



# Measuring the Performance of Neural Models

Oliver Schoppe<sup>1,2\*</sup>, Nicol S. Harper<sup>1</sup>, Ben D. B. Willmore<sup>1</sup>, Andrew J. King<sup>1</sup> and Jan W. H. Schnupp<sup>1\*</sup>

<sup>1</sup> Department of Physiology, Anatomy, and Genetics, University of Oxford, Oxford, UK, <sup>2</sup> Bio-Inspired Information Processing, Technische Universität München, Garching, Germany

## OPEN ACCESS

### Edited by:

Ad Aertsen,  
University of Freiburg, Germany

### Reviewed by:

Alexander G. Dimitrov,  
Washington State University  
Vancouver, USA  
James McFarland,  
University of Maryland, College Park,  
USA

Jannis Hildebrandt,  
Carl von Ossietzky University of  
Oldenburg, Germany

### \*Correspondence:

Oliver Schoppe  
oliver.schoppe@tum.de;  
Jan W. H. Schnupp  
jan.schnupp@dpag.ox.ac.uk

**Received:** 27 September 2015

**Accepted:** 21 January 2016

**Published:** 10 February 2016

### Citation:

Schoppe O, Harper NS,  
Willmore BDB, King AJ and  
Schnupp JWH (2016) Measuring the  
Performance of Neural Models.  
Front. Comput. Neurosci. 10:10.  
doi: 10.3389/fncom.2016.00010

Good metrics of the performance of a statistical or computational model are essential for model comparison and selection. Here, we address the design of performance metrics for models that aim to predict neural responses to sensory inputs. This is particularly difficult because the responses of sensory neurons are inherently variable, even in response to repeated presentations of identical stimuli. In this situation, standard metrics (such as the correlation coefficient) fail because they do not distinguish between explainable variance (the part of the neural response that is systematically dependent on the stimulus) and response variability (the part of the neural response that is not systematically dependent on the stimulus, and cannot be explained by modeling the stimulus-response relationship). As a result, models which perfectly describe the systematic stimulus-response relationship may appear to perform poorly. Two metrics have previously been proposed which account for this inherent variability: Signal Power Explained (*SPE*, Sahani and Linden, 2003), and the normalized correlation coefficient (*CC<sub>norm</sub>*, Hsu et al., 2004). Here, we analyze these metrics, and show that they are intimately related. However, *SPE* has no lower bound, and we show that, even for good models, *SPE* can yield negative values that are difficult to interpret. *CC<sub>norm</sub>* is better behaved in that it is effectively bounded between  $-1$  and  $1$ , and values below zero are very rare in practice and easy to interpret. However, it was hitherto not possible to calculate *CC<sub>norm</sub>* directly; instead, it was estimated using imprecise and laborious resampling techniques. Here, we identify a new approach that can calculate *CC<sub>norm</sub>* quickly and accurately. As a result, we argue that it is now a better choice of metric than *SPE* to accurately evaluate the performance of neural models.

**Keywords:** sensory neuron, receptive field, signal power, model selection, statistical modeling, neural coding

## 1. INTRODUCTION

Evaluating the performance of quantitative models of neural information processing is an essential part of their development. Appropriate metrics enable us to compare different models and select those which best describe the data. Here, we are interested in developing improved metrics to assess models of the stimulus-response relationships of sensory neurons, in the challenging (but common) situation where the stimulus-response relationship is complex, and neuronal responses are highly variable. In this case, the development of appropriate performance metrics is not trivial, and so there is a lack of consensus about which metrics are to be used.

The classical way to record and model neural responses has been to repeatedly present an animal with a small, well-defined set of stimuli (such as sinusoidal gratings of different orientations, or sounds of different frequencies). The neural responses to repeated presentations of each stimulus are then averaged. Using a small stimulus set, it may be possible to present the same stimulus enough times that this averaging succeeds in reducing the effect of neuronal response variability (Döerrscheidt, 1981). It may then be possible to produce models which accurately describe the relationship between the stimulus and the averaged responses. These models can then be accurately evaluated by comparing the modeled and actual neuronal responses using standard metrics such as correlation coefficient. Under these circumstances, the correlation coefficient may be appropriate and can easily be interpreted—a poor model will have a correlation coefficient close to 0, a perfect model will have a correlation coefficient close to 1, and the square of the value of the correlation coefficient equals the proportion of the variance in the neural responses that the model is able to account for.

However, recent work in sensory neuroscience has increasingly focused on the responses of neurons to complex stimuli (Atencio and Schreiner, 2013; David and Shamma, 2013), and even natural stimuli (Prenger et al., 2004; Asari and Zador, 2009; Laudanski et al., 2012). For such stimuli, even very sparse sampling of the stimulus space may require the presentation of very large numbers of different stimuli (at least of order  $2^d$  for  $d$  stimulus dimensions; also see Shimazaki and Shinomoto, 2007). This makes it difficult to repeatedly present stimuli enough times for response variability to simply average out. Estimating mean responses for a particular stimulus is thus subject to sampling noise, and in addition to that, the neuron under study may also be “intrinsically noisy” in the sense that only a small proportion of the response variability may be attributable to variability of the stimulus. Such situations are very common in sensory neuroscience, and they can render the use of correlation coefficients to evaluate the performance of models that map stimuli to responses very misleading. If only a fraction of the neural response variability is stimulus linked, then even a perfect model of that stimulus linkage will only ever be able to account for some fraction of the variance in the observed neural response data. This places a limit on the maximum correlation coefficient that can be achieved, and the interpretation of the raw correlation coefficients becomes ambiguous: for example, a relatively low correlation coefficient of 0.5 might be due to an excellent model of a noisy dataset, or of a rather poor model of a dataset with very low intrinsic and sampling noise, or something in between.

Different approaches for taking neural variability into account when measuring model performance have been developed. To get an unbiased estimate of *mutual information*, Panzeri and Treves (1996) suggested a method to extrapolate information content to an infinite number of trials (also see Atencio et al., 2012). Roddey et al. (2000) compared the coherence of pairs of neural responses to independent stimulus repetitions to derive a *minimum mean square error (MMSE)* estimator for an optimal model. The difference between the model prediction error and the MMSE of an optimal model allows the quantification of the

model performance relative to the best possible performance given the neural variability.

Based not only on pairs, but even larger sets of neural responses to independent stimulus repetitions, Sahani and Linden developed the very insightful decomposition of the recorded signal into *signal power* and *noise power* (Sahani and Linden, 2003). This has led to the *signal power explained (SPE)*, a measure based on *variance explained* which discounts “unexplainable” neural variability. While the work of Roddey et al. (2000) was already based on the differentiation between explainable and unexplainable neural response components, Sahani and Linden (2003) provided explicit estimations for those components. The SPE measure has been widely adopted, albeit under various names such as *predictive power*, *predicted response power*, and *relative prediction success* (Sahani and Linden, 2003; Machens et al., 2004; Ahrens et al., 2008; Asari and Zador, 2009; Rabinowitz et al., 2012). Also, it has been used as a basis for specific variants of measures for model performance (Haefner and Cumming, 2009).

Focusing on coherence and the correlation coefficient, Hsu and colleagues developed a method to normalize those measures by their upper bound ( $CC_{max}$ ), which is given by the inter-trial variability (Hsu et al., 2004). This yields the *normalized correlation coefficient* ( $CC_{norm}$ ). Following their suggestion, the upper bound can be approximated by looking at the similarity between one half of the trials and the other half of the trials ( $CC_{half}$ ). This measure has also been used by Gill et al. (2006) and Touryan et al. (2005). Others used the absolute correlation coefficient and controlled for inter-trial variability by comparing the absolute values with  $CC_{half}$  (Laudanski et al., 2012).

The two metrics *SPE* and  $CC_{norm}$  have been developed independently, but they both attempt—in different ways—to provide a method for assessing model performance independent of neuronal response variability. Here, we here analyze these metrics, show for the first time that they are closely related, and discuss the shortcomings of each. We provide a new, efficient way to directly calculate  $CC_{norm}$  and show how it can be used to accurately assess model performance, overcoming previous shortcomings.

## 2. CRITERIA OF MODEL EVALUATION

Neural responses are often measured as the membrane potential (Machens et al., 2004; Asari and Zador, 2009) or as the time-varying firing rate (Sahani and Linden, 2003; Gill et al., 2006; Ahrens et al., 2008; Rabinowitz et al., 2011; Laudanski et al., 2012; Rabinowitz et al., 2012) (which we will use without loss of generality). Thus, a measure of performance for such models should quantify the similarity of the predicted firing rate  $\hat{y}$  and the recorded firing rate  $y$  (also known as the peri-stimulus time histogram, PSTH):

$$y(t) = \frac{1}{N} \sum_{n=1}^N R_n(t) \quad (1)$$

where  $R_n$  is the recorded response of the  $n$ th stimulus presentation and  $N$  is the total number of stimulus presentations

(trials). Both  $R_n(t)$  and  $y(t)$  are a function of the time bin  $t$ , but the argument  $t$  will not be shown for rest of the manuscript. Each value of the vector  $R_n$  contains the number of spikes that were recorded in the corresponding time bin. Note that, given the trial-to-trial variability of sensory responses, the recorded firing rate  $y$  is only an approximation of the true (but unknown) underlying firing rate function that is evoked by the presentation of a stimulus (also see Kass et al., 2003). It is a sample mean which one would expect to asymptote to the true mean as the number of trials increases ( $N \rightarrow \infty$ ). As will be discussed in detail at a later point, the difference between the recorded firing rate  $y$  and the true underlying firing rate is considered to be noise under the assumption of rate coding. This is the unexplainable variance that reflects the variability of the neuronal response. As the number of trials increases, the difference between  $y$  and the true underlying firing rate decreases and so does the non-deterministic and thus unexplainable variance in the signal.

With the recorded firing rate  $y$  being the target variable for the prediction  $\hat{y}$ , a measure of model performance needs to quantify the similarity between both signals, i.e., the prediction accuracy. Note that model performance is not necessarily the same as prediction accuracy (see next section).

### 3. SIGNAL POWER EXPLAINED

Two somewhat related metrics which are widely applied in statistics are the “coefficient of determination” ( $CD$ ) and the “proportion of variance explained” ( $VE$ ). Both these metrics essentially incorporate the assumption that the quantitative observations under study—in our case the responses of a sensory neuron or neural system—are the sum of an essentially deterministic process which maps sensory stimulus parameters onto neural excitation, plus an additive, stochastic noise process which is independent of the recent stimulus history (Sahani and Linden, 2003). Consequently, if a model is highly successful at predicting the deterministic part, subtracting the predictions from the observations should leave only the noise part, but if its predictions are poor, the residuals left after subtracting predictions from observations will contain both noise and prediction error. Thus, smaller residuals are taken as a sign of better prediction. The  $CD$  is an index that quantifies the size of the residuals relative to the size of the original observation in a quite direct manner as a sum of squares, and subtracts that unaccounted for proportion from 100% to give an estimate of the proportion of the signal that is accounted for by the model. Thus

$$CD = 1 - \frac{\sum_t (y(t) - \hat{y}(t))^2}{\sum_t y(t)^2} \quad (2)$$

The  $VE$  quantifies prediction accuracy in a largely analogous manner, but instead of using the “raw” sum of squares of the observations and the residuals, it instead uses the respective sample variances, measured around their respective sample means:

$$VE = 1 - \frac{Var(y - \hat{y})}{Var(y)} \quad (3)$$

This makes the  $VE$  insensitive to whether the mean of the predicted responses closely corresponds to the mean of the observed responses over all  $t$ , which can sometimes be an advantage. Even small errors (biases) in the mean of the prediction can be penalized quite heavily by the  $CD$  measure as these will accumulate over every sample. The  $VE$  measure can be thought of as deeming such biases as unimportant, and focusing solely on how well the model predicts the trends in the responses as a function of  $t$ .

$CD$  and  $VE$  have a long established history in statistics, but neither provide an unambiguous measure of model performance because large amounts of residual variance, and therefore low  $VE$  or  $CD$  values, could arise either if the model provides a poor approximation to underlying deterministic and predictable aspects of the process under study, or if the model captures the deterministic part of the process perfectly well, but large amounts of fundamentally unpredictable noise in the system nevertheless cause the amount of residual variance to be large. In other words, even a perfect model cannot make perfect predictions, because the neuronal response has a non-deterministic component. Even if the model was completely identical to the neuron in every aspect, it would nevertheless be unable to explain 100% of the variance in the neuronal responses because the PSTHs collected over two separate sets of stimulus presentations cannot be expected to be identical and the first set does not perfectly predict the second. Furthermore, since the number of trials  $N$  used to determine any one PSTH is often rather low for practical reasons, observed PSTHs are often somewhat rough, noisy estimators of the underlying neural response function (also see Döerrscheidt, 1981; Kass et al., 2003; Shimazaki and Shinomoto, 2007). A good measure of model performance for sensory neural systems should take these considerations into account and judge model performance relative to achievable, rather than total, prediction accuracy. Such considerations led Sahani and Linden (2003) to introduce metrics which split the variance in an observed PSTH, the *total power* ( $TP$ ), into the *signal power* ( $SP$ ), which depends deterministically on recent stimulus history, and the *stochastic noise power* ( $NP$ ). Only the  $SP$  is explainable in principle by a model, and the *signal power explained* ( $SPE$ ) thus aims to quantify model performance relative to the best achievable performance.  $SPE$  is defined as:

$$SPE = \frac{Var(y) - Var(y - \hat{y})}{SP} \quad (4)$$

$$SP = \frac{1}{N-1} (N \times Var(y) - TP)$$

$$TP = (N-1) \times \sum_{n=1}^N Var(R_n) \quad (5)$$

$SPE$  is quantified as the ratio of the *explained* signal power relative to the *explainable* signal power<sup>1</sup>. The *explained* signal power is

<sup>1</sup>Please note that we do not use the notation of Sahani and Linden (2003). However, all definitions are identical. Sahani and Linden define the *power*  $P$  of a signal  $r$  as the “average squared deviation from the mean:  $P(r) = \langle (r_t - \langle r_t \rangle)^2 \rangle$ ” where  $\langle \cdot \rangle$  denotes the mean over time. This is identical to the variance of the signal, which we use.

calculated by subtracting the variance of the residual (the error) from the total variance in the observed firing rate. The *explainable* signal power *SP* is calculated according to formulas developed in Sahani and Linden (2003) and reproduced below (Equation 13). Good models will yield small error variance and thus a large *SPE* - and vice versa. However, this measure lacks an important characteristic: it is not bounded. While a perfect model would yield an *SPE* of 100%, the measure has no lower bound and can go deeply into negative values when the variance of the error is bigger than the variance of the neural signal. This shortcoming of the *SPE* metric can be exposed by reformulating parts of the equation. First, observe that for two random variables *X* and *Y* the variance of their difference can be expressed as :

$$\text{Var}(Y - X) = \text{Var}(Y) + \text{Var}(X) - 2 \times \text{Cov}(X, Y) \quad (6)$$

Applying this reformulation to Equation 5 reveals that:

$$\text{SPE} = \frac{\text{Var}(y) - \text{Var}(y - \hat{y})}{\text{SP}} = \frac{2 \times \text{Cov}(y, \hat{y}) - \text{Var}(\hat{y})}{\text{SP}} \quad (7)$$

Consider a particularly bad model, which produces predictions that are no better than the output of a random number generator. The covariance between the predictions and the neural responses will then be close to zero, but the variance (i.e., the power of the predicted signal) of the predicted signal may nevertheless be large. The *SPE* for such a model would be a negative number equal to  $-\text{Var}(\hat{y})/\text{SP}$ . This is a counterintuitive property of the *SPE* metric: the “proportion of the signal power that is explained” by a list of random numbers should be zero, not negative. Also, two bad models that are equally unable to capture the trends of the signal they are trying to predict and thus have near zero covariance may nevertheless have widely different negative *SPE* values, but how negative their *SPE* values are may have little systematic relationship to how large their prediction errors are on average, which makes small or negative *SPE* values very difficult to interpret.

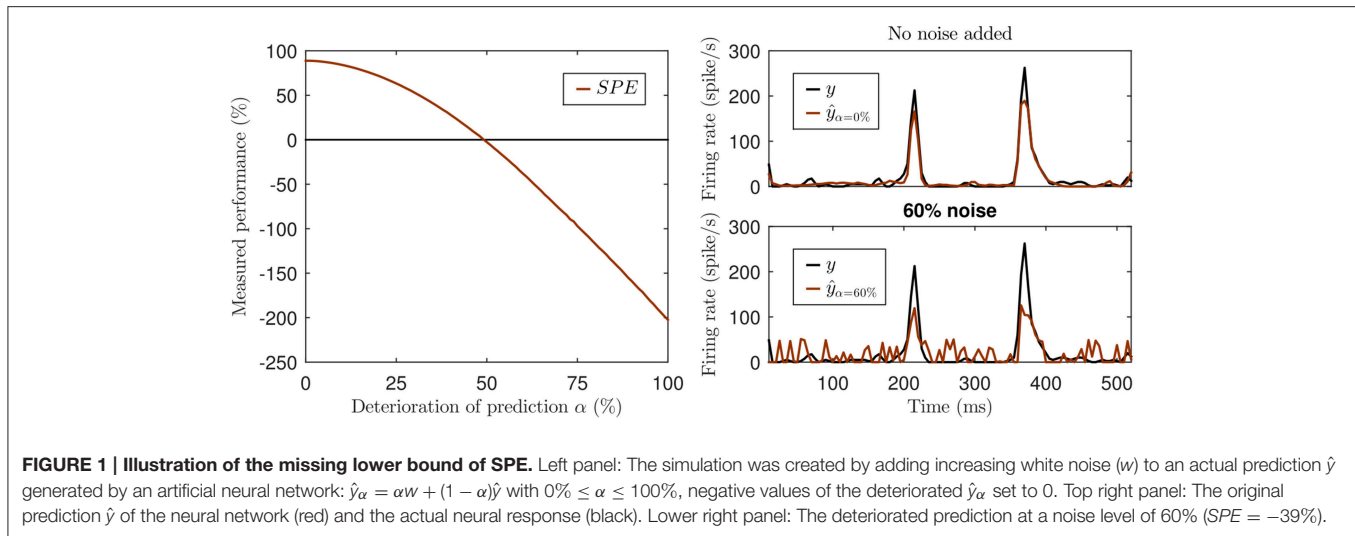
This can be illustrated with a simple hypothetical example. Imagine a visual cortex simple cell responding to a sinusoidal contrast grating stimulus with a sinusoidal modulation of its firing rate, so its observed response is a sine wave, let's say, of an amplitude of  $\pm 1$  spikes/s around a mean firing rate of 10 spikes/s at a modulation rate of 1 Hz. Let us further assume that model A predicts sinusoidal firing at a 2 Hz modulation rate with an amplitude of  $\pm 2$  spikes/s around a mean of 10 spikes/s, and model B predicts a sinusoidal firing at 2 Hz with an amplitude of  $\pm 1$  spikes/s around a mean of 100 spikes/s. Since neither model A nor B correctly predicted the rate of the sinusoidal firing rate modulations, and because sine waves of different frequencies are orthogonal, both models will have covariance of zero with the observed data. Thus, they have a negative *SPE*, as the signal variance is greater than zero. And because model A predicted larger amplitude fluctuations than model B, and thus has greater variance, the *SPE* of model A will be more negative than that of model B, which one might be tempted to interpret to mean that model A performed worse. However, the

discrepancy or prediction error between observed and predicted rates for model A will never be more than 3 spikes/s, while that of model B will never be less than 88 spikes/s, and the more negative *SPE* of model A contrasts sharply with the fact that model A produces a much smaller mean squared prediction error than model B. Furthermore *SPE* can yield negative values even when there is a reasonable amount of covariance between model and prediction, if the variance in the predicted signal is also sizable. This is illustrated in **Figure 1**. Not only is such a measure rather hard to interpret, but it can be misleading. Due to the missing lower bound the values can not only become negative, but the exact value also depends on the variance of the prediction. Consider the prediction with 60% noise in the lower right panel of **Figure 1**. While this prediction is surely not a good one, the fact that data and model prediction co-vary to a fair degree is nevertheless readily apparent, and it would be hard to argue that a model predicting a flat, arbitrary, constant firing rate (say 800 spikes/s) would be a better alternative. Yet the variance of any predicted constant firing rate would be zero and so would be their *SPE*, which may seem indicative of a “better explanatory power” of the constant rate model compared to the “60% noise” added model of **Figure 1** with its *SPE* =  $-39\%$ , but the noisy model clearly captures some of the major peaks in the data while constant rate models don't even try.

These examples illustrate that models can be thought of as being wrong in different ways. They can be “biased,” predicting an incorrect overall mean response rate, they can be “scaled wrong,” predicting fluctuations that are too small or too large, or they can fail to predict the trends and dynamics of the data, leading to near zero covariance between observation and prediction. Different metrics of model performance will differ in how sensitive they are to these different types of error. *SPE* is sensitive both to poor scaling and poor covariance, but not to bias. Some might argue, quite reasonably, that this combined sensitivity to two types of error is a virtue: When *SPE* values are large then we can be confident that the model achieves both good covariance and good scaling. However, the downside of this joint sensitivity is that small or negative *SPE* values have limited diagnostic value because they could be due to small covariance or to overestimated (but not underestimated) predicted variance, or some combination of the two. Consequently, as we will illustrate further in section 6, *SPE* values below about 0.4 become very difficult to interpret, and may be much at odds with other commonly used measures of model performance.

Negative values of the *SPE* have been previously reported (Machens et al., 2004; Ahrens et al., 2008) and have been interpreted as a sign of overfitting of the model. Overfitting usually manifests itself as a decline in covariance between data and predictions in cross-validation tests, and as such would result in small or negative *SPEs*, but because *SPE* will become negative for any prediction which has a residual variance that is larger than the variance of the target signal, negative *SPE* is not a specific diagnostic of overfitting. Also negative *SPEs* do not necessarily imply that a model performs worse than a “null model” which predicts constant responses equal to the mean firing rate. In fact,





any model predicting any arbitrary constant value (even a “dead neuron model” predicting a constant firing rate of 0 spikes/s) will have an *SPE* of zero and might on that basis be judged to perform better than other models generating noisy but fairly reasonable predictions (see **Figure 1**).

Of the three different types of error just discussed, large bias, poor scaling, small covariance, *SPE* is sensitive to two, covariance and scaling, although it is particularly excessively large, but not excessively small, scaling, that will drive *SPE* values down. Perhaps it is inevitable that single performance measures which are sensitive to multiple different types of error become very difficult to interpret as soon as performance becomes suboptimal. To an extent, whether one deems it preferable to have an error metric that is sensitive to bias, scaling and low covariance all at once, or whether one chooses a metric that is more specific in its sensitivity to only one of type of error is a matter of personal preference as well as of what one is hoping to achieve, but joint sensitivity to multiple different types of error is certainly problematic when the measure is to be used for model comparison, given that the relative weighting of the different types of error in the metric may not be readily apparent and it is unlikely to reflect how problematic the different types of error are in modeling. A constant bias, which would, for example, be heavily penalized by the *CD* metric discussed at the beginning of this section, can be easily fixed by adding or subtracting a constant value from the predictions. Similarly, scaling errors can be easily fixed by multiplication by a scalar. These two types of error pertain only to the relatively uninteresting stationary statistical properties of the data. They are in some sense trivial, and easily remedied through a simple linear adjustment. Low covariance, in contrast, is indicative of a much more profound inability of the model to capture the nature or dynamics of the neural stimulus-response relationships. In our opinion, the assessment of model performance should therefore rely first and foremost measures which are highly sensitive to poor covariance and insensitive to bias or scaling, and we discuss measures which have these properties in the next section. If needed, these could then be

supplemented with additional metrics that can diagnose biases or scaling errors.

#### 4. ABSOLUTE AND NORMALIZED CORRELATION COEFFICIENT

Another measure widely used in statistics, Pearson’s product-moment correlation coefficient can also be used to assess the similarity of two time-varying signals. The correlation coefficient quantifies the linear correlation and maps it to a value between  $-1$  and  $+1$ . To distinguish it from a normalized variant that will be used later in this section, the (absolute) correlation coefficient will from now on be abbreviated as  $CC_{abs}$ . It is defined as:

$$CC_{abs} = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} \quad (8)$$

$CC_{abs}$  satisfies many of the criteria that one might desire in a good measure of model performance. It quantifies the similarity between observation and prediction, it is bounded between  $-1$  and  $+1$ , and it can be interpreted easily and unambiguously. The normalization by the square root of the variances makes  $CC_{abs}$  insensitive to scaling errors, and the formulae for  $Var()$  and  $Cov()$  have subtractions of means built in that make  $CC_{abs}$  insensitive to bias, so that only the ability of  $Y$  to follow trends  $X$  is being quantified. However, like *VE*, it does not isolate model performance from prediction accuracy, which is inevitably limited by neural variability. In other words  $CC_{abs}$  might be small either because the model predictions  $Y$  are poor, or because the measured neural responses  $X$  are fundamentally so noisy that even an excellent model cannot be expected to achieve a large  $CC_{abs}$ . This was also noted by Hsu and colleagues who went on to develop an approach to quantify and account for the inherent noise in neural data (Hsu et al., 2004). Specifically, they introduced a method for normalizing coherence and correlation to the neural variability, which has later been applied as a performance measure (Touryan et al., 2005; Gill et al., 2006). Hsu

and colleagues define the normalized correlation coefficient as follows (Hsu et al., 2004)<sup>2</sup>:

$$CC_{norm} = \frac{CC_{abs}}{CC_{max}} \text{ with } CC_{max} = \sqrt{\frac{2}{1 + \sqrt{\frac{1}{CC_{half}^2}}}} \stackrel{CC_{half} > 0}{=} \sqrt{\frac{2}{1 + \frac{1}{CC_{half}}}} \quad (9)$$

Where  $CC_{max}$  is the maximum correlation coefficient between the recorded firing rate  $y$  and the best prediction  $\hat{y}$  that a perfect model could theoretically achieve. More specifically,  $CC_{max}$  is the correlation coefficient between the recorded firing rate  $y$  (which is based on  $N$  trials) and the true (but unknown) underlying firing rate function, which could only be determined precisely if the system was completely stationary and an infinite number of trials could be conducted ( $N \rightarrow \infty$ ). Even though the true underlying firing rate function can therefore usually not be determined with high accuracy through experiments, useful estimates of  $CC_{max}$  can nevertheless be calculated using the formulae in Equation 9. Following the methods of Hsu et al. (2004),  $CC_{half}$  is determined by splitting the data set into halves, and calculating the correlation coefficient between the PSTH constructed from the first half and the PSTH constructed from the second half of the trials. This approach determines  $CC_{max}$  by effectively extrapolating from  $N$  trials to the value that would be expected for  $N \rightarrow \infty$ .

Note that there are  $\frac{1}{2} \binom{N}{N/2}$  different ways to choose  $N/2$  out of  $N$  trials, and each such split of the data will yield a slightly different value for  $CC_{half}$ . Thus, in theory, the best estimate would average over all possible values of  $CC_{half}$  calculated for each possible split. In practice, this resampling technique can be computationally expensive, given the fact that there are already 92, 378 combinations for  $N = 20$  trials. Averaging over a smaller number of randomly chosen splits may often be sufficient, but this yields an imprecise estimation of  $CC_{max}$ .

In summary,  $CC_{norm}$  provides a feasible method for capturing model performance independently of noise in the neural responses. It gives values bounded between -1 and +1 (in practice, they are bounded between 0 and +1, as model predictions are either correlated or not correlated, but typically not anti-correlated to the firing rate). Furthermore, the measure lends itself to unambiguous interpretation, and its limitations are well-known. Finally, it is normalized so that its value does not depend on the variability of a particular data set. Thus, the normalized correlation coefficient  $CC_{norm}$  fulfills the criteria for a useful measure of model performance, but its current definition is based in a laborious and potentially imprecise resampling technique.

<sup>2</sup>The expression for  $CC_{max}$  can be derived from the work of Hsu et al. (2004) in two steps. First, Equations 6 and 8 from Hsu et al. (2004) are combined and solved for  $\gamma_{AR_M}$ . Second, the analogy of the coherence  $\gamma^2$  and the squared correlation coefficient  $CC^2$  allows to replace  $\gamma_{AR_M}$  with  $CC_{max}$  and  $\gamma_{R_1, M/2, R_2, M/2}$  with  $CC_{half}$ . In the notation of Hsu and colleagues  $\gamma_{AR_M}^2$  denotes the coherence of the mean response over  $M$  trials with the true (but unknown) underlying firing rate  $A$ , i.e., the maximum achievable coherence of a perfect model.

## 5. A CONSOLIDATED APPROACH TO QUANTIFYING NEURAL VARIABILITY

As will have become clear in the previous sections, the two measures  $SPE$  and  $CC_{norm}$  follow the same logic in that both measure prediction accuracy and normalize it by a quantification of the inherent reproducibility of the neural responses that are to be modeled ( $SP$  or  $CC_{max}$ , respectively). In this section we will show that these two approaches of normalization not only follow the same logic, but are mathematically largely equivalent. This provides a deeper insight into the underlying concept and gives rise to a more elegant and efficient technique to normalize the correlation coefficient.

Following the methods of Sahani and Linden (2003)<sup>3</sup>, the signal power  $SP$  (i.e., the deterministic part of the recorded firing rate  $y$ ) can be expressed as:

$$SP = \frac{1}{N-1} (N \times \text{Var}(y) - TP) \quad (10)$$

$$= \frac{1}{N-1} \left( N \times \text{Var} \left( \frac{1}{N} \sum_{n=1}^N R_n \right) - \frac{1}{N} \sum_{n=1}^N \text{Var}(R_n) \right) \quad (11)$$

$$= \frac{1}{N-1} \left( N \times \frac{1}{N^2} \text{Var} \left( \sum_{n=1}^N R_n \right) - \frac{1}{N} \sum_{n=1}^N \text{Var}(R_n) \right) \quad (12)$$

$$= \frac{1}{N-1} \left( \frac{1}{N} \times \text{Var} \left( \sum_{n=1}^N R_n \right) - \frac{1}{N} \sum_{n=1}^N \text{Var}(R_n) \right) \quad (13)$$

Where  $TP$  is the total power (i.e., the average variance of a single trial) and  $R_n$  is the recorded neural response of the  $n$ th trial. Since the normalization factor of  $SPE$  is the inverse of  $SP$  it will be convenient to express it as:

$$\frac{1}{SP} = \frac{N(N-1)}{\text{Var} \left( \sum_{n=1}^N R_n \right) - \sum_{n=1}^N \text{Var}(R_n)} \quad (14)$$

Furthermore, using Equation 14 the ratio of the noise power  $NP$  over  $SP$  can be expressed as:

$$\frac{NP}{SP} = \frac{TP - SP}{SP} = \frac{TP}{SP} - 1 = \frac{(N-1) \times \sum_{n=1}^N \text{Var}(R_n)}{\text{Var} \left( \sum_{n=1}^N R_n \right) - \sum_{n=1}^N \text{Var}(R_n)} - 1 \quad (15)$$

For  $CC_{norm}$  the normalization factor is the inverse of  $CC_{max}$  and, following the methods of Hsu et al. (2004), it is currently determined with an indirect resampling method using Equation

<sup>3</sup>Again, please note that Sahani and Linden (2003) use  $\overline{r^{(n)}}$  to denote the average over trials. In order to facilitate the reformulation of the equation we do not use this abbreviated notation. Despite this difference in notation, this definition of  $SP$  is identical to the definition provided by Sahani and Linden (Equations 1 on Page 3).

9. We will now show how  $CC_{max}$  can be computed directly by exploiting the relation between  $SPE$  and  $CC_{norm}$ .

The coherence  $\gamma_{AB}^2$  between a source signal  $A$  and a noisy recording  $B$  of this signal can be related to the signal-to-noise ratio, i.e., the coherence is just a function of the noise process itself (see Marmarelis, 1978 for details). In the context of neural recordings, Hsu et al. (2004) used this relation to express the coherence of the true (but unknown) underlying firing rate function (the source  $A$ ) to the observed PSTH (the noisy recording  $B$ ) as a function of the signal-to-noise ratio of the recording. They quantified this in terms of signal power of the frequency domain signals, but since the power of corresponding time and frequency domain signals is identical, we can rewrite their expression (see formulas 5 and 6 of Hsu et al., 2004) directly in terms of  $NP$  and  $SP$  to get:

$$\gamma_{AB}^2 = \frac{SP}{SP + \frac{1}{N}NP} \quad (16)$$

The derivation of the coherence function between the true underlying firing rate function and the observed neural response is analogous for the squared correlation coefficient between both signals (also see Hsu et al., 2004 for details on this analogy). Thus, we can apply the same principle to express the the inverse of  $CC_{max}$  as:

$$\frac{1}{CC_{max}} = \sqrt{1 + \frac{1}{N} \times \frac{NP}{SP}} \quad (17)$$

Combining Equation 17 with Equation 15 now allows us to express the inverse of  $CC_{max}$  as:

$$\frac{1}{CC_{max}} = \sqrt{1 + \frac{1}{N} \left( \frac{(N-1) \times \sum_{n=1}^N Var(R_n)}{Var\left(\sum_{n=1}^N R_n\right) - \sum_{n=1}^N Var(R_n)} - 1 \right)} \quad (18)$$

$$= \sqrt{1 - \frac{1}{N} + \frac{(1 - \frac{1}{N}) \times \sum_{n=1}^N Var(R_n)}{Var\left(\sum_{n=1}^N R_n\right) - \sum_{n=1}^N Var(R_n)}} \quad (19)$$

Based on Equation 8 and 9 the normalized correlation coefficient  $CC_{norm}$  between the recorded firing rate  $y$  and the model prediction  $\hat{y}$  can now be expressed as:

$$CC_{norm} = \frac{CC_{abs}}{CC_{max}} = \frac{Cov(y, \hat{y})}{\sqrt{Var(y)Var(\hat{y})}} \frac{1}{CC_{max}} \quad (20)$$

$$= \frac{Cov(y, \hat{y})}{\sqrt{Var(y)Var(\hat{y})}} \sqrt{1 - \frac{1}{N} + \frac{(1 - \frac{1}{N}) \times \sum_{n=1}^N Var(R_n)}{Var\left(\sum_{n=1}^N R_n\right) - \sum_{n=1}^N Var(R_n)}} \quad (21)$$

$$= \frac{Cov(y, \hat{y})}{\sqrt{Var(y)Var(\hat{y})}} \sqrt{1 - \frac{1}{N}} \sqrt{1 + \frac{\sum_{n=1}^N Var(R_n)}{Var\left(\sum_{n=1}^N R_n\right) - \sum_{n=1}^N Var(R_n)}} \quad (22)$$

$$= \frac{Cov(y, \hat{y})}{\sqrt{Var(\hat{y})}} \frac{\sqrt{1 - \frac{1}{N}}}{\sqrt{\frac{1}{N^2} Var\left(\sum_{n=1}^N R_n\right)}} \sqrt{1 + \frac{\sum_{n=1}^N Var(R_n)}{Var\left(\sum_{n=1}^N R_n\right) - \sum_{n=1}^N Var(R_n)}} \quad (23)$$

$$= \frac{Cov(y, \hat{y})}{\sqrt{Var(\hat{y})}} \sqrt{\frac{N(N-1)}{Var\left(\sum_{n=1}^N R_n\right)}} \sqrt{1 + \frac{\sum_{n=1}^N Var(R_n)}{Var\left(\sum_{n=1}^N R_n\right) - \sum_{n=1}^N Var(R_n)}} \quad (24)$$

$$= \frac{Cov(y, \hat{y})}{\sqrt{Var(\hat{y})}} \sqrt{N(N-1)} \sqrt{\frac{1}{Var\left(\sum_{n=1}^N R_n\right) - \sum_{n=1}^N Var(R_n)}} \quad (25)$$

$$= \frac{Cov(y, \hat{y})}{\sqrt{Var(\hat{y})}} \sqrt{\frac{N(N-1)}{Var\left(\sum_{n=1}^N R_n\right) - \sum_{n=1}^N Var(R_n)}} \quad (26)$$

$$= \frac{Cov(y, \hat{y})}{\sqrt{Var(\hat{y})}} \sqrt{\frac{1}{SP}} \quad (27)$$

In other words, we can now express  $CC_{norm}$  as a simple function of  $SP$ . The previous derivation also shows that both methods,  $SPE$  and  $CC_{norm}$ , use the covariance to quantify the prediction accuracy and take the neural variability into account

by normalizing with the signal power  $SP$ . This has several implications. First,  $SPE$  will not reveal more about the prediction accuracy than  $CC_{norm}$ , because  $SPE$  and  $CC_{norm}$  quantify the similarity of the prediction and the neural response solely based on the covariance of both signals. It is well known that the (normalized) correlation coefficient is based on covariance, but it has hitherto not been made explicit that this is also the case for  $SPE$ . Note that  $SPE$  uses only the covariance to assess prediction accuracy and thus, cannot reveal more information about the similarity of both signals than  $CC_{norm}$ . Second, how both measures quantify neural variability is not only related, but mathematically equivalent. Third, in order to calculate  $CC_{norm}$  it is not necessary to laboriously compute an approximation to  $CC_{max}$  from repeated subsampling of the data to generate computationally inefficient and potentially imprecise estimates of  $CC_{half}$ . Instead, the normalization factor can be explicitly calculated with Equation 27, using Equation 13 for  $SP$  as suggested by Sahani and Linden (2003). The close relationship between both measures can also be visualized by squaring  $CC_{norm}$  (left panel of Figure 2).

In summary,  $CC_{norm}$  as defined in Equation 27 provides an insightful measure of model performance. It quantifies the prediction accuracy using the covariance and isolates model performance by taking the amount of intrinsic variability in the observed neural responses into account. It is in theory bounded between -1 and 1, and in practice values below zero are very rarely observed. If they do occur, their interpretation is unambiguous: negative  $CC_{norm}$  implies anticorrelation between prediction and data.  $CC_{norm}$  thus behaves uniformly well whether called upon to quantify the performance of good and of poor models, in contrast to  $SPE$  which behaves well, and very similarly to  $CC_{norm}$ , for good models, but becomes increasingly harder to interpret as model performance declines.

## 6. EXPERIMENTAL VALIDATION

The previous sections show the problems caused by the missing lower bound of  $SPE$  from a theoretical point of view and illustrate them with a simulation (Figure 1). This section demonstrates the implications from a practical point of view by comparing the predictive performance of models for the activity of single neurons in the auditory system in three different experimental settings.

### 6.1. Neural Recordings

All animal procedures were approved by the local ethical review committee and performed under license from the UK Home Office. Ten adult pigmented ferrets (seven female, three male; all >6 months of age) underwent electrophysiological recordings under anesthesia. Full details are as in the study by Bizley et al. (2010). Briefly, we induced general anesthesia with a single intramuscular dose of medetomidine (0.022 mg/kg/h) and ketamine (5 mg/kg/h), which was maintained with a continuous intravenous infusion of medetomidine and ketamine in saline. Oxygen was supplemented with a ventilator, and we monitored vital signs (body temperature, end-tidal  $CO_2$ , and the electrocardiogram) throughout the experiment. The

temporal muscles were retracted, a head holder was secured to the skull surface, and a craniotomy and a durotomy were made over the auditory cortex. We made extracellular recordings from neurons in primary auditory cortex (A1) and the anterior auditory field (AAF) using silicon probe electrodes (Neuronexus Technologies) with 16 or 32 sites (spaced at 50 or 150  $\mu m$ ) on probes with one, two, or four shanks (spaced at 200  $\mu m$ ). We clustered spikes off-line using klustakwik (Kadir et al., 2014); for subsequent manual sorting, we used either spikemonger (an in-house package) or klustaviewa (Kadir et al., 2014). The time-discrete neuronal firing rate was approximated by binning spikes in 5 ms windows and averaging the spike count in each bin over all trials (compare to Equation 1).

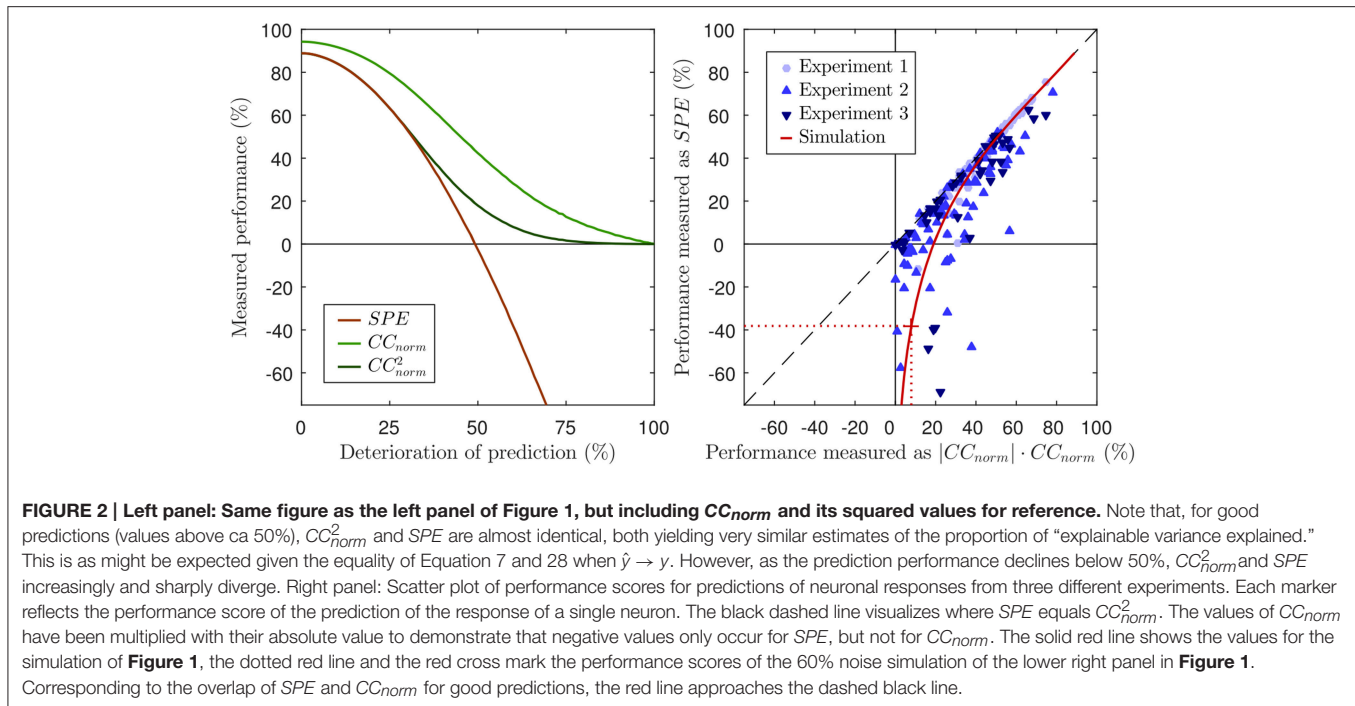
### 6.2. Acoustic Stimuli

Natural sounds were presented via Panasonic RPHV27 earphones, which were coupled to otoscope specula that were inserted into each ear canal, and driven by Tucker-Davis Technologies System III hardware (48 kHz sample rate). The sounds had root mean square intensities in the range of 75–82 dB SPL. For Experiment 1, we presented 20 sound clips of 5 s duration each, separated by 0.25 s of silence. Sound clips consisted of animal vocalizations (ferrets and birds), environmental sounds (water and wind) and speech. The presentation of these stimuli was repeated in 20 trials. For Experiments 2 and 3, we presented 45 sound clips of 1 s duration, again separated by gaps of silence. The sound clips consisted of animal vocalizations (sheep and birds), environmental sounds (water and wind) and speech. The presentation of these stimuli was repeated in 10 trials. The silent gaps and the first 0.25 s thereafter have been removed from the data set.

### 6.3. Neuronal Modeling

For Experiment 1, the responses of 119 single neurons were predicted with an LN model, a widely used class of models comprising a linear and a nonlinear stage (Chichilnisky, 2001; Simoncelli et al., 2004). The linear stage fits a spectro-temporal receptive field (STRF), which is a linear filter that links the neuronal response to the stimulus intensities of 31 log-spaced frequency channels (with center frequencies ranging from 1 to 32 kHz) along the preceding 20 time bins (covering a total of 100 ms stimulus history). The linear stage was fitted using GLMnet for Matlab (Qian et al.; see [http://web.stanford.edu/~hastie/glmnet\\_matlab/](http://web.stanford.edu/~hastie/glmnet_matlab/)). The nonlinear stage fits a sigmoidal nonlinearity to further maximize the goodness of fit to the neural response using minFunc by Mark Schmidt (University of British Columbia, British Columbia, Canada; <http://www.di.ens.fr/~mschmidt/Software/minFunc.html>). For Experiment 2, the same model class was used to predict the response of 77 single neurons. For Experiment 3, the responses of 43 single neurons were model with a standard neural network comprising 620 units in the input layer (31 frequency channels times 20 time bins of stimulus history), 20 hidden units and a single output unit. Hidden units and the output unit comprised a fixed sigmoidal nonlinearity. The connection weights of the





network were fitted with backpropagation using the Sum-of-Functions Optimizer (Sohl-Dickstein et al., 2013). Both, the STRF weights of the LN models and the connection weights of the neural networks were regularized with a penalty term on the L2-norm in order to avoid overfitting. In all cases, models were trained and tested using a cross-validation procedure. All free model parameters were fitted on a training set comprising 90% of all data. The predictive performance of a model for a given neuron was assessed by measuring  $SPE$  and  $CC_{norm}$  for the model predictions of the neural response to the remaining 10% of the data set. This procedure was repeated 10 times, each time with a distinct 10% of data. The model performance was computed as the mean across all 10 performance measurements.

## 6.4. Results

We predicted neuronal responses to acoustic stimuli with different model classes in order to address the question how the choice of a performance measure affects the interpretability of the results in a practical setting. To this end, we measured the predictive performance of models with two different methods,  $SPE$  and  $CC_{norm}$ . The right panel of Figure 2 shows a scatter plot in which each marker indicates the performance scores that the respective measures assign to a given prediction for a given neuron. Instead of raw  $CC_{norm}$  values, here we chose to plot the signed square of  $CC_{norm}$  as a percentage on the x-axis. This choice is motivated by the fact that the square of the correlation coefficient, also known as the coefficient of determination, quantifies the “proportion of variance explained” by a statistical regression model, and  $CC_{norm}^2 \times 100$  should thus be interpretable directly as a measure of “percent explainable variance explained” by the model. We plot the signed square

to ensure that there are no artificial constraints keeping the  $x$ -values positive: the fact that there  $x$ -range of the data is entirely positive while the  $y$ -range extends well into negative territory veridically reflects the way the respective underlying metrics,  $CC_{norm}$  and  $SPE$ , behave in practice. For those cases in which the model predicts the actual neuronal response quite well, one can observe a very tight relation between the  $SPE$  value and the signed squared value of  $CC_{norm}$ , i.e., both provide very similar, sensible measures of “percent explainable variance explained.” However, as expected from the theoretical analysis of both measures in the previous sections, this relation diminishes for cases in which the models poorly predicted the neuronal response. For those cases where there is little or no correspondence between the prediction and the response, the value of  $CC_{norm}$  approaches zero (by definition), and for some of those cases, the value of  $SPE$  also approaches zero, but for many others the  $SPE$  value becomes a large negative number. Substantially negative  $SPE$ s are seen even for some cases for which the  $|CC_{norm}| \times CC_{norm}$  indicates that the model was able to capture as much as 20–30% of the explainable, stimulus driven variability in the neural firing rate. Thirty percent variance explained may not be a stellar performance for a model, but it certainly does not seem deserving of a negative test score. Indeed, the experimental results are generally in accordance with the simulation in general, shown as a red line in the right panel of Figure 2. The simulation is identical to the one in Figure 1. To simulate  $SPE$  and  $CC_{norm}$  for a wide range of good and bad predictions, a good prediction was deteriorated by adding an increasing amount of white noise. Just as for the data from the three experiments,  $SPE$  values match the square of  $CC_{norm}$  for good predictions, but go deep into negative values for noisy predictions. For comparison, the  $SPE$  and  $CC_{norm}$  values of the example in the bottom right panel of Figure 1 (60%

noise added) are marked with dotted lines in the right panel of **Figure 2**. In summary, the analysis of the experimental data from three experiments validate the theoretical analysis of the previous sections.

**Figure 2** also visualizes the practical implications of the missing lower bound of *SPE*. *SPE* was from its inception described to be a “quantitative estimate of the fraction of stimulus-related response power captured by a given class of models” (Sahani and Linden, 2003). This interpretation is in conflict with values below zero because a fraction of a signal power cannot be negative. Furthermore, as was discussed in the previous sections, it is even difficult to assign an unambiguous interpretation to small or negative *SPE* values because a variety of poor models which vary widely in the size of their residual error can have similar small or negative *SPE*s, and may have *SPE*s below those of constant mean firing rate models of arbitrary value with an *SPE* of zero (including the “dead neuron model”), even if their residual error is smaller than that of these null models. If researchers are trying to quantify how well a particular class of models can describe the response properties of a sizeable sample population of neurons, a small number of somewhat spurious very negative values can heavily affect the overall population mean. For instance, the mean *SPE* value across the population of 77 neurons in Experiment 2 is just 15%, because a few very negative values drag down the average. But, as we have discussed in section 6, much of the negativity in those *SPE* values simply reflects a large variance in the predictions, which on its own is not very relevant, and constraining the *SPE* to values of zero or above would raise the mean performance by more than a quarter to over 19%.

## 7. CONCLUSION

Inter-trial variability of neural responses to repeated presentations of stimuli poses a problem for measuring the performance of predictive models. The neural variability inherently limits how similar one can expect the prediction of even a perfect model to be to the observed responses. Thus, when using prediction accuracy as a measure of performance, inherent response variability is a confound, and the need to control for this has been widely acknowledged (e.g., Panzeri and Treves, 1996; Sahani and Linden, 2003; Hsu et al., 2004; David and Gallant, 2005; Laudanski et al., 2012).

Different approaches for taking neural variability into account when measuring model performance have been developed. To get an unbiased estimate of *mutual information*, Panzeri and Treves (1996) have suggested a method to extrapolate information content to an infinite number of trials (also see Atencio et al., 2012). Sahani and Linden have developed the very insightful decomposition of the recorded signal into *signal power* and *noise power* (Sahani and Linden, 2003). This has lead to the *signal power explained (SPE)*, a measure based on *variance explained* which discounts “unexplainable” neural variability. This measure has been widely adopted, albeit under various names such as *predictive power*, *predicted response power*, and *relative prediction success* (Sahani and Linden, 2003;

Machens et al., 2004; Ahrens et al., 2008; Asari and Zador, 2009; Rabinowitz et al., 2012). Also, it has been used as a basis for specific variants of measures for model performance (Haefner and Cumming, 2009). Focusing on coherence and the correlation, Hsu and colleagues have developed a method to normalize those measures by their upper bound ( $CC_{max}$ ), which is given by the inter-trial variability (Hsu et al., 2004). This yields the *normalized correlation coefficient* ( $CC_{norm}$ ). Following their suggestion, the upper bound can be approximated by looking at the similarity between one half of the trials and the other half of the trials ( $CC_{half}$ ). This measure has also been used by Gill et al. (2006) and Touryan et al. (2005). Others have used the absolute correlation coefficient and controlled for inter-trial variability by comparing the absolute values with  $CC_{half}$  (Laudanski et al., 2012).

In this study we have analyzed in detail two measures of model quality that account for neural response variability, *SPE* and  $CC_{norm}$ . We have revealed the shortcomings of *SPE*, which has no lower bound and can yield undesirable negative values even for fairly reasonable model predictions. Furthermore, we have uncovered the close mathematical relationship between *SPE* and  $CC_{norm}$ , consolidated both approaches and arrived at several insights. First, both measures quantify prediction accuracy using the covariance (and *only* using covariance). Second, both measures quantify neural variability using the *signal power* (*SP*) (and *only* using *SP*). Third, when the variance of the prediction error approaches zero, *SPE* becomes identical to the square of  $CC_{norm}$ . And finally, it is not necessary to approximate  $CC_{max}$  using computationally expensive and inexact resampling methods because  $CC_{norm}$  can be calculated directly via *SP*:

$$CC_{abs} = \frac{Cov(y, \hat{y})}{\sqrt{Var(\hat{y})Var(y)}} \quad CC_{norm} = \frac{Cov(y, \hat{y})}{\sqrt{Var(\hat{y})SP}} \quad (28)$$

$$SP = \frac{Var\left(\sum_{n=1}^N R_n\right) - \sum_{n=1}^N Var(R_n)}{N(N-1)} \quad (29)$$

This consolidated definition of  $CC_{norm}$  is not only more elegant, precise, and efficient, but it also sheds light on how  $CC_{norm}$  can be interpreted. It is almost identical to the well-known Pearson's correlation coefficient  $CC_{abs}$ , but the variance (power) of the recorded signal is replaced with the *signal power* *SP*, i.e., the deterministic and thus predictable part of the signal. As demonstrated, using *SPE* as a measure of model performance can yield misleading results and will limit interpretability of the results. However,  $CC_{norm}$  has been shown to fulfill the criteria of Section 2 for insightful measures: it is bounded, interpretable, and comparable across data sets. Thus,  $CC_{norm}$  is a well-defined and helpful tool to assess model performance<sup>4</sup>.

<sup>4</sup>Matlab code for all measures can be found on GitHub: <https://github.com/OSchoppe/CCnorm>.

Note, however, that  $CC_{norm}$  cannot be estimated accurately if the data are excessively noisy. Equation 28 requires  $SP$  to be large enough to estimate with reasonable accuracy. For very noisy data or too few trials, observed  $SP$  values can become dominated by sampling noise, and may then behave as near zero random numbers. This would render  $CC_{norm}$  estimates unstable, allowing them to become spuriously large (if  $SP$  is small and underestimates the true value) or even imaginary (if the  $SP$  underestimate is severe enough to become negative). Thus, if  $SP$  or  $CC_{max}$  are small or have a very wide confidence interval,  $CC_{norm}$  values must be treated with caution.

## REFERENCES

- Ahrens, M. B., Linden, J. F., and Sahani, M. (2008). Nonlinearities and contextual influences in auditory cortical responses modeled with multilinear spectrotemporal methods. *J. Neurosci.* 28, 1929–1942. doi: 10.1523/JNEUROSCI.3377-07.2008
- Asari, H., and Zador, A. M. (2009). Long-lasting context dependence constrains neural encoding models in rodent auditory cortex. *J. Neurophysiol.* 102, 2638–2656. doi: 10.1152/jn.00577.2009
- Atencio, C. A., and Schreiner, C. E. (2013). *Stimulus Choices for Spike-Triggered Receptive Field Analysis*, Chapter 3. New York, NY: Nova Biomedical.
- Atencio, C. A., Sharpee, T. O., and Schreiner, C. E. (2012). Receptive field dimensionality increases from the auditory midbrain to cortex. *J. Neurophysiol.* 107, 2594–2603. doi: 10.1152/jn.01025.2011
- Bizley, J. K., Walker, K. M., King, A. J., and Schnupp, J. W. (2010). Neural ensemble codes for stimulus periodicity in auditory cortex. *J. Neurosci.* 30, 5078–5091. doi: 10.1523/JNEUROSCI.5475-09.2010
- Chichilnisky, E. (2001). A simple white noise analysis of neuronal light responses. *Network* 12, 199–213. doi: 10.1080/713663221
- David, S. V., and Gallant, J. L. (2005). Predicting neuronal responses during natural vision. *Network* 16, 239–260. doi: 10.1080/09548980500464030
- David, S. V., and Shamma, S. A. (2013). Integration over multiple timescales in primary auditory cortex. *J. Neurosci.* 33, 19154–19166. doi: 10.1523/JNEUROSCI.2270-13.2013
- Döerscheidt, G. H. (1981). The statistical significance of the peristimulus time histogram (PSTH). *Brain Res.* 220, 397–401. doi: 10.1016/0006-8993(81)91232-4
- Gill, P., Zhang, J., Woolley, S. M. N., Fremouw, T., and Theunissen, F. E. (2006). Sound representation methods for spectro-temporal receptive field estimation. *J. Comput. Neurosci.* 21, 5–20. doi: 10.1007/s10827-006-7059-4
- Haefner, R. M., and Cumming, B. G. (2009). “An improved estimator of variance explained in the presence of noise,” in *Advances in Neural Information Processing Systems 21*, eds D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou (Red Hook, NY: Curran Associates, Inc.), 585–592.
- Hsu, A., Borst, A., and Theunissen, F. (2004). Quantifying variability in neural responses and its application for the validation of model predictions. *Network* 15, 91–109. doi: 10.1088/0954-898X-15-2-002
- Kadir, S. N., Goodman, D. F., and Harris, K. D. (2014). High-dimensional cluster analysis with the masked em algorithm. *Neural Comput.* 26, 2379–2394. doi: 10.1162/NECO-a-00661
- Kass, R. E., Ventura, V., and Cai, C. (2003). Statistical smoothing of neuronal data. *Network* 14, 5–16. doi: 10.1088/0954-898X/14/1/301
- Laudanski, J., Edeline, J.-M., and Huetz, C. (2012). Differences between spectro-temporal receptive fields derived from artificial and natural stimuli in the auditory cortex. *PLoS ONE* 7:e50539. doi: 10.1371/journal.pone.0050539
- Machens, C. K., Wehr, M. S., and Zador, A. M. (2004). Linearity of cortical receptive fields measured with natural sounds. *J. Neurosci.* 24, 1089–1100. doi: 10.1523/JNEUROSCI.4445-03.2004
- Marmarelis, P. (1978). *Analysis of Physiological Systems: the White-Noise Approach*. New York, NY: Plenum Press. doi: 10.1007/978-1-4613-3970-0
- Panzeri, S., and Treves, A. (1996). Analytical estimates of limited sampling biases in different and information measures. *Network* 7, 87–107. doi: 10.1088/0954-898X/7/1/006
- Prenger, R., Wu, M. C.-K., David, S. V., and Gallant, J. L. (2004). Nonlinear V1 responses to natural scenes revealed by neural network analysis. *Neural Netw.* 17, 663–679. doi: 10.1016/j.neunet.2004.03.008
- Rabinowitz, N. C., Willmore, B. D. B., Schnupp, J. W. H., and King, A. J. (2011). Contrast gain control in auditory cortex. *Neuron* 70, 1178–1191. doi: 10.1016/j.neuron.2011.04.030
- Rabinowitz, N. C., Willmore, B. D. B., Schnupp, J. W. H., and King, A. J. (2012). Spectrotemporal contrast kernels for neurons in primary auditory cortex. *J. Neurosci.* 32, 11271–11284. doi: 10.1523/JNEUROSCI.1715-12.2012
- Roddey, J. C., Girish, B., and Miller, J. P. (2000). Assessing the performance of neural encoding models in the presence of noise. *J. Comput. Neurosci.* 8, 95–112. doi: 10.1023/A:1008921114108
- Sahani, M., and Linden, J. F. (2003). “How linear are auditory cortical responses?” in *Advances in Neural Information Processing Systems 15*, Vol. 15, eds S. Becker, S. Thrun, and K. Obermayer (MIT Press), 109–116.
- Shimazaki, H., and Shinomoto, S. (2007). A method for selecting the bin size of a time histogram. *Neural Comput.* 19, 1503–1527. doi: 10.1162/neco.2007.19.6.1503
- Simoncelli, E. P., Paninski, L., Pillow, J., and Schwartz, O. (2004). “Characterization of neural responses with stochastic stimuli,” in *The Cognitive Neurosciences, 3rd Edn.*, ed M. Gazzaniga (Cambridge, MA: MIT Press), 327–338.
- Sohl-Dickstein, J., Poole, B., and Ganguli, S. (2013). “Fast large-scale optimization by unifying stochastic gradient and quasi-newton methods,” in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, eds T. Jebara and E. P. Xing (Beijing), 604–612.
- Touryan, J., Felsen, G., and Dan, Y. (2005). Spatial structure of complex cell receptive fields measured with natural images. *Neuron* 45, 781–791. doi: 10.1016/j.neuron.2005.01.029

## 8. AUTHOR CONTRIBUTIONS

OS: initiated the project; developed methodology; wrote and tested code implementing methods; analyzed method performance both analytically and through experiment; lead author on paper. NH, BW, AK, JS: guided research, co-wrote manuscript.

## ACKNOWLEDGMENTS

This work was supported by a Wellcome Trust grant (WT076508AIA) and a BBSRC grant (BB/H008608/1). OS was supported by the German National Academic Foundation.

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Schoppe, Harper, Willmore, King and Schnupp. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.