Biol. Cybern. 59, 91–102 (1988)

Biological Cybernetics

Multinomial Pulse-Number Distributions for Neural Spikes in Primary Auditory Fibers: Theory

M. C. Teich and R. G. Turcott

Center for Telecommunications Research, Department of Electrical Engineering, and Fowler Memorial Laboratory, Department of Otolaryngology, Columbia University, New York, NY 10027, USA

Abstract. We previously reported experimental short- and long-counting-time pulse-number distributions (PND's) for the neural spike train in cat primary auditory nerve fibers. Data were obtained for spontaneous activity, pure-tone stimuli with a wide range of frequencies and intensity levels, and Gaussian noise. The irregular shapes of the PND's are an indication of the presence of spike clusters of various sizes in the neural impulse train. We develop a family of theoretical cluster counting distributions and examine their suitability for describing the experimental PND's. The reduced-quintinomial distribution provides theoretical results that describe the characteristics of the PND's quite well, accounting for the smooth or scalloped behavior of short-counting-time data, the jagged nature of long-counting-time data, and the Poisson-like character of very-short-counting-time data. This family of distributions admits values for the spike-number mean-to-variance ratio that are independent of stimulus level, in agreement with experimental observation. A number of procedures for fitting the theoretical distributions to the experimental PND's are studied. These include the use of a minimum mean-square error criterion, the factorial moments of the data, and the discrete Fourier transform of the PND. The first of these techniques appears to be the most useful.

1 Introduction

We recently reported the results of a series of nerve-spike experiments carried out on cat primary auditory fibers (Teich and Khanna 1985; Teich 1988). Particular attention was focused on the pulse-number distribution (PND), which represents the relative frequency $p(n, T)$ for the occurrence of n spikes in an arbitrary counting time T. Two counting (observation) windows T were used in collecting the data: a short counting window ($T = 51.2$ ms) and a long counting window ($T = 204.8$ ms). The appearance of the experimental short- and long-counting-time PND's were quite different even though both were simultaneously constructed from the same spike trains. The short-counting-time PND's were relatively smooth in comparison with their long-counting-time counterparts, which generally had an irregular or jagged appearance (multiple sub-peaks and valleys). Scalloping of this kind in the PND can be an indication of the presence of spike clusters in the neural pulse train. Such clusters are present in the cat optic-nerve pulse train at low light levels (Barlow et al. 1971; Mastronarde 1983; Saleh and Teich 1985). Westerman and Smith (1984) collected auditory-nerve PND's with a very-short counting window ($T \approx 1$ ms). These data displayed a mean-to-variance ratio (MVR) of order unity, as for the Poisson distribution.

The object of this paper is to investigate the suitability of a theoretical family of counting distributions for describing the experimentally observed PND's. These theoretical distributions are derived from the multinomial family, of which the binomial is the most widely known. In such distributions, each event of a primary process (which may be derived, for example, from level crossings or the onset of relaxation-type oscillations) generates a cluster of secondary events. The number of events in each cluster is specified by a particular probability law, i.e., probabilities for zeros, singlets, pairs, triplets, etc. In the general case, the secondary events are splayed out in time with respect to the primary event. These distributions are closely associated with those derivable from cluster point processes (see, e.g., Neyman and Scott 1958).

The reduced-quintinomial probability distribution, described in Sect. 2, generates PND's with behavior that is in good accord with experimental

observations. The procedure used to fit the reduced-quintinomial distribution to the experimental PND's is presented in Sect. 3. The parameters of the theoretical distribution that most closely describe experimental PND's are presented and discussed in Sect. 4. The conclusions are provided in Sect. 5.

2 The Reduced-Quintinomial Probability Distribution

It is well known that the binomial probability distribution describes the results of a probabilistic experiment in which each of L trials has two possible outcomes. An example of such an experiment is L tosses of a coin, where the result is either heads or tails on any given toss. The binomial distribution $p_b(n, L)$ represents the probability of observing a given outcome (heads or tails) n times out of L tosses.

In a similar way, the quintinomial distribution provides the probability of observing various combinations of five possible outcomes of L repeated trials (Thomas 1971; Scheaffer and McClave 1986). In applying the quintinomial to the neural spike train, we consider each trial as capable of generating a cluster of 0, 1, 2, 3, or 4 spikes; these are the five possible outcomes. This is illustrated in Fig. 1, where L trial occurrences are shown on the first line and the (randomly delayed) spike clusters associated with each occurrence appear on the second line. In this hypothetical example, the 10 repeated trials (L) give rise to the spike-cluster sequence $\{2, 1, 2, 3, 4, 3, 1, 2, 0, 3\}$.

The occurrences of spike clusters has been reported by Kiang et al. (1965) for a sequence of click stimuli. Typically, a single click elicits somewhere between 0 and 4 spikes. In the continuous-stimulus case, the consecutive trials provide a shorthand way to characterize successive impulse responses of the auditory system which occur at its own natural period. For simplicity of illustration in Fig. 1, the trials are shown as a deterministic sequence of events and the clusters associated with each trial are shown as occurring sequentially. This idealization, which enables us to present the notion clearly in graphical form, is not required in the mathematical model. Indeed, the occurrence times of the trials are expected to exhibit randomness and the clusters may well overlap so that the spikes residing in different clusters will interdigitate with each other.

The use of the quintinomial, rather than the binomial, trinomial, or quadrinomial results from the following reasoning. The higher the order of the reduced-multinomial distribution, the more jagged the behavior that is possible. All of the short-counting-time PND's (which are either smooth or exhibit spike pairs that are evidenced by enhanced probabilities at even numbers of counts; see Teich and Khanna 1985) can, in fact, be described by the reduced-trinomial distribution. This suggests that a counting time of 50 ms is insufficient to easily capture clusters of three or four spikes. However, the long-counting-time PND's are not well fit by the reduced-trinomial and reduced-quadrinomial distributions. Apparently these PND's can capture clusters of three or four spikes (as well as smaller clusters). This encouraged us to use the reduced-quintinomial which admits spike clusters up to size 4 which is the maximum cluster size observed by Kiang et al. (1965) for click stimuli.

If N_i is a random variable representing the number of spikes i generated on a single trial, the quintinomial

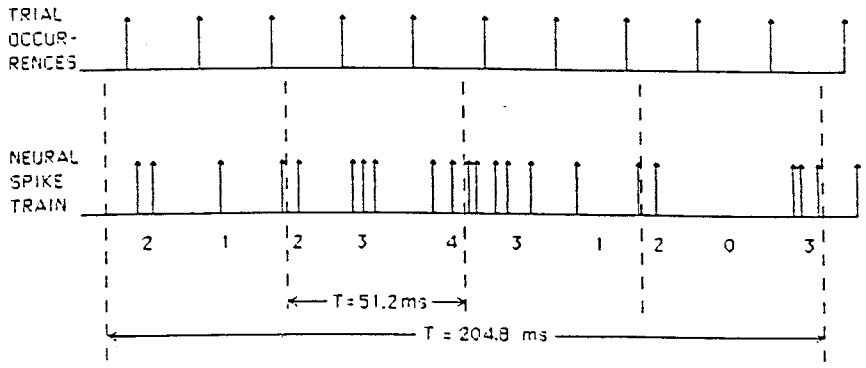

Fig. 1. L trial occurrences are shown on the first line. The spike clusters associated with each occurrence appear on the second line. In this idealized case, the trials are considered to be a deterministic sequence of events and the clusters associated with each trial are taken to occur sequentially. Realistically, the occurrence times of the trials are expected to exhibit some degree of randomness and the clusters may interdigitate. The 10 repeated trials (L) give rise to a hypothetical sequence of spike clusters with the following sizes: $\{2, 1, 2, 3, 4, 3, 1, 2, 0, 3\}$. Both short ($T \approx 50$ ms) and long ($T \approx 200$ ms) counting windows are shown. In recording a short-counting-time PND more of the spike clusters are cut apart than in recording a long-counting-time PND. In the limit in which the window is so short ($T \approx 1$ ms) that only zero spikes or one spike may be counted, all of the clusters are cut apart into zeros and singlets. This leads to a Bernoulli (Poisson-like) counting distribution with a mean-to-variance ratio of unity, as experimentally observed by Westerman and Smith (1984)

distribution expresses the joint probability $\Pr(N_0=n_0, N_1=n_1, \ldots, N_4=n_4)$ of observing zero spikes n_0 times, one spike n_1 times, etc. Since each trial must give rise to one of the permissible outcomes N_i, we have $\Sigma_i n_i = L$ (the analog for the binomial is that the number of heads plus the number of tails equals the number of trials L).

However, individual spike clusters are not observed; rather the observable is the total number of spikes N measured in the counting time T by the PND, as seen in Fig. 1. Thus, the five random variables N_i of the quintinomial distribution must be mapped onto a single random variable N representing the total number of spikes observed in L trials. This mapping is accomplished by considering the outcome of each of the L trials as contributing i counts to the total N, so that $n=\Sigma_i i n_i = 0n_0+1n_1+2n_2+3n_3+4n_4$.

For a given number of trials L and a given value of the spike number n, it turns out that specifying the values of N_2, N_3, and N_4 for a quintinomial distribution uniquely indentifies a combination of outcomes since $L=n_0+n_1+n_2+n_3+n_4$ (the total number of outcomes equals the number of trials) and $n=(n_1+2n_2+3n_3+4n_4)$ (the total number of spikes generated is equal to n). The reduced-quintinomial probability distribution $p_q(n, L)$ is then obtained by summing the quintinomial across all configurations that satisfy the condition $2n_2+3n_3+4n_4 \leqq n$, i.e.

$$p_q(n,L)=\sum_{n_4}\sum_{n_3}\sum_{n_2}\frac{L!\,\pi_0^{L-n+n_2+2n_3+3n_4}\pi_1^{n-2n_2-3n_3-4n_4}\pi_2^{n_2}\pi_3^{n_3}\pi_4^{n_4}}{(L-n+n_2+2n_3+3n_4)!\,(n-2n_2-3n_3-4n_4)!\,n_2!\,n_3!\,n_4!}, \qquad (1)$$

where n_4 is summed from 0 to $n/4$, n_3 is summed from 0 to $n/3-4n_4/3$, and n_2 is summed from 0 to $n/2-2n_4-3n_3/2$.

Here it suffices to consider the results for the reduced-quintinomial spike-number mean and variance which are, respectively,

$$\langle n\rangle = L(\pi_1+2\pi_2+3\pi_3+4\pi_4), \qquad (2)$$

$$\mathrm{Var}(n)=L[\pi_1+4\pi_2+9\pi_3+16\pi_4-(\pi_1+2\pi_2+3\pi_3+4\pi_4)^2], \qquad (3)$$

where π_i is the probability of observing outcome i for a given trial. These expressions are reaildy obtained from the moment-generating function. For this class of models the theoretical mean-to-variance ratio R is

$$R\equiv\langle n\rangle/\mathrm{Var}(n), \qquad (4)$$

which is independent of L. Thus, if stimulus level were encoded into the number of trials L, R would be independent of stimulus level. This family of distributions therefore admits behavior that is in agreement with experiment [the mean-to-variance ratio is relatively independent of stimulus level for all of the PND's that have been measured in primary auditory fibers (Teich and Khanna 1985; Teich 1988; Westerman and Smith 1984)].

Both short and long counting windows are shown in Fig. 1. It is apparent that the short window cuts apart more of the spike clusters than does the long window. Indeed, each of the four short windows illustrated in Fig. 1 cuts spike clusters apart. In the limit in which the window is so short that only zero or one spike may be counted, all of the clusters are cut apart into zeros and singlets. This leads to the Bernoulli distribution which, with π_1 small, gives a mean-to-variance ratio of unity and Poisson-like behavior. This is indeed the result that was experimentally observed by Westerman and Smith (1984). The presence of phase locking (Rose et al. 1967) does not affect the PND to a great extent (as time jitter and time quantization do not affect it). This is because the details of precise spike occurrence times on a short time scale do not affect the spike count, as can be seen from Fig. 1.

In the special case of deterministic clusters, each trial would always produce the same outcome and a given cluster size would always appear, i.e., $\pi_j=1$ with $\pi_{i\neq j}=0$. For example, if triplets are consistently produced on every trial then $\pi_3=1$ and $\pi_0=\pi_1=\pi_2=\pi_4=0$. Equation (3) then gives zero variance, as expected for a deterministic situation.

In Figs. 2 and 3, we display a number of reduced-quintinomial distributions, $p_q(n, L)$ vs n. Various values of L and $\pi_i=\pi_0, \pi_1, \pi_2, \pi_3, \pi_4$, were specially selected for the purposes of illustration and are indicated at the side of each of the panels. These plots reveal a number of characteristics of these distributions that are neither obvious from the expression for $p_q(n, L)$ in (1) nor from the underlying probability model: (*a*) Although the various distributions in Fig. 2 have quite different appearances, all have the same mean ($\langle n\rangle=12$); (*b*) In the deterministic limit ($\pi_j=1, \pi_{i\neq j}=0$), different sets of parameters are capable of generating distributions that are identical, as illustrated in the topmost row of Fig. 2. However in the probabilistic case, different sets of parameters will always generate distinct distributions even though they may have similar appearance; (*c*) If only spike doublets and quadruplets are possible, the result is always an even spike count. Similarly, the exclusive presence of singlets and triplets leads to an even spike count if L is even (see Fig. 2e, f and h); (*d*) If only zeros and a particular cluster size i are present, the reduced-quintinomial distribution displays nonzero probability only for counts that are integral multiples

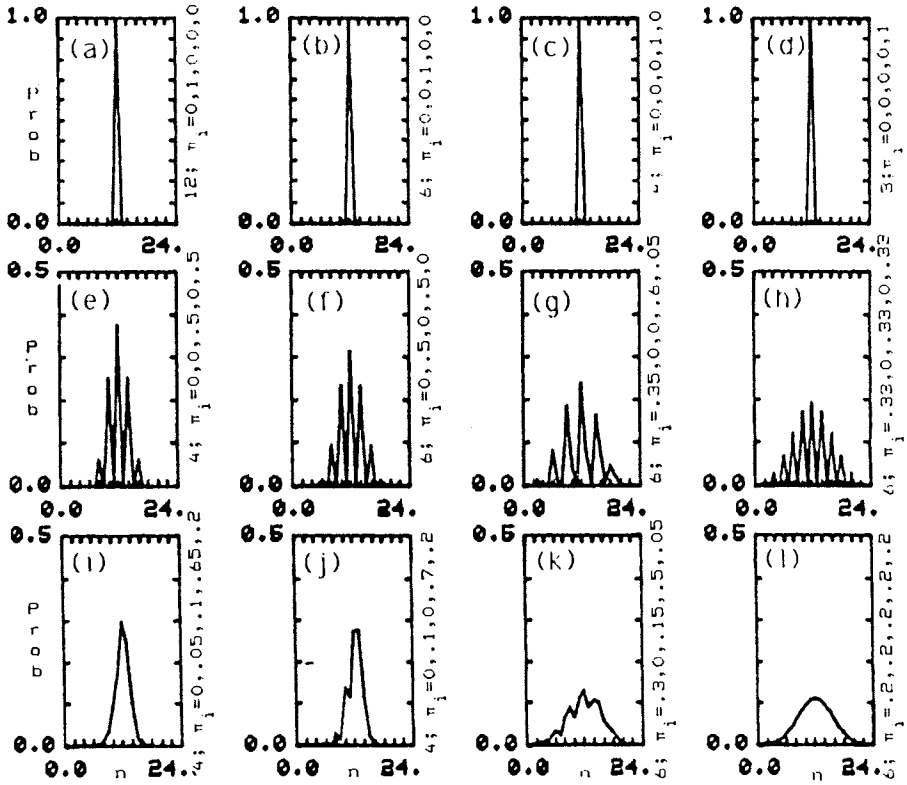

Fig. 2a–l. Theoretical reduced-quintinomial distributions, $p_q(n, L)$ vs. n. Various values of L and $\pi_i = \pi_0$, π_1, π_2, π_3, π_4 (indicated beside the individual panels) were specially selected to illustrate the range of behavior exhibited by this distribution. It can display smooth, scalloped, or jagged behavior. In all cases, the mean count $\langle n \rangle = 12$

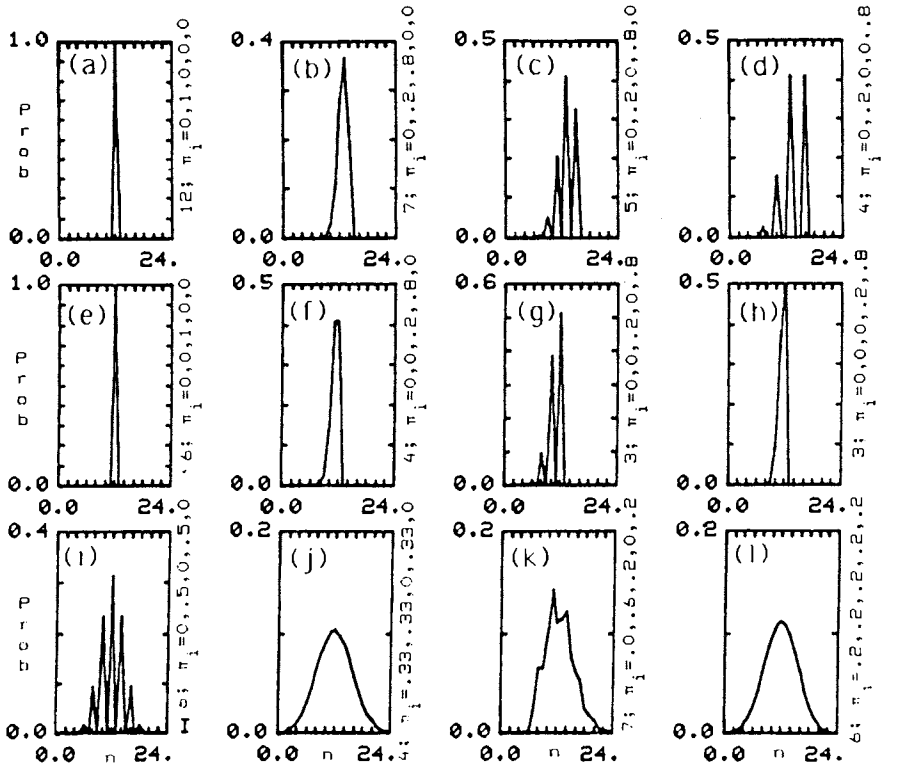

Fig. 3a–l. Additional reduced-quintinomial distributions, $p_q(n, L)$ vs. n. Different values of L and $\pi_i = \pi_0$, π_1, π_2, π_3, π_4 (indicated beside the individual panels) were selected to further illustrate the range of behavior exhibited by this distribution

of i. For example, if only zeros and triplets are permitted, $p_q(n, L)$ displays nonzero values only at $n=3$, 6, 9, 12, etc. This situation is shown approximately in Fig. 2g; in this particular example, however, the small admixture of quadruplets permits small probabilities at intermediate values of n; and (e) The admixture of substantial probabilities for even and odd cluster sizes reduces the extreme scalloping of the distributions, as is evident from the third row of Fig. 2. Equal probabilities for every cluster size leads to the smooth distribution shown in Fig. 2l.

The role of various combinations of spike cluster sizes on the reduced-quintinomial distribution is further clarified in Fig. 3. Scanning across the first row first sees singlets admitted, then singlets and doublets, followed by singlets and triplets, and finally singlets and quadruplets. The variance increases as we pass from left to right. A similar pattern is followed in the second row where distributions for various combinations of doublets, triplets and quadruplets are displayed. Other instructive combinations are presented in the third row.

It is clear from these results that irregular pulse-number distributions can (but need not) be generated by the reduced-quintinomial model, depending on the cluster probabilities and L. Furthermore, different shaped distributions will be produced as the value of T is changed and the clusters are cut apart to different degrees. For *very* small T ($\lesssim 1$ ms) the reduced-quintinomial reduces to a Poisson-like distribution, for small T (≈ 50 ms) it reduces to a binomial-like or trinomial-like distribution (the latter if spike pairs are permitted), whereas for larger T ($\gtrsim 200$ ms) it can take on a jagged character.

3 Procedures for Fitting Reduced-Quintinomial Distributions to Experimental PND's

A minimum-mean-squared error (MMSE) criterion was used to numerically estimate the parameters of the reduced-quintinomial distribution that best fits a given experimental PND. Briefly, the allowed cluster probabilities π_0 through π_4 were quantized to integer multiples of 0.1, from 0 to 1.0. Theoretical reduced-quintinomial distributions, with all reasonable values of these cluster probabilities and with the number of trials L ranging from 1 to 15, were computed with extended precision (15 decimal digits) and compared with each PND. This range of L was found to be adequate for achieving good fits to the data. Since the experimental uncertainty of the data is lowest where the sample size is largest, the theoretical distributions were fit across one standard deviation of the PND to either side of the mean rather than across its entire width. In this way the high-uncertainty tails of the distribution were ignored and the fit was based on the central portion of the PND where the scalloping was most pronounced and the spike-cluster information richest. Typically, the mean of the best-fitting reduced-quintinomial was within half a count of the PND mean.

There are other ways of fitting the theoretical distributions to the PND's, and two of these deserve discussion. In principle, one way in which this can be achieved is by analytically matching the first five theoretical and experimental factorial moments (Saleh 1978; Olkin et al. 1980). Because a fitting procedure of this kind relies on higher moments, however, it emphasizes the tails of the PND where accuracy is problematical in nerve-fiber experiments. The MMSE numerical fitting procedure, which relies on only one standard deviation to either side of the mean, turns out to be superior with data of limited accuracy.

Another estimation technique is suggested by the scalloped appearance of the PND's, in particular the periodically spaced maxima of the short-counting time data for units displaying spike pairs. This suggests that frequency analysis may be useful for gaining insight into the structure of the PND (Cox and Lewis 1966). Indeed, complete information about a set of discrete samples is contained in its discrete Fourier transform (DFT). Since there is a unique set of underlying parameters associated with each reduced-quintinomial distribution, it follows that these parameters can also be obtained from the DFT of the distribution. The DFT is calculated by means of the fast Fourier transform (FFT) method (Schwartz and Shaw 1975). The procedure we used, i.e., calculating the DFT of an experimental PND (Cox and Lewis 1966), should not be confused with the periodogram (Schwartz and Shaw 1975).

Typical theoretical frequency-analysis results are illustrated in Fig. 4. In Fig. 4a we present a reduced-quintinomial distribution, $p_q(n, L)$ vs count number n, with parameters $L=6$ and $\pi_i=\{0.35, 0.0, 0.0, 0.6, 0.05\}$. This is an expanded plot of the distribution represented in Fig. 2g. The cluster probability for triplets was purposely chosen to be quite high. The DFT magnitude of $p_q(n, L)$ is shown in Fig. 4b. It readily displays the dominant frequency component of this distribution, i.e. a period of 3 counts (lower abscissa) or equivalently $\omega=2\pi/3$ radians/count (upper abscissa). The zero-frequency component of the DFT (at dc) results from the rather broad envelope of the reduced-quintinomial distribution, about which the probability values fluctuate. The maximum value of the DFT is unity since the reduced-quintinomial distribution is normalized.

The envelope of the distribution illustrated in Fig. 4a has an approximately binomial shape. The

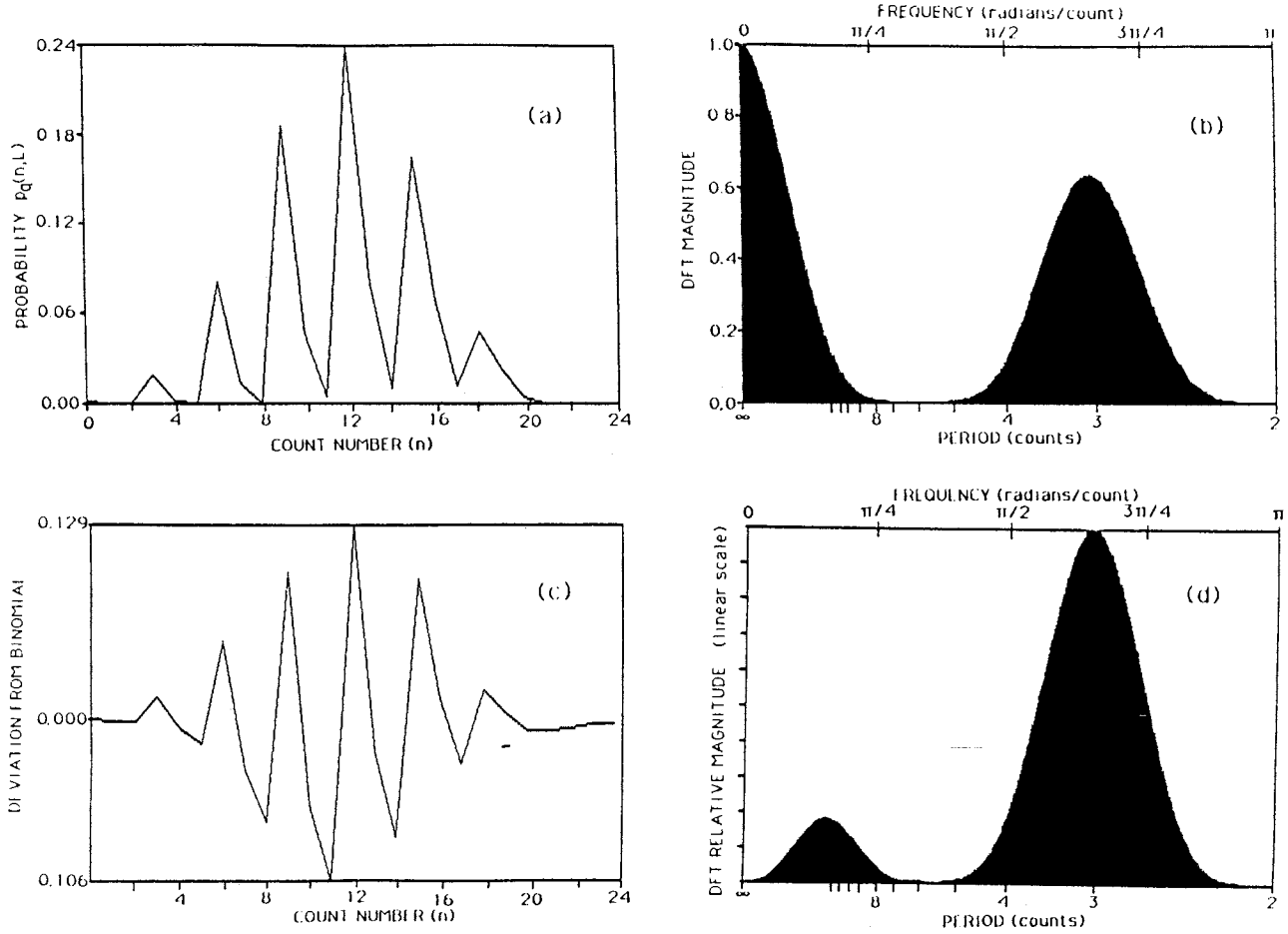

Fig. 4. a Theoretical reduced-quintinomial distribution, $p_q(n, L)$ vs. count number n, with parameters $L=6$, and $\pi_i=\{0.35, 0.0, 0.0, 0.6, 0.05\}$. The cluster probability for triplets is quite high. **b** The DFT magnitude of $p_q(n, L)$. The dominant frequency component with a period of 3 counts (lower abscissa), or equivalently $\omega=2\pi/3$ radians/count (upper abscissa), is clearly visible. The low-frequency component of the DFT results from the rather broad binomial-like envelope of the reduced-quintinomial distribution, about which the probability fluctuates. The maximum value of the DFT is unity since the reduced-quintinomial distribution is normalized. **c** Deviation of the reduced-quintinomial from a binomial distribution. The filtering leaves the probability fluctuations intact but removes the smooth envelope of the PND. **d** DFT (relative scale) of the binomial-filtered distribution in **c**. Comparison with **b** reveals that the low-frequency component is removed and the component with a period of three counts is enhanced. Binomial filtering can therefore be useful for increasing the prominence of the cluster components

scallops of the PND, which carry information about the spike clusters, can be accentuated by subtracting a binomial envelope from the PND (binomial filtering). This procedure converts the PND into a zero-sum distribution with no dc (zero-frequency) component, thereby increasing the prominence of the cluster components.

Short-counting-time PND's, by virtue of their tendency to minimize the observation of clusters, were used to estimate the appropriate binomial distribution to be subtracted from the long-counting-time PND's. Most short-counting-time PND's exhibit a lack of structure and are adequately approximated by the binomial distribution, which has count mean and variance given by

$$M_{50}=L_{50}\pi_1 \tag{5}$$

and

$$V_{50}=L_{50}\pi_1(1-\pi_1). \tag{6}$$

These equations can be inverted to provide π_1 and L_{50} in terms of the short-counting-time empirical spike mean and variance, i.e.,

$$\pi_1=1-V_{50}/M_{50}, \tag{7}$$

$$L_{50}=M_{50}/\pi_1. \tag{8}$$

The number of trials L_{50} is then multiplied by 4 to obtain the appropriate long-counting-time parameter L_{200}. The probability of a spike π_1 is taken to be independent of the counting time T. The long-counting-time binomial distribution generated by these parameters was then subtracted from the $T=204.8$ ms PND's, and the result was Fourier trans-

formed to provide the filtered DFT's. The subscripts 50 and 200 stand for $T \approx 50$ ms and $T \approx 200$ ms, respectively.

The deviation of the counting distribution from the binomial is shown in Fig. 4c. This processing eliminates the dc (zero-frequency) component in the DFT, as is evident in Fig. 4d, which is the DFT of Fig. 4c. Binomial filtering can therefore be used to increase the prominence of the cluster components of interest. Indeed, comparing Fig. 4d with Fig. 4b shows that the component with a period of three counts is emphasized and the zero-frequency component is removed. However, this nonlinear filtering technique introduces a new peak near 12 counts. Although the DFT technique is of interest from a theoretical point of view, we have concluded that it is useful principally when scalloping is plainly apparent by visual inspection of the PND. The MMSE fitting procedure discussed above turns out to operate in a more effective manner.

4 Results

All of the long-counting-time experimental PND's display jagged behavior, regardless of the characteristics of the unit or the stimulus conditions (Teich 1988). We present a detailed comparison of the best fitting reduced-quintinomial theoretical distributions with the experimental PND's for two units. One of these units (No. 8/52304.PN, 5/5/82), shown in Figs. 5 and 6, did not display visually apparent spike pairs in the short-counting-time PND's. The other unit (No. 44/81426.PN, 2/17/82), shown in Figs. 7 and 8, did display such spike pairs in the short-counting-time PND's.

The solid curves in Figs. 5 and 7 represent experimental short-counting-time PND's drawn from Teich and Khanna (1985) whereas the solid curves in Figs. 6 and 8 are long-counting-time PND's drawn from Teich (1988). The data in Figs. 5 and 6 are from a *high-*

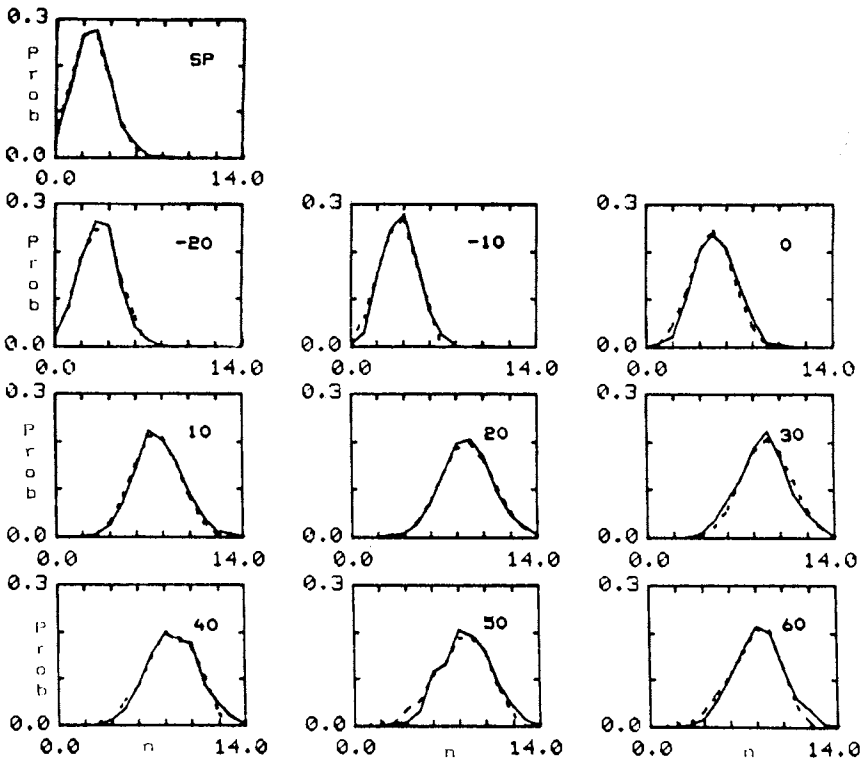

Fig. 5. The solid curves represent experimental short-counting-time ($T = 51.2$ ms) PND's for a *high-spontaneous* (53.7 counts/s), *high-frequency* (CF = 6113 Hz) unit (No. 8/52304.PN, 5/5/82) with a threshold of about 20 dB SPL. The data are drawn from Teich and Khanna (1985). The corresponding long-counting-time PND's for this unit are shown in Fig. 6. The upper-left element in the figure represents the PND collected in the absence of external stimulus (spontaneous PND, denoted SP). The other elements in the figure represent PND's recorded with a pure-tone stimulus applied at the CF; the stimulus level (in dB:re FTC) is indicated in each element. The relatively smooth character of the distributions means that this unit did not display visually apparent spike pairs in the short-counting-time PND's (Teich and Khanna 1985). The best-fitting reduced-quintinomial distributions are shown as the dotted curves. The reduced-quintinomial distribution is capable of taking on the smooth bell-shaped forms shown here for appropriate parameter values. The fits of the theory to the data are quite good

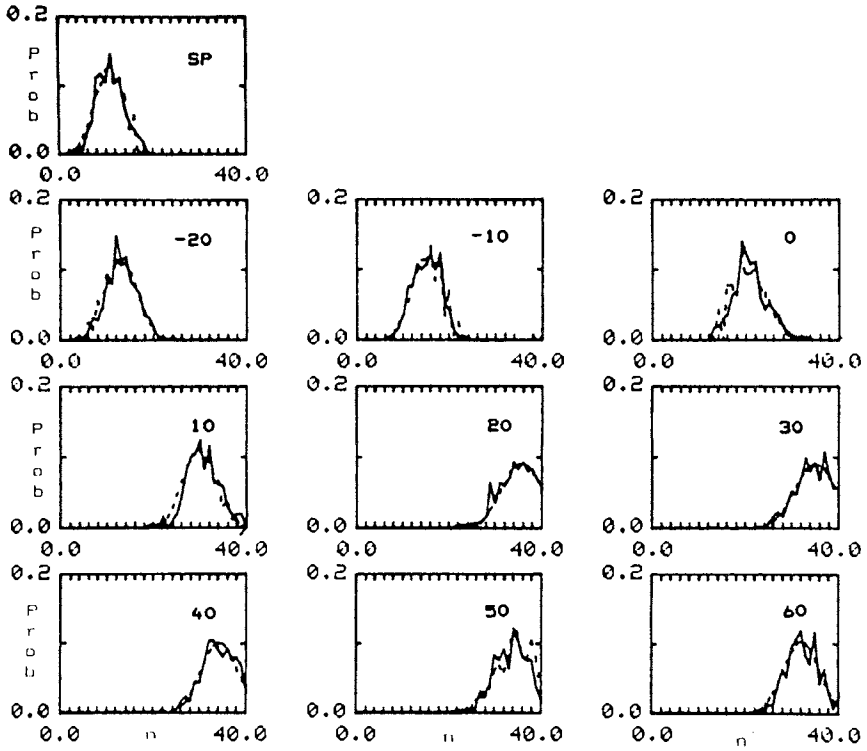

Fig. 6. The solid curves represent experimental long-counting-time ($T = 204.8$ ms) PND's for a *high-spontaneous* (53.7 counts/s), *high-frequency* (CF = 6113 Hz) unit (No.8/52304.PN, 5/5/82) with a threshold of about 20 dB:re SPL. The data are drawn from Teich (1988). The corresponding short-counting-time PND's for this unit are shown in Fig. 5; this unit did not display visually apparent spike pairs. Nevertheless, the long-counting-time PND's exhibit irregular shapes indicating the presence of spike clusters. The best-fitting reduced-quintinomial distributions (dotted curves) take on the jagged character of the data for appropriate parameter values. The fits of the theory to the data are reasonably good. Other counting distributions commonly used in auditory modeling, such as the binomial or the dead-time-modified Poisson distribution, cannot exhibit such inflections or local extrema

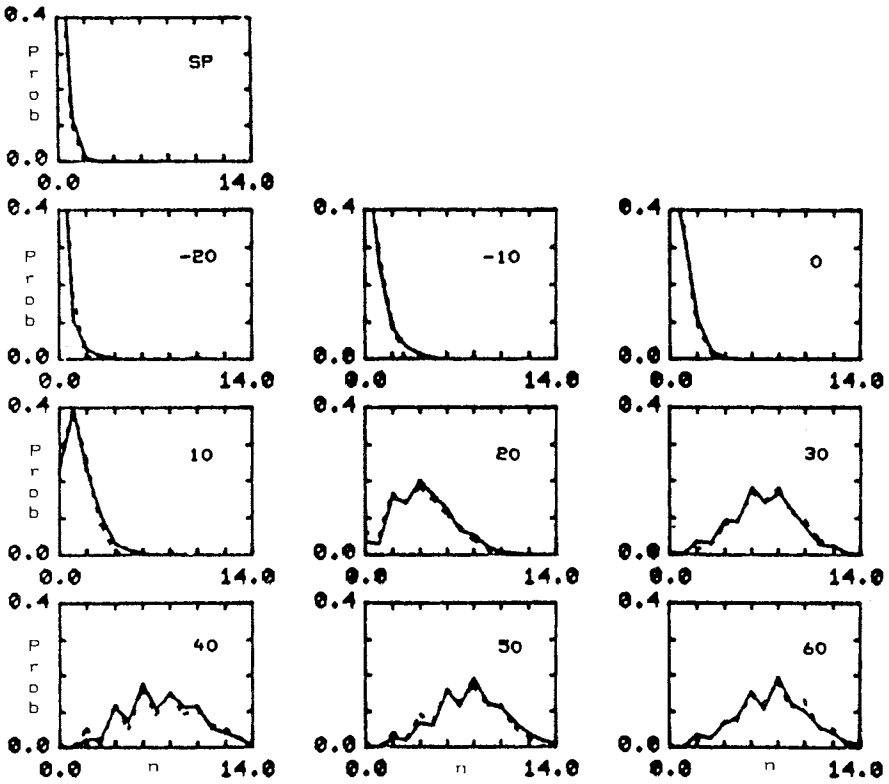

Fig. 7. The solid curves represent experimental short-counting-time ($T = 51.2$ ms) PND's for a *medium-spontaneous* (4.1 counts/s), *high-frequency* (CF = 7227 Hz) unit (No. 44/81426.PN, 2/17/82) with a threshold of about 23 dB SPL. The data are drawn from Teich and Khanna (1985). The short-counting-time PND's indicate the presence of spike pairs (even spike numbers are enhanced). The corresponding long-counting-time PND's for this unit are shown in Fig. 8. The best-fitting reduced-quintinomial distributions (dotted curves) take on the scalloped character of the data for appropriate parameter values. The fits of the theory to the data are excellent. Counting distributions commonly used in auditory modeling, such as the binomial or the dead-time-modified Poisson distribution, do not exhibit scallops

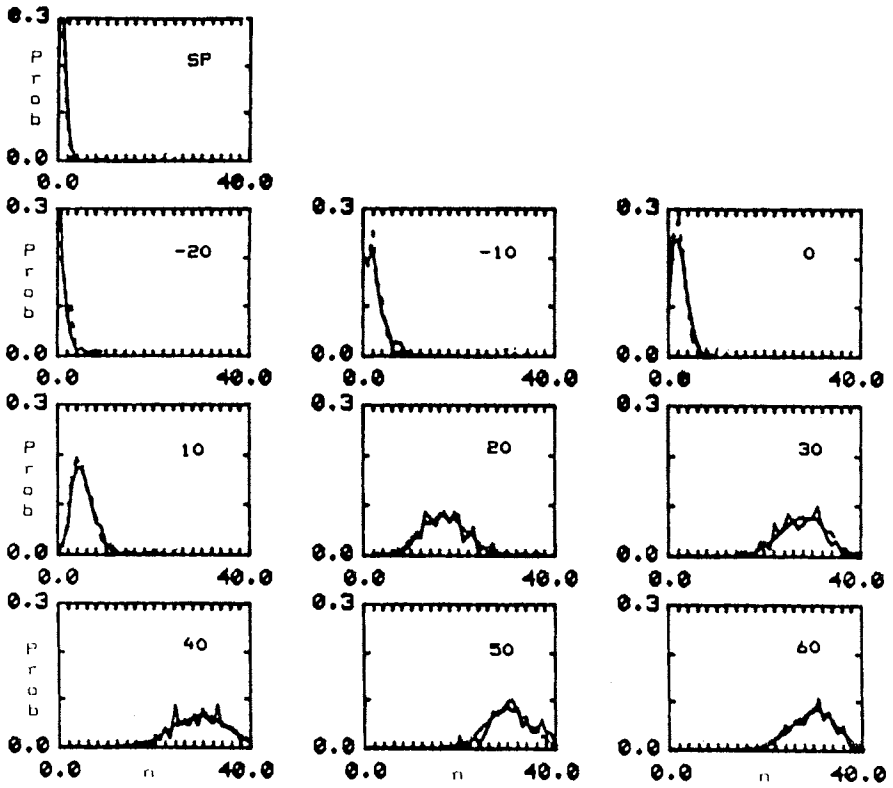

Fig. 8. The solid curves represent experimental long-counting-time ($T = 204.8$ ms) PND's for a *medium-spontaneous* (4.1 counts/s), *high-frequency* (CF = 7227 Hz) unit (No. 44/81426.PN, 2/17/82) with a threshold of about 23 dB SPL. The data are drawn from Teich (1988). The corresponding short-counting-time PND's for this unit are shown in Fig. 7. The spike pairs evident in the short-counting-time PND's are masked by the presence of other clusters in the long-counting-time data. The best-fitting reduced-quintinomial distributions (dotted curves) take on some of the jagged character of the data for appropriate parameter values. The fits of the theory to the data are satisfactory

spontaneous (53.7 counts/s), *high-frequency* (characteristic frequency CF = 6113 Hz) unit with a threshold of about 20 dB sound pressure level (SPL). The data in Figs. 7 and 8 data are from a *medium-spontaneous* (4.1 counts/s), *high-frequency* (CF = 7227 Hz) unit with a threshold of about 23 dB SPL.

The upper-left element in each figure represents the PND collected in the absence of external stimulus (spontaneous PND, denoted SP) and the best-fitting reduced-quintinomial distribution (dotted curve). The other elements represent PND's recorded with a pure-tone stimulus applied at the CF. The stimulus level (in dB: re FTC, i.e., dB with respect to the unit rate threshold) is indicated in each element of the figure. The same procedure was carried out for PND's generated by Gaussian noise, and by pure tones above and below CF. We do not explicitly present these data since the results shown in Figs. 5–8 are representative for these stimuli as well. Furthermore, the fits of reduced-quintinomial distributions to PND's collected from other units, including low-frequency units, are similar to those displayed in Figs. 5–8.

The fits of theory to the data are reasonably good; the reduced-quintinomial distribution follows many of the inflections and local extrema present in the data. It can assume a broad variety of shapes depending on its parameters (including a smooth bell-shaped distribution). Other counting distributions commonly used in auditory modeling, such as the binomial or the dead-time-modified Poisson distribution, cannot display inflections or local extrema.

Two observations emerge from an examination of the parameters of the reduced-quintinomial that best describe the experimental PND's represented in Figs. 5–8. The cluster probabilities for triplets and quadruplets, π_3 and π_4, as well as the number of trials L, generally increase with increasing counting time and increasing stimulus level. The increase in π_3 and π_4 with counting time can be understood in terms of the presentation in Fig. 1. In particular, an example is provided in which a cluster of 4 spikes is split between two adjacent 50 ms counting periods. Large clusters are therefore more likely to appear in long-counting-time PND's, explaining the increase of π_3 and π_4 with increasing counting time. This suggests, by the way, that the reduced-quintinomial parameters that describe the long-counting-time PND's more accurately portray the character of the underlying neural spike train. The increase in π_3 and π_4 with stimulus level reflects the fact that spike rate generally increases with stimulus level, in accordance with (2). The weighting factors associated with π_3 and π_4 demonstrate that their increase with stimulus level can maintain good fits of theory to data.

For the long-counting-time data, the best fitting number of trials L generally takes on values between 5

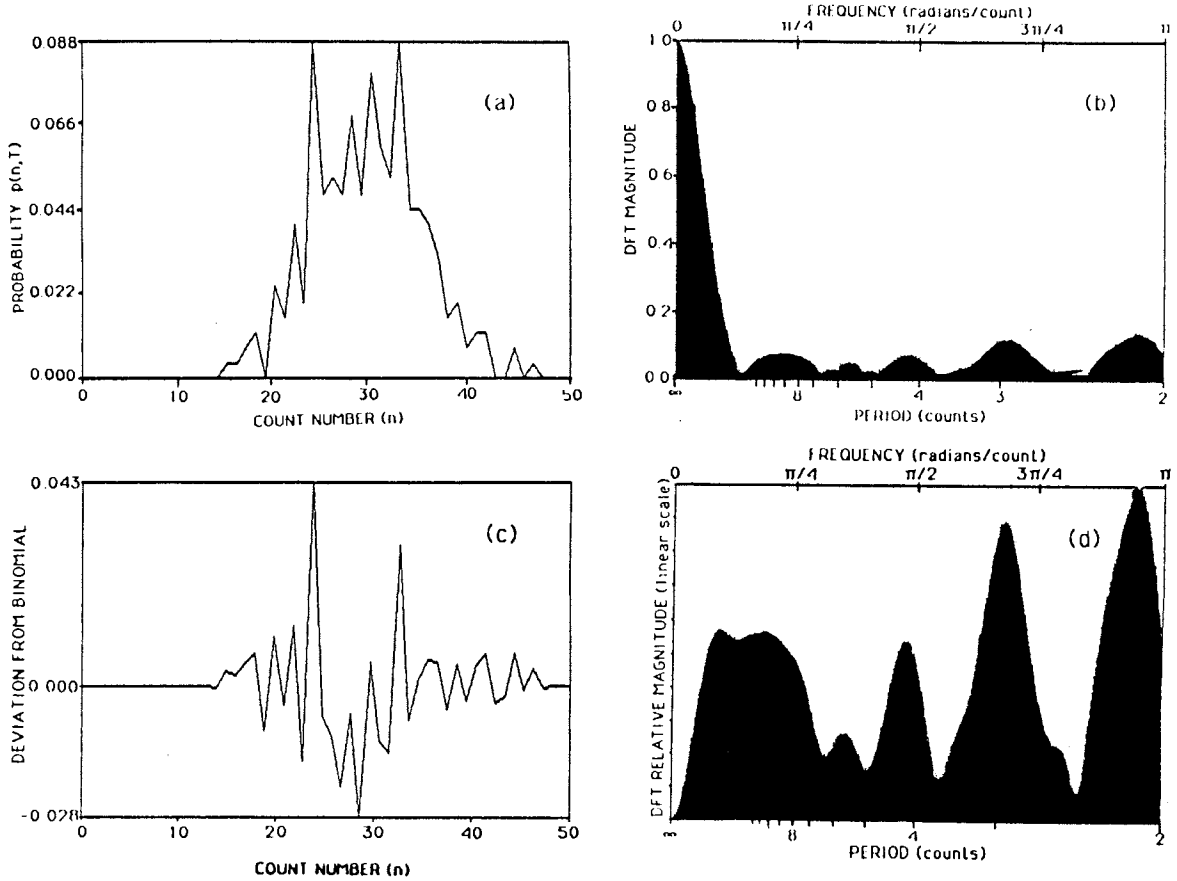

Fig. 9. a Long-counting-time ($T = 204.8$ ms) PND for Unit No. 44/81426.PN stimulated by a pure-tone at 7227 Hz (CF) at 40 dB:re FTC. **b** The DFT magnitude of the PND shown in **a**. Several dominant frequency components are visible in the DFT. The low-frequency component results from the rather broad binomial-like envelope of the PND, about which the probability fluctuates. The maximum value is unity since the reduced-quintinomial distribution is normalized. **c** Deviation of the PND from a binomial distribution. The filtering leaves the probability fluctuations intact but removes the smooth envelope of the PND. **d** DFT (relative scale) of the binomial-filtered distribution in **c**. Comparison with **b** reveals that the low-frequency component is removed and that the strong frequency components at periods of 2, 3, 4, and 9 counts are enhanced

and 12, whereas for the short-counting-time data the average value of L is about 4. The optimally fitted value of L is rarely at either extreme of the range that was tested ($1 \leqq L \leqq 15$). At first thought, it might be expected that L should increase in approximately linear fashion with T because the longer counting time captures a greater number of trials, as schematically displayed in Fig. 1. However, L will be overestimated for the short-counting-time data because the large clusters are cut apart by the measuring window and therefore the data appear as a larger number of smaller clusters (i.e., L is overestimated and the large-cluster probabilities are underestimated). This is analogous to the quadrat-sampling problem of spatial processes in ecology (Pielou 1957; Gleeson and Douglas 1975). Thus L should increase with counting time, but at a rate less than linear, in accordance with our observations. The increase in L with stimulus level again reflects the fact that spike rate generally increases with stimulus level, in accordance with (2).

A characteristic inter-trial time can be roughly estimated by taking the ratio of the long counting time T ($= 204.8$ ms) to the empirical average number of trials L_{200} (≈ 10), as portrayed in Fig. 1. The typical time splay-out of neural spikes within an individual cluster therefore has a duration of the order of $204.8/10 \text{ ms} \approx 20$ ms. This time course is not unlike the characteristic auditory response times determined in a broad variety of other experiments such as click responses (Kiang et al. 1965), impulse-response function measurements (Littlefield 1973), and relative-refractory periods (Gray 1967). It is also of the same order of magnitude as the value experimentally observed in the cat's retinal ganglion cell, which is ≈ 27 ms (Barlow et al. 1971; Saleh and Teich 1985).

We could not discern any special relationships between the parameters of the theory (π_i's and L) and the parameters of the experiment (unit type, unit CF, stimulus level, stimulus frequency), aside from those expected on the basis of constraints on the mean,

variance, and maximum count number associated with the reduced-quintinomial distribution. Furthermore, all of the data that we fitted were collected from experiments of duration $D = 51.2$ or 102.4 s. Since an increase in the experimental duration leads to an increase in the count variance, and a decrease in the count mean and mean-to-variance ratio (Teich 1988), the results obtained here are likely to be specific to the experimental duration D as well as the counting time T.

Finally, we carried out a frequency analysis of the experimental PND's, using the FFT procedure discussed earlier. This technique is useful for PND's that exhibit structure, i.e., principally the long-counting-time PND's. In Fig. 9, we display a long-counting-time PND (Fig. 9a) and its DFT (Fig. 9b), along with their binomially filtered versions (Fig. 9c and d), for Unit No. 44/81426.PN (stimulated by a 40 dB:re FTC pure tone at 7227 Hz). This is the same distribution represented in the last row/first column of Fig. 8. The DFT magnitude, particularly the filtered version, shows a substantial degree of structure (at or near 2, 3, 4, and 9 counts) which is associated with the rather large and quasi-regularly spaced fluctuations present in the PND. Although the cluster probabilities can be estimated from these DFT's, these estimates are likely to be less accurate than those obtained from the numerical fit, as indicated earlier.

5 Conclusion

We have investigated the reduced-multinomial family of theoretical cluster counting distributions in terms of their suitability as a model for neural-spike counting distributions in primary auditory fibers. These distributions are generalizations of the binomial distribution. The reduced-quintinomial distribution provides theoretical results that describe the characteristics of the PND's quite well, accounting for the smooth or scalloped (spike-pair) behavior of short-counting-time PND's, the jagged nature of long-counting-time PND's, and the Poisson-like character of very-short-counting-time PND's. The theory appears to be applicable for a broad variety of unit types and stimulus conditions. The results are far more satisfactory than those achievable with other commonly used auditory counting distributions, such as the binomial or the dead-time-modified Poisson distribution (Teich et al. 1978). They may also be useful for describing the spike counting distributions in other neural systems. Nevertheless, a model that specifies only the counting distribution is necessarily incomplete. Ideally, it is desirable to identify the underlying neural point process in terms of its moment generating functional. This, in turn, would permit any desired measure (e.g., the PND, PST, and PID) to be calculated and, in the auditory case, the mapping between the space $p(n, L)$ and $p(n, T)$ to be rigorously established.

The reduced-multinomial family of distributions admits values for the spike-number mean-to-variance ratio that are independent of stimulus level, in agreement with experimental observation. A number of procedures for fitting the theoretical distributions to the experimental PND's were studied. These include the use of a minimum-mean-square error criterion, the factorial moments of the data, and the discrete Fourier transform of the PND. The first of these techniques appears to be the most useful.

Use of the theory permits a characteristic time course for the splay-out of neural spikes within a cluster to be extracted from the data. This time turns out to be ≈ 20 ms which is not unlike the characteristic response times determined by a broad variety of other auditory experiments such as click responses, impulse-response functions, and relative-refractory periods.

By extending the calculations to a reduced-multinomial of higher order than the quintinomial, the theoretical treatment could be expanded to allow for the possible presence of spike clusters of size larger than 4 in the time window. However, this would involve the introduction of another parameter into the model and would be somewhat tedious from an algebraic point of view. But it could improve the fits of theory to data, and might be useful in describing the PND's for yet longer-counting-time experiments and for other neural systems. It is our view that, on balance, the reduced-quintinomial distribution represents an appropriate theoretical middle ground that is capable of suggesting an explanation for the nature of 204.8 ms experimental PND's in the auditory system without invoking undue computational complexity.

Finally, we point out that a mathematical model for the point process underlying the reduced-quintinomial distribution could be developed. One way to achieve this would be to construct a cascade of two stochastic processes in the form of a primary process driving a secondary process. An appealing alternative theoretical approach to achieving spike clusters that we have pursued involves compounding a simple discrete distribution with a fractal stochastic process. This has the merit of accounting for both the T and D dependences of our data.

Acknowledgements. We are grateful to Ian Storper, David Feld, and Mark Rosovsky for a great deal of assistance with the calculations. This work was supported by the National Science Foundation through the Columbia Center for Telecommunications Research and by the National Institutes of Health through the Columbia College of Physicians & Surgeons.

References

Barlow HB, Levick WR, Yoon M (1971) Responses to single quanta of light in retinal ganglion cells of the cat. Vision Res [Suppl 3] 11:87–101

Cox DR, Lewis PAW (1966) The statistical analysis of series of events. Methuen, London

Gleeson AC, Douglas JB (1975) Quadrat sampling and the estimation of Neyman type-A and Thomas distributional parameters. Austr J Stat 17:103–113

Gray PR (1967) Conditional probability analysis of the spike activity of single neurons. Biophys J 7:759–777

Kiang NY-S, Watanabe T, Thomas EC, Clark LF (1965) Discharge patterns of single fibers in the cat's auditory nerve. Res Monograph no 35. MIT Press, Cambridge, Mass

Littlefield WM (1973) Investigation of the linear range of the peripheral auditory system. Ph.D. thesis, Department of Electrical Engineering, Washington University, St. Louis, Mo

Mastronarde DN (1983) Correlated firing of cat retinal ganglion cells. II. Responses of X- and Y-cells to single quantal events. J Neurophysiol 49:325–349

Neyman J, Scott EL (1958) A statistical approach to problems of cosmology. J R Stat Soc Ser B 20:1–43

Olkin I, Gleser LJ, Derman C (1980) Probability models and applications. Macmillan, New York, pp 415–430

Pielou EC (1957) The effect of quadrat size on the estimation of the parameters of Neyman's and Thomas's distributions. J Ecol 45:31–47

Rose JE, Brugge JF, Anderson DJ, Hind JE (1967) Phase-locked response to low-frequency tones in single auditory nerve fibers of the squirrel monkey. J Neurophysiol 30:769–793

Saleh BEA (1978) Photoelectron statistics. Springer, Berlin Heidelberg New York, pp 9–10, 16–17

Saleh BEA, Teich MC (1985) Multiplication and refractoriness in the cat's retinal-ganglion-cell discharge at low light levels. Biol Cybern 52:101–107

Scheaffer RL, McClave JT (1986) Probability and statistics for engineers. Prindle, Weber & Schmidt, Boston, pp 139–140

Schwartz M, Shaw L (1975) Signal processing: discrete spectral analysis, detection, and estimation. McGraw-Hill, New York, pp 28–60

Teich MC (1988) Long-counting-time pulse-number distributions for neural spikes in cat primary auditory fibers. IEEE Trans Biomed Eng [special issue on neural signal processing] (submitted for publication)

Teich MC, Khanna SM (1985) Pulse-number distribution for the neural spike train in the cat's auditory nerve. J Acoust Soc Am 77:1110–1128

Teich MC, Matin L, Cantor BI (1978) Refractoriness in the maintained discharge of the cat's retinal ganglion cell. J Opt Soc Am 68:386–402

Thomas JB (1971) An introduction to applied probability and random processes. Wiley, New York, pp 10, 25

Westerman LA, Smith RL (1984) Adaptation and variability of spike discharge of auditory nerve. J Acoust Soc Am [Suppl 1] 75:S13

Received: August 18, 1987

Prof. M. C. Teich
Department of Electrical Engineering
Columbia University
New York, NY 10027
USA