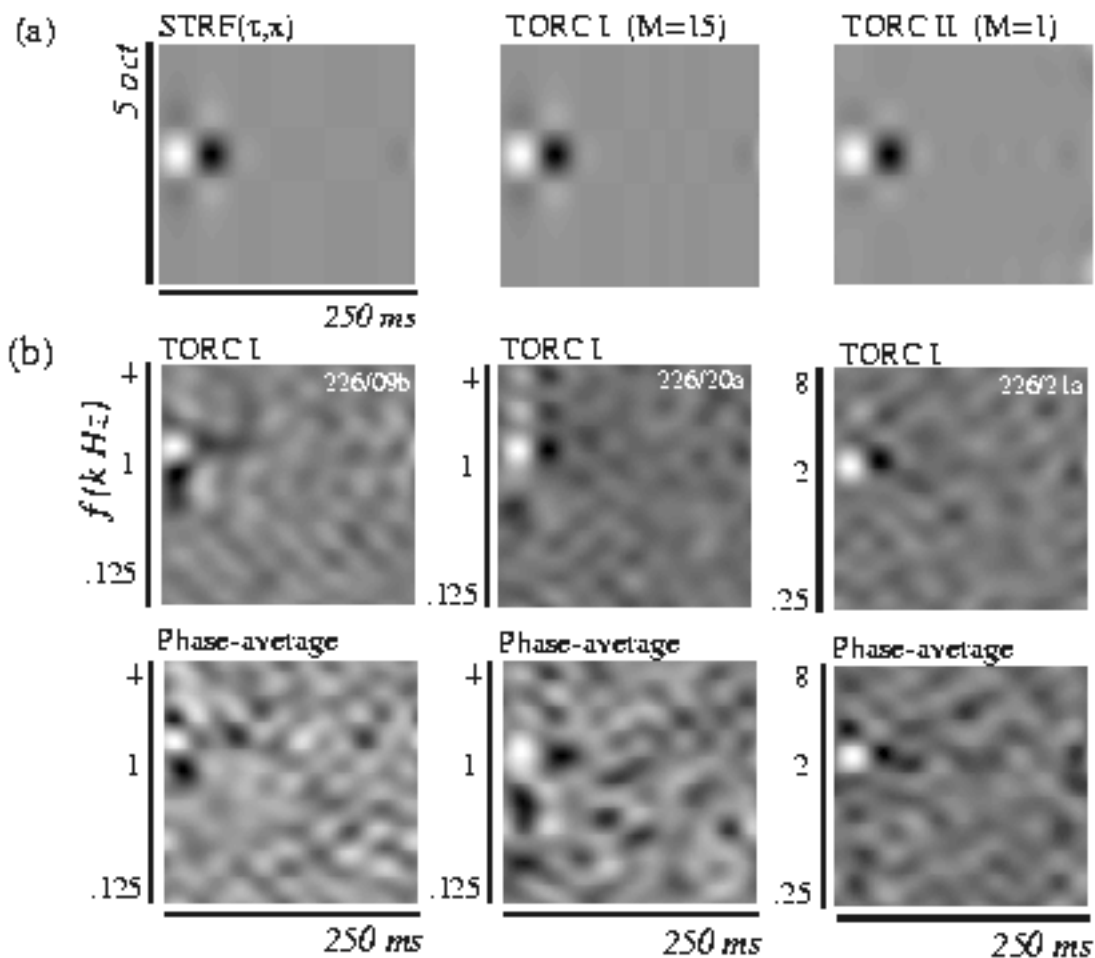
Figure 10: (a) Noiseless, linear simulation of TORC method I ($M = 15$) and TORC method II ($M = 1$). (b) Comparison of experimental results from ferret AI, produced with TORC method I ($M = 15$), and by phase-averaging with stimuli akin to that used for Figure 8 ($M = 25$). For all stimuli, $T = 250$ ms., except $T = 4$ sec. for the TORC II stimulus.

With increases the stimulus per-component power, (39) predicts that P_E will decrease quadratically. While this may be true for the random portion of the error, the deterministic (non-linear) portion of the error term, being systematically related to the stimulus, cannot, in general, be reduced by increasing the stimulus strength; if anything, its contribution is multiplied.

As for the structure of this systematic error, one can consider the difference between the first-order Volterra v_1 and Wiener w_1 kernels for a third-order, single-input system (Eggermont, 1993):

$$w_1(\tau) = v_1(\tau) + 3P \int v_3(\tau_1, \tau_2, \tau_2) \, d\tau_2. \tag{40}$$

Similarly, there is expected to be a systematic difference between the estimated and actual STRFs if the system contains higher-order non-linearities of the same parity as the STRF.

This phenomenon can also be understood from a sum-of-sinusoids standpoint. For example,
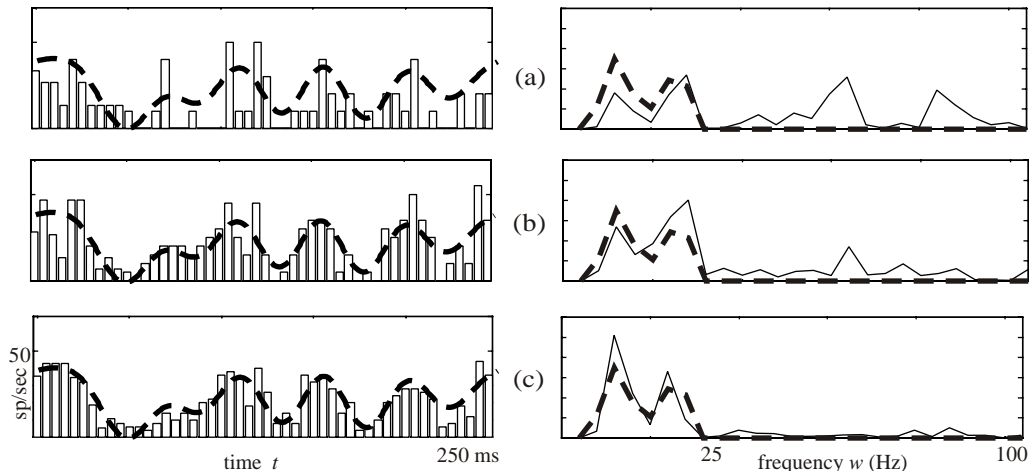
Figure 11: The response error due to the PSTH approximation improves with the number of spikes. Spike histograms (right), created with a 5 ms bin width, and their corresponding power spectra were created with 20 (a), 40 (b), and 100 (c) periods of a $250ms$ TORC (method I), and contain 65, 191, and 472 spikes, respectively. The superimposed dashed curve corresponds to the smooth response predicted by this neuron's STRF.

the response of a cubic non-linearity to a sum of sinusoids contains frequencies that result from additive or subtractive combinations of any three input frequencies ($w_k \pm w_l \pm w_m$), which always includes frequencies that overlap with the linear response (Victor and Shapley, 1980; Victor, 1991). Combination frequencies due to quadratic non-linearities ($w_k \pm w_l$) may also overlap with the linear response; however, unlike the cubic contribution, they can be removed relatively easily, e.g., using the inverse-repeat method (Swerup, 1978) (with respect to the dynamic spectrum). Alternatively, it is possible to choose the input frequencies so that there is no overlap between the quadratic and linear portions of the response (Victor and Shapley, 1980). Of course, if non-linearities are evident, they should ultimately be incorporated into the model. This, however, is beyond the scope of this paper.

5   Discussion and summary

Spectro-temporal reverse correlation was developed as a means to simultaneously measure the interdependence of several important physiological properties of auditory neurons—including frequency tuning, lateral inhibition, latency, and modulation rate tuning—related to both spectral and temporal aspects of stimuli (Aertsen and Johannesma, 1980; Aertsen et al., 1980a; Aertsen et al., 1980b). The spectro-temporal receptive field (STRF), thence born of empiricism, promised descriptive capabilities transcending those of separate spectral and temporal measures (Aertsen et al., 1980a; Hermes et al., 1981; Eggermont et al., 1981). Subsequently, the theory of the STRF as a functional property of neurons, characterizing a comprehensive 'spectro-temporal sensitivity,' independent of the means used to measure it, was further developed, and relations were drawn between the STRF and the Volterra and Wiener functional expansions (Aertsen and Johannesma, 1981b; Hermes et al., 1981; Eggermont et al., 1983c).

The Volterra and Wiener series, in all of their mathematical elegance, are not biologically

motivated constructs. Therefore, the close correspondence between the second-order Volterra functional and the STRF functional represents a unique opportunity to join empiricism with theoretical rigor in the study of the auditory system. In doing so, the Volterra parallel suggests that the input to central auditory neurons may be represented by members of a specific class of (quadratic) joint time-frequency representations of the stimulus. If the representation actually employed is a more complicated functional of the stimulus waveform (such as a that produced by a non-linear cochlear model (Carney and Friedman, 1998)), it may be difficult to strictly compare the corresponding STRF to any one term in the Volterra or Wiener expansions (with respect to the waveform). Of course, the resulting model may be more concise; intuition should not be abandoned just to fulfill the requirements of the Volterra series. Intuition brought the STRF into use; the Volterra and Wiener series have only been useful insofar as they have aided in evaluating the completeness of the STRF, and the challenges inherent with its measurement.

The evaluation of the white-noise approach has led to the conclusion that stationary Gaussian noise stimulation is not well suited for STRF estimation in most auditory areas, and particularly those most central. It is apparent that improvements should be made in at least three aspects. First, the stationarity of the stimulus is often cited as being an undesirable; it is thought that stimulation should be more 'dynamic' and 'natural,' (Smolders et al., 1979; Eggermont et al., 1983c; Yeshurun et al., 1985; Nelken et al., 1997; Nelken et al., 1999). A second (and related) improvement is that, for efficient laboratory use, stimulation should be brief and restricted to that which elicits responses that are reliably measured. Finally, while fulfilling these criteria, it is obviously desired for the structure of the stimulus ensemble to be such that the simple act of spectro-temporal reverse correlation produces an accurate reconstruction of the STRF.

The realization of these improvements depends strongly on the adopted theoretical framework. For example, the stochastic and multiple-input framework in which STRF estimation is typically cast suggests that power fluctuations across discrete frequency bands, i.e., 'channels' of the dynamic spectrum should be uncorrelated, to prevent cross-channel mixing in the computation of the stimulus-response cross-correlation C (Eggermont et al., 1983b). While enforcing this condition, improvements inevitably focus solely on the temporal structure within the individual channels (Eggermont et al., 1983b; deCharms et al., 1998; Kvale et al., 1998).

We have adopted a new framework, under which further improvements are realized by manipulating both spectral and temporal aspects of stimuli simultaneously. Our analysis, inspired by the linearity of the STRF functional, was based upon a Fourier decomposition of the spectro-temporal domain into two-dimensional sinusoidal components, 'moving ripples,' each of which embodies a unique intersection of spectral-peak density, modulation rate, and modulation direction. This put the analysis of the system on equal footing with that of single-input linear systems; the stimulus has a 'ripple spectrum,' and the system has a 'ripple transfer function' which describes its linear response to moving ripples. This led directly to an analytical expression, in terms of these quantities, for C which holds for *any* given stimulus. From within this framework, we have addressed each of the proposed improvements.

The labels 'dynamic' and 'natural' are related not only to the total power of the dynamic spectrum, but also the way that it is distributed among its ripple components. White noise, for example, has power spread thin over *all* ripple components; presumably, it gives rise to a stationary percept because the power relegated to the narrow range of spectral-peak densities and modulation rates at which human observers can perceive changes (Chi et al., 1999) is

relatively weak. The same can be said for auditory neurons; at the level of AI, the range of densities and rates to which neurons are responsive is quite narrow. Interestingly, it also over this range that natural sounds seem to hold their energy (Attias and Schreiner, 1997; Chi et al., 1999). Thus, in realizing dynamic stimulation whose structure approaches that of natural sounds, the power of the dynamic spectrum should be focused over this relevant range of ripple components. This is not possible while keeping the channels uncorrelated; the ripple spectrum of such a stimulus is dispersed over all ripple densities.

Besides the prospect of augmenting the applicability of the STRF estimates (e.g., for predicting responses), there are other practical reasons for restricting the ripple spectrum of stimuli as such. By dedicating the stimulus power wholly to those components that are relevant to a given locus, and thus maximizing the strength of the stimulus 'seen' by that system, the strength of the STRF-mediated response in that locus is maximized. This is especially helpful in biological systems identification, because there is a significant amount of measurement noise that has to be overcome. Further gains in signal-to-noise ratio are made possible by the brief and periodic nature of the sum-of-ripples stimuli, which brings a computational advantage as well; the stimulus-response cross-correlation need only be performed between the period-averaged response and one period of the stimulus, whose structure is known ahead of time. In total, all of these improvements bring to the laboratory the greater possibility of reliably measuring meaningful STRFs in a short period of time, in loci where previous attempts have failed.

Paramount to such concerns, in order for C to be an undistorted representation of the STRF, it was shown, after separating C into 'self terms' and 'cross terms,' that two basic criteria must be satisfied by the delivered stimulus ensemble. First, it is necessary for the ripple spectrum of the ensemble to be flat over the significant extent of the ripple transfer function, so that the self terms reduce to STRF components, scaled by a known factor. Subsequently, the cross terms were identified as a primary source of estimation error, especially with brief stimuli, thus warranting their removal.

Two strategies were outlined for removal of the cross terms. First, the phase-averaging method, for which the stimulus' ripple spectrum has randomly varying phase, offers insight into how STRF estimation can be accomplished with traditional, stochastic (and ergodic) stimulation. However, since there is a limit to the amount of averaging that can be performed in an actual experiment, there is a practical lower limit to the estimation error caused by the cross-terms. Due to the relatively large initial magnitudes of the cross terms which, interestingly, partially depend on the structure of the STRF, this limit may not be acceptable in all cases.

A preferable strategy, through which the cross terms were completely avoided, involves the use of deterministic stimuli whose ripple components all differ in modulation rate—TORCs— and thus all evoke different response frequencies. With the TORC method, no stimulus averaging is needed and, therefore, it is possible to achieve error-free STRF estimates with stimuli having a duration comparable to the memory of the system. Also, due to this brevity, and the predetermined structure of the stimuli, it has been possible, once recording the response, to compute the STRF estimate in a few seconds, 'on the fly'. This is in stark contrast to other, stochastic estimation schemes.

That perfect STRF estimation is possible with TORC stimulation might be surprising, since different channels of a TORC can be strongly correlated. Fortunately, the TORC mechanism can be illuminated in the same light as ideal-white noise, from within a multiple-input framework. By inserting the STRF functional (1) directly into (19) and rearranging terms, one

obtains for the spectro-temporal cross-correlation function:

$$C(\tau, x) = \int \int STRF(\tau', x') \cdot \Phi(\tau - \tau', x, x') \, d\tau' \, dx' + \epsilon(\tau, x). \tag{41}$$

Here,

$$\Phi(\tau - \tau', x', x) \triangleq \int S(t - \tau', x') S(t - \tau, x) \, dt \tag{42}$$

is a function which, in the discrete channel interpretation, describes the cross-correlation between two channels x' and x of the stimulus' dynamic spectrum. Thus, a *single* channel x of C is produced by the sum of the convolutions of *every* channel x' of the STRF with the cross-correlations between the channels x' and x of the stimulus. This rather complicated expression illustrates the difficulty in disentangling the STRF from C for an stimulus of arbitrary structure.

It can be shown, however, that for both an ideal-white dynamic spectrum and for a TORC, but certainly not in general, this expression reduces to a relatively simple two-dimensional convolution between the STRF and a spectro-temporal filter $\Phi(\tau, x)$; for these two special cases, Φ depends only on the channel difference, $x - x'$, and is given by the (two-dimensional) auto-correlation of S (see Appendix B).

These 'auto-correlation filters', for an ideal-white stimulus and for the TORC of Figure 9(b) (method II), are compared in Figure 12. For white noise (a), $\Phi(\tau, x) = \delta(\tau, x)$, since the only non-zero channel cross-correlations are those between each channel and itself, and are all identical, impulse functions. For the TORC (b), the impulse is relaxed to a *sinc* function in both τ and x. This results from the restriction of the ripple spectrum to low scales and rates, and is only allowed because of the expected smooth, band-limited nature of the STRF. Besides the practical advantages that this restriction brings, another fundamental practical advantage that TORCs have over white noise is that their idealities are realizable, as long as the dynamic spectra of stimuli can be predetermined with reasonable accuracy.
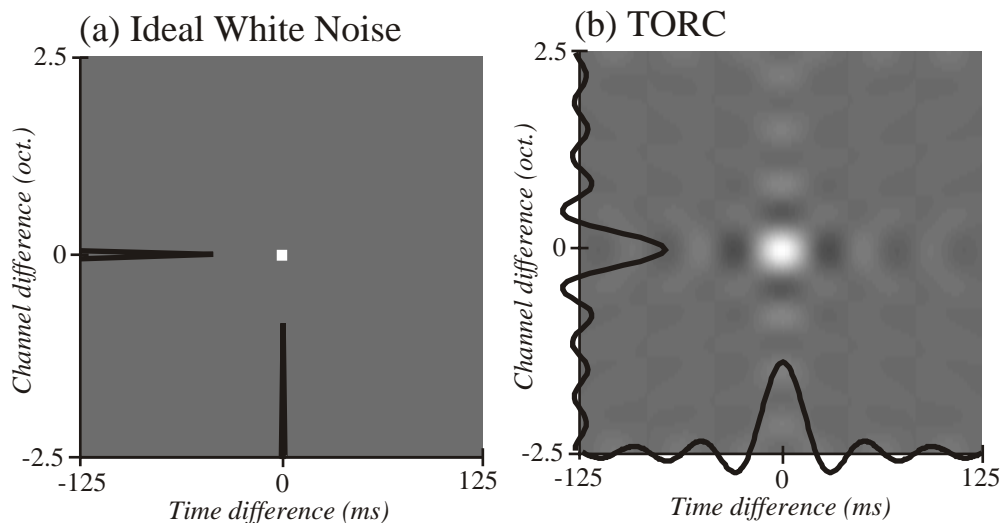

Figure 12: The two-dimensional auto-correlations of (a) ideal spectro-temporal white noise and (b) the TORC of Figure 9(b). See the text for an explanation.

It should be noted that, although it led to the development of specific sum-of-ripples stimuli like TORCs, the employed Fourier analysis has nothing specifically to do with ripple stimulation. Furthermore, its scope is not limited to any specific time-frequency representation. As long as there is some linear relationship between the employed representation S and the response of the system, as embodied by some STRF, the above conditions on S must be satisfied in order for C to be justified as an accurate STRF estimate. In addition, of course, the accuracy and completeness of the estimate depends on the satisfaction of the conditions imposed by the reverse correlation methodology, which include time-invariance, and the lack of significant higher-order non-linearities.

The use of moving ripples was originally inspired by the use of drifting sinusoidal luminance gratings in vision research (Valois and Valois, 1990). It is hoped that this development, performed in a spatio-temporal-like input domain, will further facilitate the exchange of ideas between the auditory, visual, and somatasensory sciences. For example, the principles that make TORC stimulation successful may be applicable to the design of stimuli for use in other sensory systems.

Acknowledgments

Appendix

A $\quad K_2$ and the generalized STRF functional

A large class of time-frequency representations S of a time waveform s, including the commonly used 'spectrogram', can be obtained by filtering the Wigner distribution W of s (defined in Eq. (6)) (Cohen, 1995), i.e.,

$$S(t, f) = \int \int g(t', f') \cdot W(t - t', f - f') \, dt' \, df'. \tag{43}$$

Thus, each member of this general class can be specified, via the Wigner distribution, by the structure of its corresponding filter function g. For example, for a spectrogram computed with a window function $h(t)$, g is given by the Wigner distribution of h (Cohen, 1995).

Using (43) for S in (1), one obtains

$$
\begin{aligned}
r(t) &= \int \int STRF(\tau, f) \cdot \left[\int \int g(t', f') \cdot W(t - t' - \tau, f - f') \, dt' df' \right] d\tau \, df \\
&= \int \int STRF'(t', f') \cdot W(t - t', f') \, dt' \, df',
\end{aligned}
\tag{44}
$$

where

$$STRF'(t', f') = \int \int STRF(\tau, f) \cdot g(t' - \tau, f - f') \, d\tau \, df. \tag{45}$$

Therefore, an STRF operating on the representation S that corresponds to the filter g, is equivalent to another STRF' operating upon the Wigner distribution, where STRF' is obtained by linearly filtering the original STRF by $g(t', -f')$. If the response of the system is given by (1) and, equivalently, (44), then we must have $STRF^{K_2}(t', f') = STRF'(t', f')$ (by the homogeneity of the Volterra functionals) and so, via (7), the system is also described by K_2 (5).

Working backwards is a bit trickier. If the system is described by (5) and, equivalently by (8), then it can also be described by other STRFs operating on other representations S within this general class. However, this is not true for any representation S. In this scenario, other STRFs are obtained by *inverse* filtering STRF^{K_2} with the corresponding g. Thus, g must have support everywhere STRF^{K_2} does in order for S to be useful. In terms of ripple spectra, introduced in Section 3, the ripple spectrum of g cannot be zero anywhere that the ripple spectrum of STRF^{K_2} is non-zero, or else S will not be able to represent some of the acoustic features that the system is responsive to.

B   The TORC channel cross-correlation function

We start by restating (42), and substituting (10) for S:

$$
\begin{aligned}
\Phi(\tau - \tau', x', x) &\triangleq \int S(t - \tau, x) \cdot S(t - \tau', x') \, dt \\
&= \sum_k \sum_l \sum_{k'} \sum_{l'} a_{k,l} a_{k',l'} \exp\left\{j[-2\pi(w_k \tau + w_{k'}\tau') + 2\pi(\Omega_l x + \Omega_{l'} x')]\right\} \\
&\quad \cdot \exp\left\{j(\psi_{k,l} + \psi_{k',l'})\right\} \int \exp\left\{j 2\pi(w_k + w_{k'})t\right\} \, dt \\
&= \sum_k \sum_l \sum_{l'} a_{k,-l} a_{k,l'} \exp\left\{j[2\pi w_k(\tau - \tau') + 2\pi(\Omega_l x + \Omega_{l'} x') - \psi_{k,-l} + \psi_{k,l'}]\right\}.
\end{aligned}
\tag{46}
$$

This development parallels that pursued in (23); one can now define self terms and cross terms, depending on whether or not they are the product of a stimulus component with itself, or the product of two different stimulus components. In fact, the self and cross terms in (46), via the convolution of each with the STRF in (41), are the progenitors of the self and cross terms in (23).

By definition, no two components of a TORC are temporally correlated. Thus, only the self terms will survive in (46). Stated more precisely, no two components of a TORC have the same $|w_k|$. Thus, for a given k, there is only one l such that $a_{k,-l} \neq 0$. Furthermore, there is only one l' ($= -l$) such that $a_{k,l'} \neq 0$. Therefore, if S is a TORC (46) reduces to

$$\Phi(\tau - \tau', x - x') = \sum_k \sum_l (a_{k,-l})^2 \exp\left\{j 2\pi[w_k(\tau - \tau') + \Omega_l(x - x')]\right\}. \tag{47}$$

Thus, for a TORC, the cross-correlation between two channels x and x' only depends on the channel difference $x - x'$.

The the two-dimensional auto-correlation of S can be expressed, again using (10), as

$$
\begin{aligned}
\alpha(t', x') &\triangleq \int\int S(t - t', x - x') \cdot S(t, x)\, dt\, dx \\
&= \sum_k \sum_l (a_{k,l})^2 \exp\left\{j2\pi(w_k t' + \Omega_l x')\right\}.
\end{aligned}
\tag{48}
$$

Regardless of the stimulus structure, α is solely a function of the time-difference t' and the channel difference x'. Thus, TORCs are among the special group of stimuli, which includes ideal-white noise (none other examples are known), whose channel cross-correlation functions Φ are given by their two-dimensional auto-correlation functions α.

References

Aertsen, A. and Johannesma, P. (1980). Spectro-temporal receptive fields of auditory neurons in the grassfrog. I. Characterization of tonal and natural stimuli. *Biological Cybernetics*, 38:223–234.

Aertsen, A. and Johannesma, P. (1981a). A comparison of the spectro-temporal sensitivity of auditory neurons to tonal and natural stimuli. *Biological Cybernetics*, 42:145–156.

Aertsen, A. and Johannesma, P. (1981b). The spectro-temporal receptive field: A functional characteristic of auditory neurons. *Biological Cybernetics*, 42:133–143.

Aertsen, A., Johannesma, P., and Hermes, D. (1980a). Spectro-temporal receptive fields of auditory neurons in the grassfrog. II. Analysis of the stimulus-event relation for tonal stimuli. *Biological Cybernetics*, 38:235–248.

Aertsen, A., Olders, J., and Johannesma, P. (1980b). Spectro-temporal receptive fields of auditory neurons in the grassfrog. III. Analysis of the stimulus-event relation for natural stimuli. *Biological Cybernetics*, 39:195–209.

Attias, H. and Schreiner, C. E. (1997). Temporal low-order statistics of natural sounds. In Mozer, M., Jordan, M., and Petsche, T., editors, *Advances in Neural Information Processing Systems*, volume 9, page 27. The MIT Press.

Azouz, R. and Gray, C. (1999). Cellular mechanisms contributing to response variability of cortical neurons *in vivo*. *Journal of Neuroscience*, 19:2209–2223.

Backoff, P. and Clopton, B. (1991). A spectrotemporal analysis of DCN single unit responses to wideband noise in guinea pig. *Hearing Research*, 53:28–40.

Boyd, S., Tang, Y., and Chua, L. (1983). Measuring Volterra kernels. *IEEE Transactions on Circuits and Systems*, 30:571–577.

Carney, L. and Friedman, M. (1998). Spatiotemporal tuning of low-frequency cells in the anteroventral cochlear nucleus. *Journal of Neuroscience*, 18:1096–1104.

Carney, L. and Yin, T. (1988). Temporal encoding of resonances by low-frequency auditory nerve fibers: Single-fiber responses and a population model. *Journal of Neurophysiology*, 60:1653–1677.

Chi, T.-S., Gao, Y., Guyton, M., and Shamma, S. (1999). Spectro-temporal modulation transfer functions and speech intelligibility (in press). *Journal of the Acoustical Society of America*.

Clopton, B. and Backoff, P. (1991). Spectrotemporal receptive fields of neurons in cochlear nucleus of guinea pig. *Hearing Research*, 52:329–344.

Cohen, L. (1995). *Time-Frequency Analysis*. Prentice-Hall.

de Boer, E. (1967). Correlation studies applied to the frequency resolution of the cochlea. *Journal of Auditory Research*, 7:209–217.

de Boer, E. and de Jongh, H. (1978). On cochlear encoding: Potentialities and limitations of the reverse-correlation technique. *Journal of the Acoustical Society of America*, 63:115–135.

deCharms, R., Blake, D., and Merzenich, M. (1998). Optimizing sound features for cortical neurons. *Science*, 280:1439–1443.

Depireux, D., Simon, J., Klein, D., and Shamma, S. (1998a). Representation of dynamic complex spectra in primary auditory cortex. *Abstracts of the Twenty-first ARO Mid-Winter Meeting*.

Depireux, D., Simon, J., and Shamma, S. (1998b). Measuring the dynamics of neural responses in primary auditory cortex. *Comments on Theoretical Biology*, 5:89–118.

Eggermont, J. (1993). Wiener and Volterra analyses applied to the auditory system. *Hearing Research*, 66:177–201.

Eggermont, J., Aertsen, A., Hermes, D., and Johannesma, P. (1981). Spectro-temporal characterization of auditory neurons: Redundant or necessary? *Hearing Research*, 5:109–121.

Eggermont, J., Aertsen, A., and Johannesma, P. (1983a). Prediction of the responses of auditory neurons in the midbrain of the grass frog based on the spectro-temporal receptive field. *Hearing Research*, 10:191–202.

Eggermont, J., Aertsen, A., and Johannesma, P. (1983b). Quantitative characterisation procedure for auditory neurons based on the spectro-temporal receptive field. *Hearing Research*, 10:167–190.

Eggermont, J., Johannesma, P., and Aertsen, A. (1983c). Reverse-correlation methods in auditory research. *Quarterly Review of Biophysics*, 16:341–414.

Eggermont, J. and Smith, G. (1990). Characterizing auditory neurons using the Wigner and Rihacek distributions: A comparison. *Journal of the Acoustical Society of America*, 87:246–259.

Epping, W. and Eggermont, J. (1985). Single-unit characteristics in the auditory midbrain of the immobilized grassfrog. *Hearing Research*, 18:223–243.

Escabí, M., Schreiner, C., and Miller, L. (1998). Dynamic time-frequency processing in the cat midbrain, thalamus, and auditory cortex: Spectro-temporal receptive fields obtained using dynamic ripple spectra. *Society for Neuroscience Abstracts*, 24:1879.

Hermes, D., Aertsen, A., Johannesma, P., and Eggermont, J. (1981). Spectro-temporal characteristics of single units in the midbrain of the lightly anaesthetised grass frog (Rana temporaria L.) investigated with noise stimuli. *Hearing Research*, 5:147–178.

Johnson, D. (1980). Applicability of white-noise nonlinear system analysis to the peripheral auditory system. *Journal of the Acoustical Society of America*, 68:876–884.

Kim, P. and Young, E. (1994). Comparative analysis of spectro-temporal receptive fields, reverse correlation functions, and frequency tuning curves of auditory-nerve fibers. *Journal of the Acoustical Society of America*, 95:410–422.

Korenberg, M. and Hunter, I. (1996). The identification of nonlinear biological systems: Volterra kernel approaches. *Annals of Biomedical Engineering*, 24:250–268.

Kowalski, N., Depireux, D., and Shamma, S. (1996a). Analysis of dynamic spectra in ferret primary auditory cortex. I. Characteristics of single-unit responses to moving ripple spectra. *Journal of Neurophysiology*, 76:3503–3523.

Kowalski, N., Depireux, D., and Shamma, S. (1996b). Analysis of dynamic spectra in ferret primary auditory cortex. II. Prediction of unit responses to arbitrary dynamic spectra. *Journal of Neurophysiology*, 76:3524–3534.

Kvale, M., Schreiner, C., and Bonham, B. (1998). Spectro-temporal and adaptive response to AM stimuli in the inferior colliculus. *Abstracts of the Twenty-first ARO Mid-Winter Meeting*.

Langner, G. (1992). Periodicity coding in the auditory system. *Hearing Research*, 60:115–142.

Lee, Y. and Schetzen, M. (1965). Measurement of the Wiener kernels of a non-linear system by crosscorrelation. *International Journal of Control*, 2:237–254.

Marmarelis, P. and Marmarelis, V. (1978). *Analysis of Physiological Systems: The White Noise Approach.* New York: Plenum.

Marmarelis, V. (1993). Identification of nonlinear biological systems using Laguerre expansions of kernels. *Annals of Biomedical Engineering*, 21:573–589.

Nelken, I., Kim, P., and Young, E. (1997). Linear and nonlinear spectral integration in type IV neurons of the dorsal cochlear nucleus. II. Predicting responses with the use of nonlinear models. *Journal of Neurophysiology*, 78:800–811.

Nelken, I., Rotman, Y., and Yosef, O. (1999). Responses of auditory-cortex neurons to structural features of natural sounds. *Nature*, 397:154–157.

Palm, G. and Popel, B. (1985). Volterra representation and Wiener-like identification of nonlinear systems: Scope and limitations. *Quarterly Review of Biophysics*, 18:135–164.

Papoulis, A. (1962). *The Fourier Integral and its Applications*. McGraw-Hill.

Pickles, J. (1988). *An Introduction to the Physiology of Hearing*. San Diego: Academic Press.

Ruggero, M. (1992). Physiology and coding of sound in the auditory nerve. In Popper, A. and Fay, R., editors, *The mammalian auditory pathway: Neurophysiology*, pages 34–93. New York: Springer-Verlag.

Schafer, M., Rubsamen, R., Dorrscheidt, G., and Knipschild, M. (1992). Setting complex tasks to single units in the avian forebrain. II: Do we really need natural stimuli to describe neuronal response characteristics? *Hearing Research*, 57:231–244.

Schreiner, C. and Calhoun, B. (1995). Spectral envelope coding in cat primary auditory cortex: Properties of ripple transfer functions. *Journal of Auditory Neuroscience*, 1:39–61.

Shamma, S. (1985). Speech processing in the auditory system I: The representation of speech sounds in the responses of the auditory nerve. *Journal of the Acoustical Society of America*, 78:1612–1621.

Shamma, S., Depireux, D., Klein, D., and Simon, J. (1998). Representation of dynamic broadband spectra in auditory cortex. *Society for Neuroscience Abstracts*, 24:402.

Shamma, S., Versnel, H., and Kowalski, N. (1995). Ripple analysis in the ferret primary auditory cortex. I. Response characteristics of single units to sinusoidally ripples spectra. *Journal of Auditory Neuroscience*, 1:233–254.

Smolders, J., Aertsen, A., and Johannesma, P. (1979). Neural representation of the acoustic biotope. *Biological Cybernetics*, 35:11–20.

Sutter, E. (1992). A deterministic approach to nonlinear systems analysis. In Pinter, R. and Nabet, B., editors, *Nonlinear Vision: Determination of Neural Receptive Fields, Function, and Networks*, pages 171–220. Boca Raton, FL: CRC Press.

Swerup, C. (1978). On the choice of noise for the analysis of the peripheral auditory system. *Biological Cybernetics*, 29:97–104.

Temchin, A., Recio, A., van Dijk, P., and Ruggero, M. (1995). Wiener-kernel analysis of chinchilla auditory-nerve responses to noise. *Abstracts of the Eighteenth ARO Mid-Winter Meeting*.

Theunissen, F., Sen, K., and Doupe, A. (1998). Characterizing non-linear encoding in the zebra finch auditory forebrain. *Society for Neuroscience Abstracts*, 24:402.

Valois, R. D. and Valois, K. D. (1990). *Spatial Vision*. New York: Oxford University Press.

van Dijk, P., Wit, H., and Segenhout, J. (1997). Dissecting the frog inner ear with Gaussian noise. I. Application of high-order wiener-kernel analysis. *Hearing Research*, 114:229–242.

Victor, J. (1979). Nonlinear systems analysis: Comparison of white noise and sum of sinusoids in a biological system. *Proceedings of the National Acadamy of Sciences, USA*, 76:996–998.

Victor, J. (1991). Asymptotic approach of generalized orthogonal functional expansions to Wiener kernels. *Annals of Biomedical Engineering*, 19:383–399.

Victor, J. and Knight, B. (1979). Nonlinear analysis with an arbitrary stimulus ensemble. *Quarterly of Applied Mathematics*, 37:113–136.

Victor, J. and Shapley, R. (1980). A method of nonlinear analysis in the frequency domain. *Biophysical Journal*, 29:459–484.

Volterra, V. (1930). *Theory of Functionals and of Integro-Differential Equations*. Dover, New York.

Wiener, N. (1958). *Nonlinear Problems in Random Theory*. New York: John Wiley.

Yamada, W. and Lewis, E. (1999). Predicting the temporal respsonses of non-phase-locking bullfrog auditory units to complex acoustic waveforms. *Hearing Research*, 130:155–170.

Yamada, W., Wolodkin, G., Lewis, E., and Henry, K. (1997). Wiener kernel analysis and the singular value decomposition. In Lewis, E., editor, *Diversity in Auditory Mechanics*, pages 111–118. Singapore: World Scientific.

Yeshurun, Y., Wollberg, Z., Dyn, N., and Allon, N. (1985). Identification of MGB cells by Volterra kernels. I. Prediction of responses to species specific vocalizations. *Biological Cybernetics*, 51:383–390.