Network: Computation in Neural Systems
March–June 2010; 21(1–2): 91–124

Multilinear models of single cell responses in the medial nucleus of the trapezoid body

B. ENGLITZ[1,2]†, M. AHRENS[3]†, S. TOLNAI[2]‡, R. RÜBSAMEN[2], M. SAHANI[3], & J. JOST[1]

[1]*Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, Leipzig, Germany,* [2]*Faculty of Biosciences, Pharmacy and Psychology, University of Leipzig, Talstraße 33, Leipzig, Germany, and* [3]*Gatsby Computational Neuroscience Unit, University College London, London, UK*

(Received 21 December 2009; accepted 23 March 2010)

Abstract

The representation of acoustic stimuli in the brainstem forms the basis for higher auditory processing. While some characteristics of this representation (e.g. tuning curve) are widely accepted, it remains a challenge to predict the firing rate at high temporal resolution in response to complex stimuli.

In this study we explore models for in vivo, single cell responses in the medial nucleus of the trapezoid body (MNTB) under complex sound stimulation. We estimate a family of models, the multilinear models, encompassing the classical spectrotemporal receptive field and allowing arbitrary input-nonlinearities and certain multiplicative interactions between sound energy and its short-term auditory context. We compare these to models of more traditional type, and also evaluate their performance under various stimulus representations.

Using the context model, 75% of the explainable variance could be predicted based on a cochlear-like, gamma-tone stimulus representation. The presence of multiplicative contextual interactions strongly reduces certain inhibitory/suppressive regions of the linear kernels, suggesting an underlying nonlinear mechanism, e.g. cochlear or synaptic suppression, as the source of the suppression in MNTB neuronal responses. In conclusion, the context model provides a rich and still interpretable extension over many previous phenomenological models for modeling responses in the auditory brainstem at submillisecond resolution.

Keywords: *Auditory system, sound localization, neuronal encoding*

Correspondence: B. Englitz, Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, 04317 Leipzig, Germany. E-mail: benglitz@gmail.com

†These authors contributed equally.

‡Present address: Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford OX1 3PT, UK.

ISSN 0954-898X print/ISSN 1361-6536 online/10/01-020091–124 © 2010 Informa Healthcare Ltd.
DOI: 10.3109/09548981003801996

Introduction

Projections of the medial nucleus of the trapezoid body (MNTB) are vital for computing sound source locations in the subsequent centers of sound localization, the medial and lateral superior olives (MSO, LSO, Moore and Caspary, 1983; Brand et al., 2002; Pecka et al., 2008). Modeling sound localization in the MSO and LSO might therefore require an accurate model of MNTB responses. In a previous study (Brand et al., 2002), a model of auditory nerve fibers (ANFs, Carney, 1993) has been substituted for MNTB responses. However, this model does not include the integration on the way from the ANFs to the MNTB: first, a convergence of ANFs occurs on the globular bushy cells of the cochlear nucleus, and second, the responses could be processed during synaptic transmission at the calyx of Held in the MNTB (Awatramani et al., 2004). In both locations the signal representation could be influenced by changes in rate or timing, induced by convergent excitatory and inhibitory inputs as well as cellular and synaptic dynamics (Kopp-Scheinpflug et al., 2002).

In the present study we estimate a family of phenomenological models, in each case evaluating their performance. Performance measures are important in neuronal modeling, because drawing functional conclusions from poorly-fitting models may be misleading. In contrast, interpretations of well-fitting models are far more likely to be relevant to neuronal function. Phenomenological models are useful in that they do not make many assumptions. On the downside, they often lack mechanistic interpretability; however, the models estimated here do allow us to address several important structural aspects, including stimulus representation, nonlinear scaling, and the spectral and temporal relation of suppressive contextual influences.

The multilinear models (Ahrens et al., 2008a) utilized in the present study have been shown to improve predictions for firing rate models of neurons in the auditory cortex. This class of models provides a principled and parsimonious extension over classical models, such as spectrotemporal receptive fields (STRF, Aertsen et al., 1981a; Kim and Young, 1994). Multilinear models offer a rich framework, providing flexible kernel choices in the time, frequency, and stimulus level domain with the possibility to investigate separability between each of the dimensions, thereby controlling the number of degrees of freedom. Arbitrary input nonlinearities can be estimated alongside classical time and frequency kernels. Multiplicative interactions between two spectrotemporal locations can also be implemented in the so-called context model, via the "contextual reweighting field" (CRF). Further, data-driven, Bayesian regularization schemes such as automatic smoothness control (Sahani and Linden, 2003a) are readily available. For the present study these models were adapted to the requirements of the auditory brainstem, including a 50-fold higher temporal resolution, rendering the description close to the regime of temporal coding.

We compared the performance of a variety of multilinear models based on three different stimulus representations. We find that a cochlea-like stimulus representation in conjunction with the full-fledged context model provides the best performance. On average 75% of the explainable power was predictable as a function of the stimulus. In comparison to our reference model, the STRF, the final context model provided an increase in predictive power of about 35%. The final context model had combined time-frequency dimensions, nonlinear scaling functions on the input and the output as well as an optimized choice of sampling rate and stimulus representation.

Methods

Experimental procedures

All experimental procedures were approved by the Saxonian District Government, Leipzig. The physiological methods have been described in greater detail in Tolnai et al. (2009) and are only provided biefly here.

Preparation. Thirty adult pigmented Mongolian gerbils (Meriones unguiculatus, aged 2–4 months, weighing 45–70 g) were used in the present study. Anesthesia was initialized by an intraperitoneal dose of xylazine-ketamine and maintained during the experiment by hourly, subcutaneous injections. The skull of the animal was exposed along the midsagittal line and a metal bolt was glued to the bone on bregma and stabilized with dental cement. Two holes (centered and 1.5 mm lateral, $\varnothing$ 0.5 mm) drilled into the occipital bone 2–2.3 mm caudal to the lambdoid suture allowed the insertion of a recording electrode (glass micropipette, 3 M KCl, 5–15 MΩ) and a reference electrode (silver wire, WPI) in the superficial cerebellum. Animals were placed in a sound-attenuated booth (Type 400, Industrial Acoustic Company) on a vibration-isolated table and positioned in a stereotaxic device using the metal bolt. The MNTB was approached dorsally with the animal tilted at 4–10° to the midsagittal plane.

Neuronal recordings. Stereotaxic coordinates of the MNTB were determined by online analysis of acoustically evoked multi-unit activity using low impedance micropipettes (<5 MΩ). Differentiation of the MNTB from other nuclei within the superior olivary complex was facilitated by the exclusively contralateral excitatory input to MNTB units. Single-unit, extracellular voltage recordings were performed using high-impedance glass micropipettes (8–30 MΩ, GB150TF-10, Science Products) filled with 3 M KCl. Single units were identified by the characteristic shape of their waveform. The complex of the calyx of Held (presynaptic) and the principal cells of the MNTB (postsynaptic) produces complex waveforms which distinguish them from other cell types and fibers in this nucleus (Guinan and Li, 1990; Englitz et al., 2009). Only units exhibiting such a complex waveform were included in the analysis. The voltage signal was preamplified (Neuroprobe 1600, A-M Systems, Carlsborg), band pass filtered (0.3–7 kHz), and further amplified (PC1, TDT, Alachua) to match the input voltage range of the A/D converter (RP2.1, TDT). Voltage traces were digitized (sampling rate $SR_{rec} = 97.7$ kHz) and stored for subsequent analysis.

Acoustic stimulus generation. Stimulus waveforms were generated at 97.7 kHz using custom written software (Matlab 7.3, The Mathworks, Nattick). Stimuli were then transferred to a real-time processor (RP2.1, TDT), D/A converted, and further sent to a speaker (DT 770 pro, Beyerdynamic). Sound from the speaker was funneled into a plastic tube (35 mm length, 5 mm diameter) whose other end was inserted into the outer ear canal at a distance of ~4 mm to the tympanic membrane. Acoustic calibration was performed by convolving the stimulus with the earphone's inverse impulse response prior to stimulus presentation. The impulse response of the system was estimated by presenting a 10 s white noise stimulus and computing the real part of the inverse Fourier transform of the frequency transfer function (Matlab function: *tfe*) between the original and recorded waveform (recorded with a condenser microphone, Bruel & Kjær type 2618). The calibration was subsequently verified to lie within ±5 dB of the target amplitude in the range of 0.5 to 48 kHz before the experiment.

Modeling

The functional relationship modeled here is the translation of an auditory stimulus S to a vector of instantaneous firing rates $\mathbf{r}$. The actual firing rate $\mu(i)$ at time i is not directly observed but estimated from the available trials as $r(i) = \overline{r_n(i)} = \frac{1}{N}\sum_{n=1}^{N} r_n(i) = \frac{1}{N\Delta t}\sum_{n=1}^{N} c_n(i)$, where N is the number of trials, Δt the time-step, $c_n(i)$ the counts and $r_n(i)$ the rate in the n-th trial at time i and $\bar{\cdot}$ denotes trial average. The translation is captured by a neuronal response model consisting of three stages (schematically depicted in Fig. 1(A)): (i) a time-frequency representation of the sound pressure wave, (ii) a multilinear model over the dimensions time, frequency and level, and (iii) a static output nonlinearity.

A time-frequency representation is included in the model to account for the fact that the auditory system itself generates a time-frequency representation from the scalar sound pressure wave at each ear. Motivated by the transformation properties of the cochlea a number of time-frequency representations of the sound have been in use for modeling neuronal responses, e.g. the spectrogram (Aertsen et al., 1981a), the Wigner transform (Kim and Young, 1994) and other cochlea-inspired transforms (e.g. Elhilali et al. 2004). We here compare the effectiveness of three different representations to account for the neuronal response (in combination with the following model architecture). The time-frequency representation is then filtered by a multilinear model (Ahrens et al., 2008a), which constitutes a family of models in the dimensions time, frequency, and level. This family includes several models, e.g. the spectrotemporal receptive field (STRF, Aertsen et al., 1981a) and the time-static models introduced by Young and coworkers (Yu and Young, 2000; Young and Calhoun, 2005; Bandyopadhyay et al., 2007). Finally, a sigmoidal output nonlinearity constrains the output of the model to obey the limits of neuronal discharge rates. These stages are detailed below after introducing the probe stimulus.

Broadband stimulus

The acoustic stimulus was a broadband sound with a prescribed spectrotemporal profile of amplitude modulations. Its construction was a modified version of the TORC (temporally orthogonal ripple combination) method (Klein et al., 2000) which directly prescribes the modulation spectrum and constructs the acoustic stimulus from this prescription. This method allows a restriction of the available

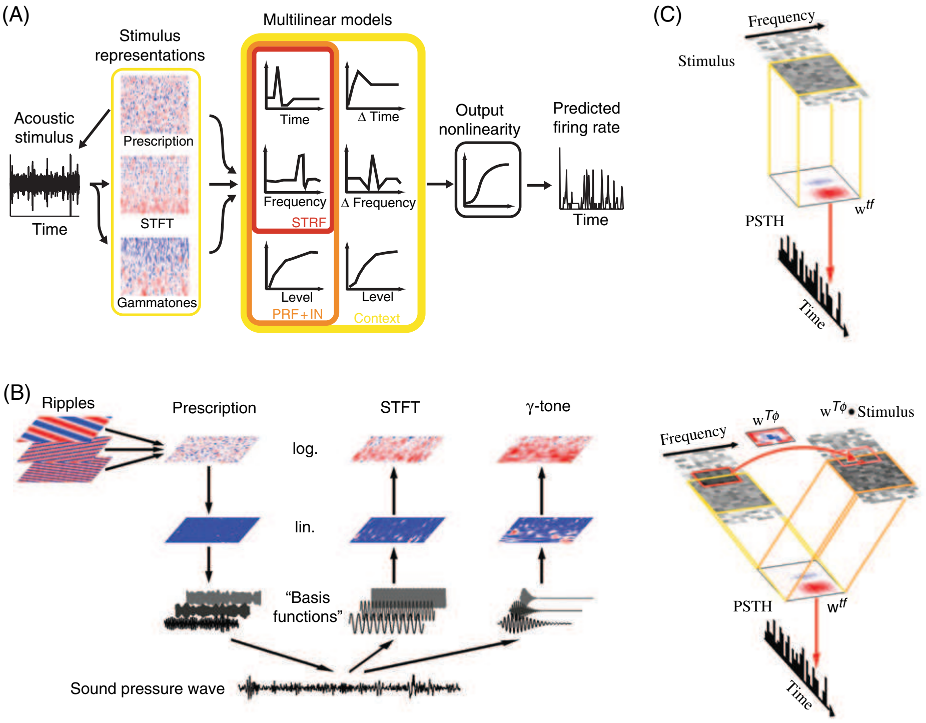

Figure 1. Overview of the overall model structure and stimulus representation. (A) Schematic overview of estimated models. An acoustic stimulus is created from the spectrotemporal prescription. This prescription or two other spectrotemporal representations of the stimulus are used as input for the following models: First, a multilinear model of dimensions time, frequency, and level is estimated, e.g. a STRF, an input nonlinearity model (IN) or a context model. Second, an estimated output linearity rescales the multilinear prediction to the final firing rate prediction. We compare the performance contributed by the individual parts. (B) Construction of the different stimulus representations. The original TORC stimulus (Prescription) is generated as the sum of 217 spectrotemporal ripples. This representation is assumed to be logarithmically scaled with respect to the sound pressure level. Rescaling leads to the linearly scaled spectrotemporal representation (middle row). Using this stimulus as an instantaneous weighting of sinusoidal carriers leads to the acoustic stimulus (bottom). The short-term Fourier transform (STFT) and the γ-tone representation (right) are computed from the acoustic stimulus based on their respective sets of time-restricted filtering functions. From these linear representations a logarithmic scaling provides the representations corresponding to the original TORC. (C) Schematic of the STRF and the context model. In the STRF model (top) the stimulus is weighted by the time-frequency kernel $\mathbf{w}^{\mathbf{ft}}$ (middle) which produces the rate prediction (bottom). In the context model (bottom) a second time-frequency filter $\mathbf{w}^{\tau\phi}$ transforms the stimulus (left) by a similar weighting around a given time-frequency point (dashed red rectangle around red point on the right). Both the original and the transformed stimulus are then weighted by $\mathbf{w}^{\mathbf{ft}}$ (middle) producing the rate prediction (bottom). The colored and darkened rectangles illustrate the weighting regions for each kernel. For displaying purposes the optional inclusion of input and output nonlinearities have been omitted here.

stimulus energy to a limited range of frequency channels and therefore produces higher modulation depths per channel. Adapting the TORC method for use in brainstem structures essentially required increasing the maximal temporal modulation rates to 800 Hz (from ≈40 Hz typical for the cortex).

Amplitude modulations are better suited than fine structure modulations to investigate the responses of MNTB units due to their dominantly high characteristic frequencies (CF). These units have been shown to modulate their firing rate in response to amplitude modulations of a carrier tone (Joris and Yin, 1998; Tolnai et al., 2008).

Construction of the TORC stimulus. In accordance with Klein et al. (2000), a TORC stimulus is a sum of spectrotemporal ripples (Fig. 1(B)), which are orthogonal in the time-domain (intended to ideally decouple their effects on the neuronal response). The i-th spectrotemporal ripple is defined as a function of time t and logarithmic frequency x

$$R_i(t,x) := R(t,x,\{w_i,\Omega_i,\phi_i\}) = \cos(2\pi(w_i t + \Omega_i x) + \phi_i)$$

where w_i denotes the i-th temporal modulation frequency, Ω_i the i-th spectral modulation frequency, and ϕ_i the i-th phase. Temporal orthogonality is defined via

$$\int_0^T R_i(t,x)\, R_j\,(t,x)\mathrm{d}t = 0, \quad \text{for } i \neq j$$

and holds as long as the temporal modulation frequencies differ in absolute value (assuming $T \gg \max(w_i^{-1}, w_j^{-1})$). The whole TORC stimulus is then given by

$$S_{\mathrm{dB}}(t,x,\{\mathbf{x},\mathbf{k},\phi\}) = \sum_{i=1}^{N} R(t,x,\{w_i,\Omega_i,\phi_{i,i}\}).$$

For the ripple parameters the following ranges were chosen:

- temporal modulation frequencies w_i: $[-800, 800]$ Hz in 15 steps.
- spectral modulation frequencies Ω_i: $[0, 4.2]$ oct^{-1} in 16 steps.
- phases $\phi_{i,i}$: $0 - 2\pi$ radians, randomly drawn for each ripple.

These parameters covered the upper half-plane of the spectrotemporal modulation space approximately uniformly. As in Klein et al. (2000), the w_i were shifted by different amounts for each Ω_i to avoid w_i's of same absolute value (see Fig. 9(B) in Klein et al. (2000) for a depiction of the stimulus arrangement.). Spectral and temporal modulation frequencies were additionally jittered (normal distribution with a S.D. of 25% of the spacing in the respective dimension) around the commonly used equal spacing to reduce long term correlations in the stimulus (due to commensurate modulation frequencies). Caution was taken to preserve temporal orthogonality, i.e. no absolute values of two w_i were allowed closer than 1 Hz.

The overall stimulus duration T was set to 5 s, i.e. >40 repetitions of the lowest temporal modulation frequency (7 Hz). In the spectral direction the stimulus spanned 3 octaves at a spacing of 0.1 octaves, amounting to a total of 31 frequency channels. Altogether 217 ripples were summed to obtain the final TORC stimulus ($7*15$ had negative w_i, $7*16$ ripples had positive w_i, where the $\Omega_i = 0$ were contained in the latter set, Fig. 1(B)). The distribution in each channel was nearly Gaussian with S.D. ≈ 8.75 dB (which has been termed spectral contrast in Reiss et al. (2007) and Bandyopadhyay et al. (2007)).

Since MNTB cells usually exhibit approximately linear scaling of response rate w.r.t. dB-scaled loudness, the amplitude modulation was by design linear in dB and converted to sound pressure in the process of generating the sound pressure waveform. Each amplitude was then logarithmically rescaled via the transformation $S_{\mathrm{Pa}}(x, t) = 10^{(S_{\mathrm{dB}}(x,t)-A_0)/20}$, where A_0 was the loudness at which calibration had been performed (80 dB). The resulting amplitudes were used for pointwise scaling of each frequency channel, i.e. each sinusoidal carrier (phase uniformly randomized) was amplitude scaled individually and the resulting signals added to produce the acoustic stimulus.

To aid comparability, the same TORC profile of amplitude modulations was used for all cells, however, on top of a different range of carrier frequencies adapted to a unit's CF. Further, the overall amplitude of the stimulus was chosen to drive the neuron to high discharge rates (132 ± 38 Hz, $n = 96$).

Choice of carrier frequencies. The choice of carrier frequencies was adjusted to each cell's CF (estimated prior to generating the TORC stimulus using a set of brief tonal stimuli). Carrier frequencies encompassed 3 octaves ideally reaching from two octaves below to one octave above the CF. This asymmetric spacing reflects the asymmetric shape of the (logarithmic) frequency tuning of MNTB principal cells. If the CF was closer than two octaves to the lower (0.5 kHz) or one octave to the upper bound (48 kHz), the frequency range reached from the respective boundary three octaves into the admissible frequency range.

Stimulus representation

The choice of stimulus representation can be an important determinant for the predictive quality of the overall model. We therefore tested three different stimulus representations to determine the one most suitable for prediction with the multilinear models. We compared the short-term Fourier transform (STFT) and a γ-tone representation to the TORC representation described above (termed 'Prescription' in the following).

STFT. The STFT performs a windowed Fourier transform (Fig. 1(B) middle). We used a Hamming window of length $L_{\mathrm{win}} \approx 1$ ms with an overlap $W_{\mathrm{over}} = L_{\mathrm{win}} - \mathrm{SR}_{\mathrm{in}}^{-1}$. This overlap effectively reduces the sampling rate by a desired factor $\mathrm{SR}_{\mathrm{rec}}/\mathrm{SR}_{\mathrm{in}}$, where $\mathrm{SR}_{\mathrm{in}}$ is the internal sampling rate of the model (see below). At $\mathrm{SR}_{\mathrm{rec}} = 97.65625$ kHz, L_{win} stretches 100 bins, i.e. the STFT will divide the frequency range equally in 50 steps from 0 Hz to the Nyquist-limit. Due to the logarithmic spacing of frequencies in the TORC, it is advisable to attempt to match these frequencies with the original frequencies, effectively sparsening the linearly spaced frequencies at the high frequency end. For comparability with the other stimulus representations, the same number of frequencies (31) was selected. Since phase information is likely to not be represented above 2.5 kHz, we used only the absolute value of the spectrogram.

γ-tone representation. The γ-tone representation was inspired by the filter types typically used for modeling early stages of cochlear signal transduction (Fig. 1(B) right). Specifically, the sound pressure was first passed through a filter bank of

γ-tones centered at the original carrier frequencies of the TORC. Each γ-tone's impulse response is given by

$$\gamma_{f,\tau}(t) = t^{\gamma-1} e^{-t/\tau(f)} \cos(2\pi f\, t)$$

with $\gamma = 3$, $\tau(f) = Q/f$, and $Q = 2.2$, similar to the cochlear model of Zhang et al. (2001). Cochlear latencies were not modeled directly to keep the delays in the estimated kernels absolutely interpretable. Next, the amplitude of each γ-tone channel was estimated by taking the absolute value of the channel's analytical signal A (Matlab function: *hilbert*). Together this transform is given by

$$S_{\gamma}(t,f) = |A(\gamma_{f,\tau} * S_{\mathrm{SPW}})|$$

All transformations were performed at $\mathrm{SR}_{\mathrm{rec}}$ and later downsampled to $\mathrm{SR}_{\mathrm{in}}$.

Amplitude scaling. On top of these qualitatively different representations, the appropriate scaling of the stimulus could influence the effectiveness of a given model (Escabi et al., 2003). Although by far not exhaustive, we compared both linear and logarithmic scaling of the amplitudes for each representation. A similar comparison of stimulus representations has been conducted in several avian auditory nuclei by Gill et al. (2006), with only partially consistent results as to the effectiveness of each representation.

Sampling rate. Finally, $\mathrm{SR}_{\mathrm{in}}$ can influence the predictive power. $\mathrm{SR}_{\mathrm{in}}$ ranging from $\approx$1 to $\approx$5 kHz were compared with respect to their predictive quality for the separated model (see below & Results, Fig. 3). A $\mathrm{SR}_{\mathrm{in}}$ of $\approx$2.2 kHz was found to maximize the predictive quality and therefore used in estimating all subsequent models. This resolution is fine enough to guarantee that each bin contains only one spike per trial, leading to binomial sampling distributions. For comparability the same $\mathrm{SR}_{\mathrm{in}}$ was chosen for all cells. Unfortunately this resolution is still too low to resolve interaural time differences (ITDs). Resolving ITDs in the MNTB could be required, given that its projections to the MSO/LSO are important for computing sound location (Brand et al., 2002; Pecka et al., 2008), possibly based on ITDs of the stimulus envelope (Bernstein, 2001; Bernstein and Trahiotis, 2002).

Multilinear models

We have investigated a hierarchy of models which differ in their model order (first or second order), separability of dimensions (separated or combined), and input scaling (linear or nonlinear). This range of models is accommodated within the framework of multilinear models (Ahrens et al., 2008a). Details on model structure, estimation, and regularization have been provided previously (Sahani and Linden, 2003a; Ahrens et al., 2008a) and are therefore only given briefly.

All models are conveniently described in the language of k-dimensional arrays which we either denote as tensors $\mathbf{M}^{\mathbf{a}\cdots\mathbf{z}}$ (bold fonts) or element-wise $M^{\mathbf{a}\ldots\mathbf{z}}_{a\ldots z}$ (regular fonts), where the upper letters specify the 'physical' dimensions of the tensor. For example, a time kernel will be denoted as $\mathbf{w}^{\mathbf{t}}$, a time-frequency kernel as $\mathbf{w}^{\mathbf{tf}}$, and a multi-dimensional stimulus matrix as $\mathbf{M}^{\mathbf{itfl}}$, where the letters **i**, **t**, **f**, **l**, represent time, response delay, frequency, and level, respectively. The lower letters specify the current (multi)index in the respective dimensions. All indices run from 1 to a maximal integer, usually denoted by the corresponding capital letter, e.g. i usually runs from 1 to I. Two tensor operations are required for the present model, briefly (a more detailed description is provided in Appendix B): (i) the generalized inner product sums out certain dimensions (e.g. for generating a prediction), for example a 2D matrix can be contracted with a 1D kernel to yield another 1D kernel

$$\mathbf{Q}^{\mathbf{f}} = \mathbf{w}^{\mathbf{t}} \bullet \mathbf{M}^{\mathbf{tf}} \text{ with } Q^{\mathbf{f}}_{k} = \sum_{j} w^{\mathbf{t}}_{j} M^{\mathbf{tf}}_{jk}.$$

(ii) The generalized outer product combines smaller tensors into larger ones (e.g. for generating a less separated model), for example a 2D kernel can be constructed from two 1D kernels

$$\mathbf{w}^{\mathbf{tf}} = \mathbf{w}^{\mathbf{t}} \otimes \mathbf{w}^{\mathbf{f}} \text{ with } w^{\mathbf{tf}}_{jk} = w^{\mathbf{t}}_{j} w^{\mathbf{f}}_{k}.$$

The multilinear models linearize potentially nonlinear transformations by representing the stimulus in a way that allows formulating and estimating a number of nonlinear transformations via linear operations. Usually this is achieved by expressing nonlinear or noninstantaneous maps as linear combinations of suitable basis functions. In general, a multilinear model is given by

$$\hat{r}^{\mathbf{i}}_{i} = \sum_{j_1,\ldots,j_n} w^{k_1}_{j_1} \ldots w^{k_n}_{j_n} Q^{\mathbf{ik_1\ldots k_n}}_{ij_1,\ldots,j_n} \quad \text{or} \quad \hat{\mathbf{r}}^{\mathbf{i}} = (\mathbf{w}^{\mathbf{k_1}} \otimes \cdots \otimes \mathbf{w}^{\mathbf{k_n}}) \bullet \mathbf{Q}^{\mathbf{ik_1\ldots k_n}},$$

where the $\mathbf{w}^{\mathbf{k}_\bullet}$ denote the linear kernels for dimension $\mathbf{k}_\bullet$ and $\mathbf{Q}^{\mathbf{ik_1\ldots k_n}}$ the new stimulus representation. The index i for the time dimension $\mathbf{i}$ runs from 1 to I. While the model structure is simple and the kernels are directly interpretable, the heart of a multilinear model lies in the representation $\mathbf{Q}^{\mathbf{ik_1\ldots k_n}}$.

The set of multilinear models estimated here is almost the same as in Ahrens et al. (2008a) and will therefore be described only briefly. Broadly, two subsets of models are distinguished, the input nonlinearity models and the context models.

Input nonlinearity models. This set of models includes an arbitrary transformation of the stimulus level in addition to a linear mapping from the spectrotemporal stimulus representation. The neuronal response is then represented as

$$\hat{\mathbf{r}} = (\mathbf{w}^{\mathbf{t}} \otimes \mathbf{w}^{\mathbf{f}} \otimes \mathbf{w}^{\mathbf{l}}) \bullet \mathbf{Q}^{\mathbf{itfl}},$$

where $\mathbf{Q}^{\mathbf{itfl}}$ is a representation of the stimulus as a function of peristimulus time $\mathbf{i}$, lag time $\mathbf{t}$, stimulus frequency $\mathbf{f}$, and level $\mathbf{l}$. The level representation is based on a basis function representation of the stimulus. For the present study these basis functions simply constituted a binning of the range of levels. The kernels $\mathbf{w}^{\mathbf{t}}$, $\mathbf{w}^{\mathbf{f}}$, and $\mathbf{w}^{\mathbf{l}}$ correspond to the respective stimulus dimension and are combined by the outer product to a matrix compatible in size for the inner product with $\mathbf{Q}^{\mathbf{itfl}}$.

Context models. In contrast to the input nonlinearity model, stimulus features at different times or frequencies can be combined multiplicatively in the context model. Intuitively, these can be thought of as a local neighborhood in the stimulus – the context – which modifies each point in the stimulus multiplicatively, before the principal filters are applied to produce the response (Fig. 1(C)). This local modification is another linear kernel on the spectrogram – the contextual reweighting field (CRF, Ahrens et al., 2008a). As in the input nonlinearity model the CRF can be extended to contain another weighting of stimulus level. In the present study the principal filters were either given by a spectrotemporal kernel (similar to an STRF) or an input nonlinearity model. As detailed below, estimation of all kernels was done simultaneously rather than independently.

The full context model is given by

$$\hat{\mathbf{r}} = (\mathbf{w^t} \otimes \mathbf{w^f} \otimes \mathbf{w^l} \otimes \mathbf{w}^{\tau} \otimes \mathbf{w}^{\phi} \otimes \mathbf{w}^{\lambda}) \bullet \mathbf{Q}^{\mathbf{tfl}\tau\phi\lambda},$$

where the lag kernel $\mathbf{w}^{\tau}$, the frequency kernel $\mathbf{w}^{\phi}$, and the amplitude kernel $\mathbf{w}^{\lambda}$ form the CRF. The stimulus representation $\mathbf{Q}^{\mathbf{tfl}\tau\phi\lambda}$ is computed to contain the corresponding multiplicative interactions between different points in the time-frequency-level representation of the stimulus on which the CRF operates.

Note, that both the input nonlinearity and the context model can model a constant offset (background) rate by extending the stimulus representation appropriately (detailed in Ahrens et al. 2008a). Also, for comparison with other models certain kernels can be left out, e.g. the input nonlinearity for comparing the input nonlinearity model with the STRF. To distinguish the first-order spectro-temporal kernel in the nonlinearity/context model from the classical STRF, we will refer to it as the principal receptive field (PRF).

Grouping of stimulus dimensions. While a fully dimension-separated representation of the multilinear kernels is preferrable for parsimony in the number of parameters, interactions between certain stimulus dimensions can lead to inseparabilities (e.g. Depireux et al., 2001). These can be accounted for in the multilinear framework by grouping certain stimulus dimensions. For example, if time and frequency are assumed to be inseparable, the input nonlinearity model can be modified to $\hat{\mathbf{r}} = (\mathbf{w^{tf}} \otimes \mathbf{w^l}) \bullet \mathbf{Q^{itfl}}$, where $\mathbf{w^{tf}}$ denotes the time-frequency kernel. This grouping can be applied to any set of kernels in the aforementioned models. For notational convenience, the models will be abbreviated by their grouping structure, e.g. $\mathbf{tf}\otimes\mathbf{l}$.

Model estimation. Errors and parameter priors were assumed to follow a Gaussian distribution. Under this assumption, maximizing the likelihood of the data given the model becomes equivalent to minimizing the squared error

$$\varepsilon = \|\mathbf{r} - \mathbf{W} \bullet \mathbf{Q}\|^2 = \sum_{i=1}^{I} (r(i) - \mathbf{W} \bullet \mathbf{Q^i})^2$$

a problem which is solved by linear regression.

In the multilinear setting the full kernel $\mathbf{W}$ is given by the outer product of the individual kernels, $\mathbf{W} = \mathbf{w^1} \otimes \cdots \otimes \mathbf{w^n}$. Estimating $\mathbf{W}$ directly by linear regression would lead to an oversized representation and require a post-hoc step to estimate the constituent kernels $\mathbf{w^k}$. Ahrens et al. (2008a) proposed an iterative algorithm based on the alternating least squares method to estimate the $\mathbf{w^k}$ directly. Starting from the expanded error expression

$$\varepsilon = \|\mathbf{r} - (\mathbf{w^1} \otimes \cdots \otimes \mathbf{w^n}) \bullet \mathbf{Q}\|^2$$

the following set of equations

$$\begin{array}{ll} \mathbf{Q^{(1)}} = (\mathbf{w^2} \otimes \cdots \otimes \mathbf{w^n}) \bullet \mathbf{Q} & \mathbf{w^1} = ((\mathbf{Q^{(1)}})^T \mathbf{Q^{(1)}})^{-1} (\mathbf{Q^{(1)}})^T \mathbf{r} \\ \vdots & \vdots \\ \mathbf{Q^{(n)}} = (\mathbf{w^1} \otimes \cdots \otimes \mathbf{w^{n-1}}) \bullet \mathbf{Q} & \mathbf{w^n} = ((\mathbf{Q^{(n)}})^T \mathbf{Q^{(n)}})^{-1} (\mathbf{Q^{(n)}})^T \mathbf{r} \end{array}$$

can be derived by differentiating with respect to each $\mathbf{w^k}$. Each equation minimizes the squared error with respect to the corresponding dimension k, providing the updated kernel $\mathbf{w^k}$. Since each step reduces the lower-bounded error ε, the iteration has to converge eventually. Estimation was terminated if $\frac{\|\mathbf{w}^{\mathbf{k}}_{\text{new}} - \mathbf{w}^{\mathbf{k}}_{\text{old}}\|}{\|\mathbf{w}^{\mathbf{k}}_{\text{new}}\|} < 0.005$ for all kernels $\mathbf{w^k}$, which was usually reached after a few iterations. For the input nonlinearity model this criterion was already achieved after 3–4 iterations, while context models converged after ≈20 iterations.

Regularization. Some of the estimated linear models contain hundreds of parameters. Spurious correlations between stimulus and response are often reflected in noisy parameter estimates. This overfitting can be limited by supplying prior information on the parameters. Matching this prior to aspects of the observable data reduces the arbitrariness of this general strategy. Sahani and Linden (2003a) developed a method to adapt the covariance structure of the parameters to the evidence given by the data. This method effectively controls the smoothness between parameters and has therefore been termed Automatic Smoothness Determination (ASD, see Sahani and Linden 2003a and Ahrens et al. 2008a for details). All estimates were regularized using ASD for the first three iterations of the alternating least squares procedure after which the prior parameters were kept fixed to guarantee convergence of the kernel estimation.

Degeneracies. For the multilinear models, certain parameter choices can lead to the same global mapping. While the predictive power is not influenced by these choices, the certainty with which a parameter can be constrained often decreases severly. We implemented the counter-measures detailed in Ahrens et al. (2008a) for both the 'scaling' degeneracy (level related) and the 'additive' degeneracy (context related).

Output nonlinearity

A static output nonlinearity can improve the predictive power by (i) confining the predicted firing rate within its natural bounds (≈0–1300 Hz, based on a typical refractory period of 0.7 ms in the MNTB) and (ii) approximating certain static nonlinear transformations which cannot be captured by the linear model (see Ahrens et al. (2008b) for a dramatic example of this kind).

Since the ability for generalization is typically inversely related to the number of free parameters, we attempted to find a functional prototype with only a small number of parameters which captured the range of observed output nonlinearities. An extended sigmoidal function served this purpose well, given by

$$s_2(p) := r_0 + \frac{r_{\max}}{1 + e^{-k_1(p-p_1)} + e^{-k_2(p-p_2)}},$$

depending on 6 parameters: a minimal rate r_0, a maximal rate increase $r_{\max}$, and two pairs of shift ($p_{1,2}$) and slope parameters ($k_{1,2}$). An example is shown in Fig. 6(A). The realizable functional forms of $s_2(p)$ account for the different curvatures of the nonlinearity close to its lower and upper saturating values. More complex functions did not improve crossvalidation performance beyond the level attained with s_2.

Estimation of the nonlinearity can either be performed jointly with the linear model or sequentially. A previous comparison for cortical neurons (Ahrens et al., 2008b) suggested only slight improvements of joint over sequential estimation. Therefore we considered it acceptable to estimate the output-nonlinearity post-hoc using nonlinear least-squares minimization (Matlab function: *lsqnonlin*), i.e. between the multilinear prediction and the PSTH.

Model evaluation and analysis

Model evaluation. The performance of the different models was compared based on their predictive power β (Sahani and Linden, 2003b; Machens et al., 2004), defined as the difference between total and error power normalized by the estimated signal power, i.e.

$$\beta(\{r_n\}_{n=1,\ldots,N}) := \frac{P(\overline{r_n}) - P(\overline{r_n} - \hat{\boldsymbol{\mu}})}{\hat{P}(\boldsymbol{\mu})} \tag{1}$$

where $\overline{\cdot}$ denotes trial average, $\boldsymbol{\mu}$ is the true, but unobserved firing rate, and the power of a signal s is defined as

$$P(s) := \langle (s - s)^2 \rangle = \frac{1}{T}\sum_{i=1}^{T}\left(s(i) - \frac{1}{T}\sum_{j=1}^{T} s(j)\right)^2.$$

The estimator for the signal power $\hat{P}(\boldsymbol{\mu}) := \frac{1}{N-1}\left(NP(\overline{r_n}) - \overline{P(r_n)}\right)$ is unbiased and only assumes additivity and independence between signal and noise (Sahani and Linden, 2003b). The predictive power is a measure of the explained variance with respect to the signal, i.e. explainable variance. In comparison, the fraction of explained variance is defined as $f_{\mathrm{VE}} := \frac{P(\overline{r_n}) - P(\overline{r_n} - \hat{\boldsymbol{\mu}})}{P(\overline{r_n})}$ and is known to underestimate the predictive quality in the presence of noise since the denominator overestimates the explainable variance.

Applying model estimation and prediction to the same dataset (insample estimate) overestimates the true predictive power, since the model also learns some of the noise in the dataset. Conversely, estimating the model on one part of the dataset and predicting another part (crossvalidation estimate, 80% training set, 20% test set, 5 divisions) underestimates the true predictive power, since the learned noise worsens the prediction on the unseen data. For linear models, these estimates diverge linearly with the power of the noise, with the slope determined by the signal power and the model complexity (B.E., unpublished results). Given a model class, its true predictive power is therefore to be found between the insample and the crossvalidation estimate.

When evaluating model performance one faces two tasks: attributing a performance to individual cells and attributing a performance to the considered population of cells. Consider first the latter task: An approach taken by Sahani and Linden (2003b) and Ahrens et al. (2008a) is to depict β as a function of noise power and extrapolate (linearly or nonlinearly) to zero noise power. This extrapolation is performed separately for crossvalidation and insample predictions yielding lower and upper bounds of β respectively (examples of this extrapolation are shown in Fig. 5(A–D)). These bounds would then be assumed to quantify the level to which the given model can account for the investigated stimulus-response mapping. This, however, assumes that all instances, i.e. cells, are similarly well described by the applied models. If this assumption is severely violated, these extrapolations can give erroneous results, since the population properties dominate the noise dependence. Fig. 5(A–D) exemplifies this problem since the predictive power of the crossvalidation fits (gray) *increase* with noise power. At the same time the differences between the insample and the crossvalidation estimation for individual cells are mostly small, especially compared to the range of predictive powers in the population. Quantitatively, the average difference between insample (IS) and crossvalidation (CV) predictive powers was only 7% of the average difference of predictive powers between cells for any of the models estimated (measured as the S.D. of the differences for the input nonlinearity model, 0.8% vs. 12.7%). As theoretically predicted, the differences do increase as a function of noise power ($r = 0.7$, data not shown).

Therefore the population predictive power had to be estimated differently. First, the single cell predictive power was estimated by taking the average between the insample and crossvalidation predictive power. Since ASD-regularized estimates were obtained in both cases, the divergence rates for increasing noise should be similar (see Appendix A for a semianalytic proof of this claim). If this holds, the average should be a reasonable estimate of the true predictive power. A model's predictive power on the entire dataset was then defined as the median of these averages since their distribution across the population was skewed towards higher predictive powers. Using only regularized estimates might underestimate the model's performance for a given dataset. Regularized estimates avoid counterintuitive values of β ($\beta > 1, \beta < 0$) and provide smoother kernel estimates for the insample estimates. Note, that Sahani and Linden (2003b) used unregularized insample estimates to gauge the full potential of STRFs for explaining the data.

In cell-by-cell comparisons for two models or conditions, the average between the insample predictive power and the crossvalidation predictive power is used. In histograms the average insample predictive power is compared to the average crossvalidation predictive power, thus providing a measure of the margin between them (Predictive powers for all cells and tested models) are provided on the web (http://www.mis.mpg.de/jjost/neuro/englitz/).

Model analysis. Assigning significance to certain regions in the estimated kernels requires estimating their distribution. Under the assumption of a normal distribution, estimating the standard deviation suffices. Using bootstrap resamples (Efron and Tibshirani, 1993) these standard deviations were estimated pointwise for each model and cell (10 resamples). Deviations from zero were considered significant if they exceeded 3 standard deviations, corresponding approx. to a $p<0.01$ criterion. Kernels were oversampled by a factor or four before analysis (using Matlab function: *interp1* and *interp2*). Contours were extracted (Matlab function: *contourc*) for each region of significant deviation and quantified in several dimensions, including spectral and temporal position and bandwidth, size, and weight. All spectrotemporal positions were measured at their peak. For suppressive regions the spectral and temporal location was further measured relative to the main excitatory region.

Results

We analyzed extracellular recordings from 96 MNTB principal cells of the Mongolian gerbil. For each cell we collected 20 or more trials in response to a broadband, noise-like stimulus (Fig. 2(A1)). For comparability the same number of trials (20) was used to construct the PSTH which went into estimating the models. The first 50 ms of the response were excluded from the fitting procedure to avoid onset effects. The first trial was not excluded, since no difference in predictive power or receptive field structure was apparent.

An example of a neuronal response is shown in Fig. 2(A2). The spiking pattern varies rapidly in response to the temporal modulations in the stimulus (Fig. 2(A3), (A4)), consistent with the frequency range of rate comodulation of MNTB neurons (Joris and Yin, 1998; Tolnai et al., 2008). Many cells exhibited episodes of highly

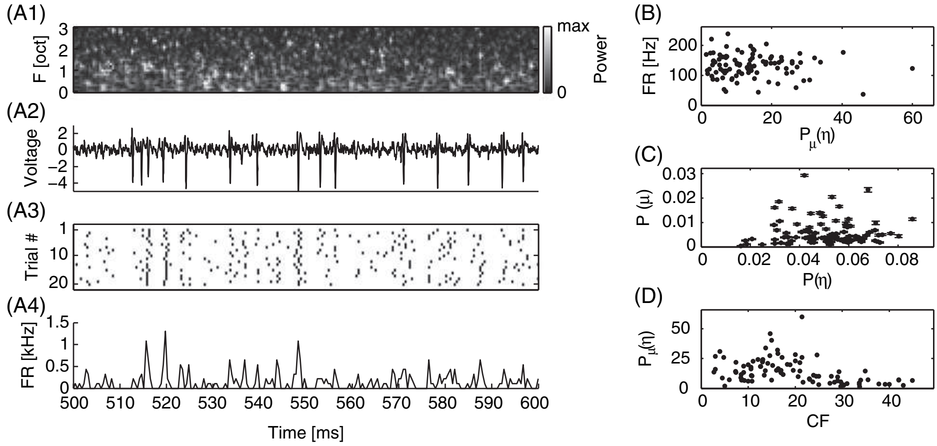

Figure 2. Single cell responses vary in temporal precision over time and across trials. The spectrotemporal stimulus (A1) elicits a neuronal response (A2). At least 20 trials (A3) were collected for each cell, yielding the PSTH (A4). The reliability of the response is strongly modulated over time with intermediate phases of high temporal precision. (B) The firing rate varied across cells, but was uncorrelated to the normalized noise power. (C) The signal powers of all cells were significantly greater than 0 (error bars do not include 0) and exhibited only weak correlation with noise power. (D) The normalized noise power decreases with CF corresponding to an increase in signal to noise ratio with CF.

reliable, precisely timed (0.5–1 ms) responses in firing rate, containing a spike in the corresponding bin of most trials.

Individual neurons discharged at a range of average firing rates, depending on their tuning and internal properties. Not all cells could be driven up to the same firing rate. Interestingly, firing rate was not predictive of a cell's overall reliability of discharge. This relationship is shown in Fig. 2(B), where the noise power $P(\eta)$ normalized by the power of the systematic response $P(\mu)$, denoted as $P_{\mu}(\eta)$, serves as an inverse measure of reliability, i.e. responses are perfectly reliable at zero noise power and become more unreliable for higher noise powers. In general, the reliability of spike timing varied strongly across cells and as a function of peri-stimulus time within a given cell.

While $P(\eta)$ always exceeded $P(\mu)$ (Fig. 2(C)), the estimated variances of $P(\mu)$ indicate that all cells modulated their firing rate significantly in response to the stimulus. The criterion for significant modulation was that zero was not within 2 S.D. of the estimated $P(\mu)$. Hence, no cells needed to be excluded based on this criterion.

Interestingly, $P_{\mu}(\eta)$ decreased significantly as a function of characteristic frequency (CF, $p < 10^{-6}$, Pearson correlation coefficient, Fig. 2(D)), corresponding to a relative increase in signal to noise ratio as a function of CF. This dependence was based on an increase of $P(\mu)$ as a function of CF ($p < 10^{-6}$), while $P(\eta)$ was uncorrelated to CF ($p = 0.3$).

For the following analysis we considered the firing rate as the stimulus dependent, deterministic variable, and the spike timing variations as stimulus-unrelated noise.

Parameter and model selection

We systematically optimized the sampling rate and the stimulus representation, as described in the following two sections. The subsequent three sections are dedicated to our choices of model structure.

Submillisecond sampling rate provides best performance

The sampling rate of the model was chosen to optimize the predictive performance. Sampling rates between 1 and 5 kHz were tested. The choice of the lower bound was guided by knowledge about the modulation properties of MNTB cells (up to 1 kHz, Tolnai et al. (2008)). The upper bound was restricted by the memory resources required to estimate the available models, especially the full context model (with a 7-dim. $\mathbf{Q}$). In this section and the next, the simple STRF model with the linearly scaled Prescription stimulus and unregularized regression was used since this setting was least time consuming. Results with other models and estimation methods were similar.

The choice of sampling rate influences the coarseness of the stimulus (Fig. 3(A)) and the response representation (Fig. 3(B)). At high sampling rates, the available number of trials can be too low to provide useful firing rate estimates, and further the models are prone to overparametrization. At low sampling rates estimates become cleaner but might miss details in both stimulus and response. For the considered range, both the fraction of explained variance f_{VE} and the predictive power β assumed a local maximum at 1.5 kHz and 2.2 kHz, respectively (Fig. 3(C)).

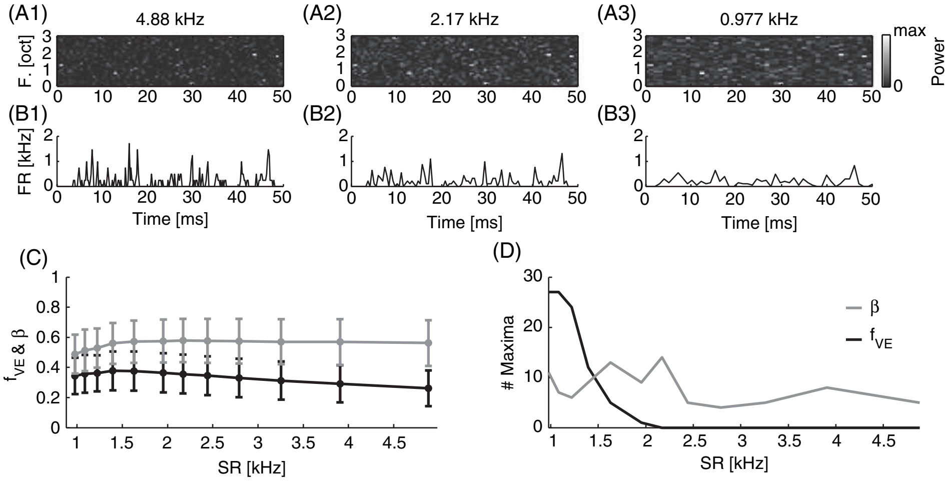

Figure 3. The sampling rate influences the model performance. Stimuli (A1–A3) and PSTHs (B1–B3) for three of the sampling rates (SRs) ranging from 1 to 4.9 kHz are depicted. The fraction of explained variance f_{VE} (black) and the predictive power β (gray) depend on SR (C). The average of f_{VE} peaks around 1.5 kHz, whereas β reaches a plateau above ≈1.6 kHz with a shallow peak at 2.17 kHz. This peak coincides with the maximum number of β-optimal SRs in the population (D). The model was an STRF with output nonlinearity based on the Prescription stimulus.

Since β assumes a plateau starting at ≈1.6 kHz, we additionally collected a histogram of sampling rates providing best β and f_{VE} for individual cells (Fig. 3(D)).

As argued before, β provides a better estimate of the predictive quality than f_{VE} we chose its optimal firing rate (2.2 kHz) as the standard sampling rate for all further models. Note that this optimum could be dependent on the available stimulus set and might be different for other sets.

Cochlea-like stimulus representation provides best performance

The stimulus representation corresponds to a linear or nonlinear preprocessing of the stimulus and can reshape the stimulus to be better adapted to the subsequent model stages. We compared three spectrotemporal stimulus representations, the description used in designing the stimulus (‘Prescription’, Fig. 4(A)), a classical short-term Fourier transform based spectrogram (Fig. 4(B)) and a γ-tone based, cochlea-like representation (Fig. 4(C)). The choice of representation and its scaling, either logarithmic (corresponding to dBs, Fig. 4(A1–C1)) or linear (corresponding to sound pressure) scaling significantly influenced the predictive power of the model.

The linearly scaled γ-tone representation (Fig. 4C2) led to the best performance among the tested alternatives. Compared to the Prescription representation the difference was 8.6% on average with improved performance essentially across the entire population (Fig. 4(D)). The STFT representation led to an average improvement of 6.3% with respect to the Prescription representation, thus 2.3% less than the γ-tone representation (Fig. 4(E)). For the γ-tone representation the linear scaling compared favorably with the logarithmic scaling at a quantitative advantage of 1.8% (Fig. 4(F)). In these comparisons the improvements were highly

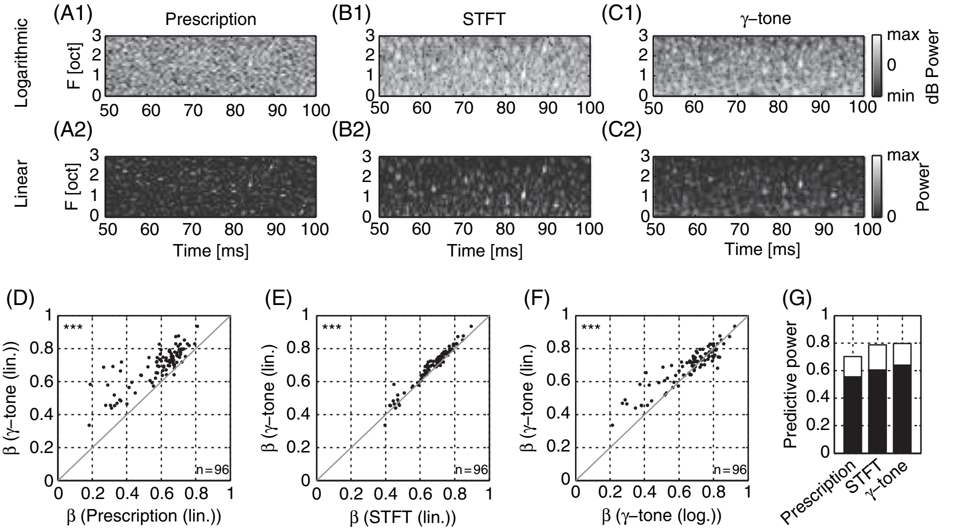

Figure 4. Comparison of different stimulus representations. Examples of each representation are shown in (A–C) in linear (A1–C1) and logarithmic scaling (A2–C2). The predictive power from STRF models was highest for the linear γ-tone representation. In comparison with the linear Prescription representation the increase was 8.6% in (D), compared to only 2.3% in comparison to the STFT in (E). (F) Linear scaling outperformed logarithmic scaling by 1.8% on average. (G) Bar plots for the crossvalidation (black) and insample (white) performance indicate that the γ-tone representation also provides more consistent performance than the STFT.

significant ($p < 0.001$). Significance was assessed with the Wilcoxon signed rank test for a nonzero median on the differences between the two conditions. Stars in the upper left corner of this and the following figures indicate significance, with ⋆, ⋆⋆, and ⋆⋆⋆, corresponding to $p < 0.05$, $p < 0.01$, and $p < 0.001$, respectively.

The histograms in Fig. 4(G) show both the average crossvalidation (black bar) and the insample (white bar) performance for linear scaling. Since both performances lie above the respective performances of the other representations, the γ-tone representation can be considered the better choice. In all further models this combination of representation and scaling was used.

This choice of representation is consistent with the result obtained by Gill et al. (2006), where a cochlear model (Lyon, 1982) provided the best performance. Concerning the scaling the results differ as dB-scaled spectrograms led to better predictive power in the aforementioned studies. A direct comparison is nontrivial since Gill et al. (2006) recorded from neurons of the avian mid- and forebrain.

Grouping dimensions improves performance

After fixing the sampling rate and stimulus representation, we step through the family of multilinear models. First, we consider variants of the input nonlinearity model based on the three stimulus dimensions time, frequency, and sound level, by grouping pairs of dimensions. While this procedure increases the number of model parameters, it allows modeling inseparabilities between the dimensions.

The population differences in predictive power usually exceeded the differences between the values for insample and crossvalidation for individual cells (Fig. 5(A–D)). In the input nonlinearity model with separated dimensions the latter differences remain small over the whole range of noise powers. Correspondingly, the extrapolations for insample and crossvalidation predictive powers stay close and both increase as a function of noise power (see Methods). The time-frequency grouped model displays substantially larger noise dependent deviation between the two estimates than the other two model variants (compare Fig. 5(B) with Fig. 5(C&D)). This seems surprising given the number of free parameters of each variant (**tf** $\otimes$ **i**: 522, **fi** $\otimes$ **t**: 822, **ti** $\otimes$ **f**: 447). However, since the ASD regularization for intensity usually extended over many more bins than either of the other two dimensions, the effective number of free parameters is lower in the latter two variants, where intensity is part of the two dimensional kernel.

Considering the best performance over all grouped models brings an average increase in predictive power of 3.3% over the fully separated model. This increase is fairly homogeneous over the entire population (Fig. 5(E)). Separated by model variant (Fig. 5(F)), the time-frequency grouped model performs better than the other models in 58% of the cells, followed by the time-intensity grouped model at 30%, and the frequency-intensity grouped model at 11%. The fully separated model never performed best, even for crossvalidation estimates. Histograms for insample

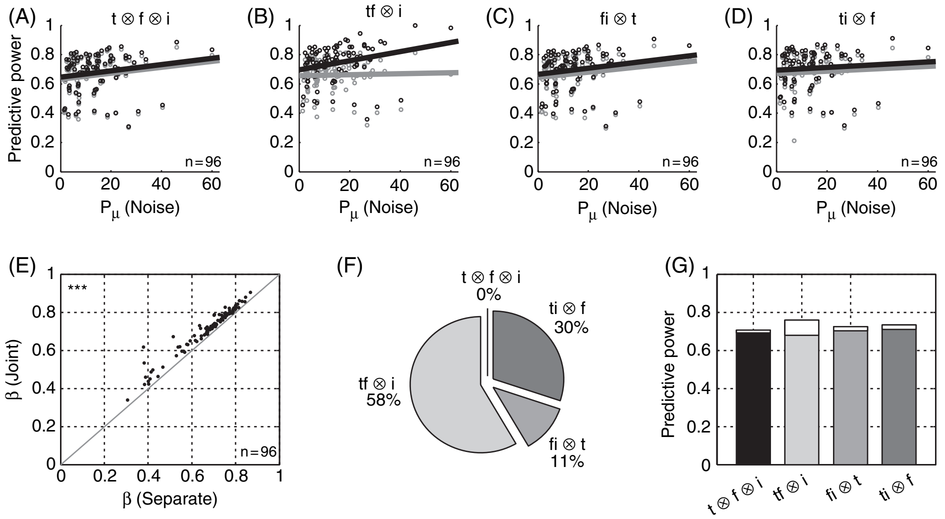

Figure 5. Insample (black) and crossvalidation (gray) results of the individual models vary less than population results. (A–D) shows the results for the fully separated (A) and two dimensions grouped input nonlinearity models (B–D). The insample and crossvalidation regressions reflect the population trend rather than the expected noise-performance relation. (E) Fusing dimensions improves model performance with the best of the three grouped models performing 3.3% (on average) better than the separated model. (F) In 58% of the cells fusing time and frequency performed better than fusing time and intensity (30%) or frequency and intensity (11%). (G) Despite the dominance of the time-frequency model for individual cells, the averages are close (insample: white, crossvalidated: gray levels). The time-frequency model exhibits the highest population variance. Interestingly, the time-intensity grouped model has the best crossvalidation performance.

(white) and crossvalidation power (colored, Fig. 5(G)) reflect differences in noise dependence observed in Fig. 5(A–D), but also show that the average differences between the three grouped models are only minor.

Effects on performance of input nonlinearity and output nonlinearity

Estimating a static, nonlinear rescaling of either stimulus level or neuronal response (or conversely the prediction) improves the representational capabilities of the multilinear models. While the nonlinearity after the multilinear model (output nonlinearity, ON) is estimated between the multilinear prediction and the data, the nonlinearity on stimulus level is directly estimated within the multilinear framework (Ahrens et al., 2008a). Hence, both the difference in location and in estimation could lead to differences in predictive performance.

As output nonlinearity we estimated the double exponential sigmoid s_2 with 6 free parameters (see Fig. 6(A) and Methods). In the depicted example the nonlinearity maps predictions below the diagonal closer to the diagonal, thus improving the predictions. On the population level, adding an output nonlinearity (Fig. 6(D)) or adding an input nonlinearity (Fig. 6(E)) to an STRF model improves performance (probably coincidentally) by a similar percentage (4.8% and 4.1% respectively), in both cases consistently across the population. Different cells benefit relatively more from one or the other nonlinearity, as illustrated by Fig. 6(F). The cell dependence suggests combining the two nonlinearities: this provides a substantial (5.4%) and consistent increase over the STRF model (Fig. 6(G)), thus emphasizing the need for nonlinear scaling at different stages of the models.

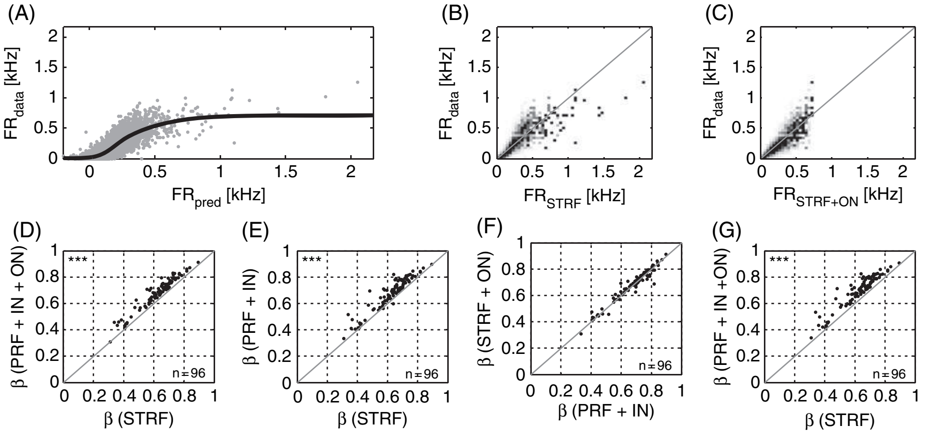

Figure 6. Input and output nonlinearity account for similar effects. (A) In the case of a STRF model the relationship between the linear prediction and the observed firing rates appears to be a sigmoidal nonlinearity, modeled here by a biexponential sigmoid (see Methods). (B) and (C) show the same data as A as density plots (normalized vertically), where the effect of the nonlinear rescaling is shown in C, moving the density closer to the diagonal. For the population the predictive power of STRF models increased by 4.8% with an additional output nonlinearity (D) and by 4.1% with the input nonlinearity (E). The performance of the STRF models with a nonlinearity either at the input or at the output were nearly identical (**f**, 0.6%). Combining the two nonlinearities provides the strongest increase (**g**, 5.4%).

CRF systematically improves performance

The context model adds a second level of kernels of the same dimensions (time, frequency and level), as the input nonlinearity model and combines them multiplicatively with the first. We here compare it with the time-frequency grouped input nonlinearity model. Predictions and model kernels of both models are shown for four cells in Fig. 7 (left column).

The context model without nonlinear scaling improves the performance by 2.8% (Fig. 8(A)); including the nonlinear stimulus scaling improves it by another 0.4% (Fig. 8(B)). Importantly, this increase is also present when considering only the crossvalidation results (Fig. 8(C)). The improvements due to the added context suggest that multiplicative interactions are relevant when modeling the present neuronal responses.

In summary, while the individual improvements in the previous sections were small, appropriate choices of sampling rate, stimulus representation, and several aspects of model structure improved the performance in total by ≈20%. Starting from about 55% (STRF, Prescription, low sampling rate, no output nonlinearity) to

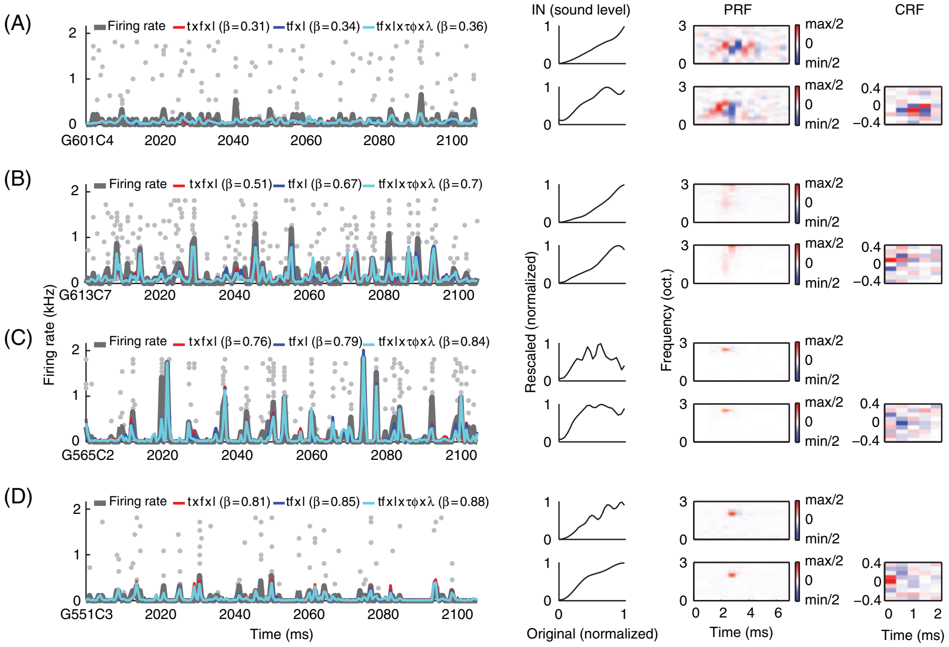

Figure 7. Predictions and kernels for individual cells and models. The left column depicts the PSTH (grey) and the respective predictions of the separated (red, $\mathbf{t} \otimes \mathbf{f} \otimes \mathbf{l}$), the time-frequency grouped (blue, $\mathbf{tf} \otimes \mathbf{i}$) and the full context model (cyan, $\mathbf{tf} \otimes \mathbf{l} \otimes \boldsymbol{\tau\phi} \otimes \boldsymbol{\lambda}$). The right hand side depicts the kernels of the time-frequency grouped model on top and the kernels of the context model on the bottom (respectively for each cell). (A) While the depicted cell was least predictable, it is an interesting example for a transfer of structures from the time-frequency kernel to the context field, keeping its tilt and relative location to the main excitatory field. (B) Substantial improvement can be achieved in a number of cells by grouping the time-frequency dimensions, here amounting to ?30% and almost 40% when the context field was included. The cases in (C) & (D) are characteristic for the majority of cells in the sample, where grouping of dimensions and adding the CRF lead to a consistent, yet, smaller increase in predictive power β.

more than 75% (context model, γ-tone, intermediate sampling rate, input or output nonlinearity), this increase corresponded to a relative increase of more than 35% in predictive power. Relative to a similarly optimized STRF model (STFT, intermediate sampling rate, output nonlinearity) the improvement by the multilinear model was 8%.

Physiological interpretation of model properties

The model parameters allow a natural interpretation in terms of classical temporal and spectral quantities, i.e. as (relative) latencies, preferred frequency as well as temporal and spectral bandwidth. We measure and attribute these quantities to parameters which are significantly different from 0, assessed by the S.D. obtained from bootstrap resampling. Neighboring parameters of same polarity are treated as a field or region.

Time-frequency kernel. An example of a PRF (from an input nonlinearity model) with an overlay of the significance boundaries of each field is depicted in Fig. 9(A). PRFs usually had only one excitatory (red) and a number of inhibitory (blue) regions. For further analysis, we subdivided the inhibitory regions into 4 classes: (i) LF = below CF, (ii) HF = above CF, (iii) CF(L) = at CF with lower latencies than the excitation and (iv) CF(H) = at CF with higher latency than excitation. While the LF and HF regions probably indicate cochlear processing or neuronal inputs, the CF regions could also correspond to internal dynamics (e.g. adaptation) or network dynamics.

The excitatory and inhibitory fields of the present neuronal sample cover essentially (up to the high frequency limit determined by speaker calibration range) the entire range of auditory frequencies audible by the gerbil (Fig. 9(B)). The excitatory fields' latencies are not strongly dependent on CF consistent with previous findings in the auditory nerve for CFs above 2 kHz (Recio-Spinoso et al., 2005). Inhibitory fields broaden temporally with CF and extend up to 6–8 ms.

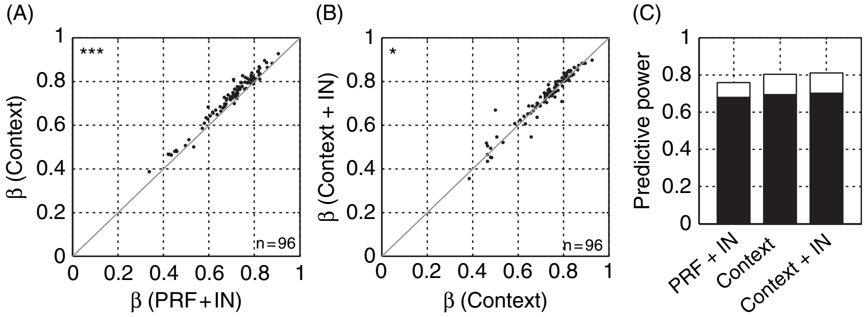

Figure 8. Including the context part systematically improves predictive performance. (A) The context model without nonlinear input scaling performs 2.8% better than the input nonlinearity model, both with grouped time-frequency kernel. (B) The context model with input nonlinearity performs only slightly better (0.4%). (C). While the variances of the context models exceed that of the input nonlinearity, they still provide better crossvalidation performance (black).

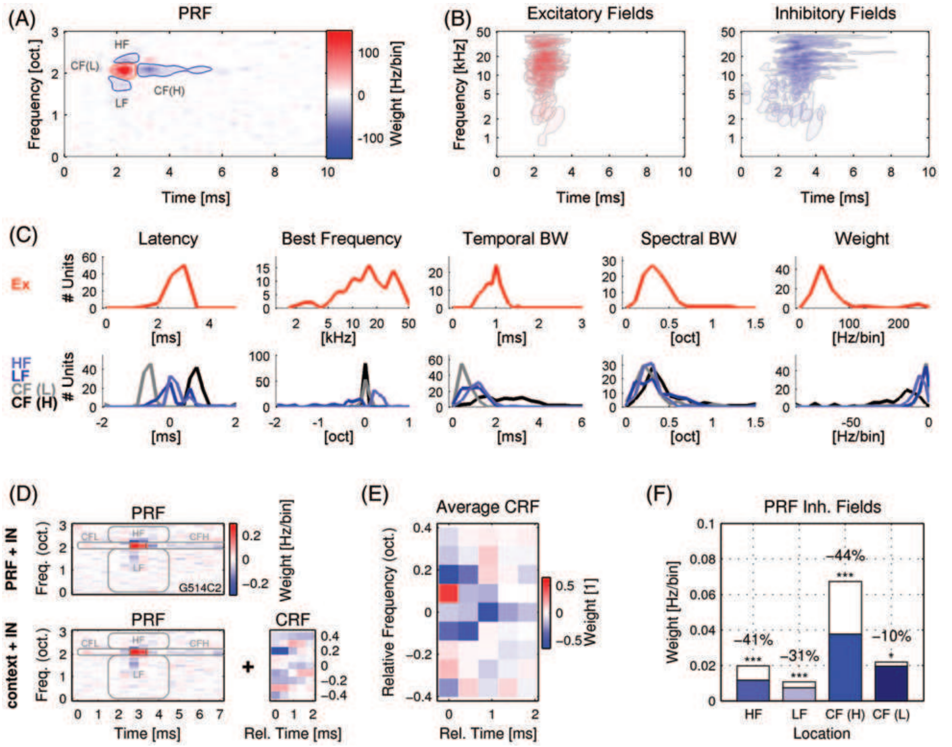

Figure 9. Spectrotemporal properties of the principal receptive field (PRF) for the input nonlinearity model (A–C) and comparison with the context model (D–F). (A) Example of an oversampled PRF with contours delimiting regions of significant deviation. In all graphs red lines indicate excitatory and blue lines inhibitory regions. Inhibitory regions were subdivided into 4 classes: (i) LF = below CF, (ii) HF = above CF, (iii) CF(L) = at CF with lower latency and (iv) CF(H) = at CF with higher latency. (B) Shape and absolute spectrotemporal position of main excitatory (left) and inhibitory (right) fields for the entire population. (C) Distribution of a range of spectral and temporal properties of excitatory (top) and inhibitory (bottom) regions. Inhibitory latencies and spectral positions are relative to the main excitation's location (see text for details). (D) The reduction in the high frequency sideband (relative to main excitatory field) coincides with the spectral position of the suppressive field in the context model. (E) The average context receptive field (CRF) exhibits a neighborhood of inhibition around the reference frequency and decays for further distances. (F) While all inhibitory sidebands in the PRF are reduced by adding a CRF to the model (different blues) when compared to the PRF in the input nonlinearity model (white bars), the high frequency (HF, −41%) and post CF (CF (H), −44%) fields show the strongest reduction, followed by the low frequency field (LF, −31%). Positive values were set to 0.

Some of the CF inhibitory fields prior to the excitatory field seem to occur at unphysiologically low latencies. They probably correspond to temporal correlations in the response which reappear as temporal correlations in $\mathbf{w}^{\mathbf{tf}}$, e.g. due to the refractory period in between spike sequences.

Histograms of spectral and temporal location and width as well as total weight are shown in Fig. 9(C). The excitatory latencies are quite constrained around an average of 2.6 ms (The average latency is slightly smaller ($\approx$0.3 ms) than usual estimates, probably due to phase delays generated by the γ-tone filtering.). The CF distribution is skewed towards higher frequencies, corresponding to the CF distribution in the MNTB (Tolnai et al., 2008). Inhibitory regions usually stay close to the excitatory region both in the spectral and the temporal direction (latency and best frequency subplots). While the temporal bandwidth (TBW, middle subplots) of the excitatory regions is tightly distributed between 0.5 and 1.3 ms, the TBW of inhibitory regions depends on their location. LF and HF fields have similar TBW distributions ranging between 0.5 and 2 ms, exceeding the very brief CF(L) fields (0.5–1.2 ms), but staying far below the broad distribution for CF(H) of 1–4 ms. The distributions of spectral bandwidths are quite similar across all fields, excitatory and inhibitory. Finally, the weights of the inhibitory fields indicate that the CF(H) field usually has the greatest influence on the neuronal response, followed by the HF field. The LF and CF(L) fields are usually least influential.

Context model. Aside from the increased predictive quality provided by the full context model, it also provides insight into the structures underlying auditory processing. Here we address whether adding the context reduces inhibitory fields in the PRF. The depicted context model included nonlinear scaling on the input (kernels not shown).

In the example shown in Fig. 9(D), the HF inhibitory region is strongly reduced (left: compare PRFs on top and bottom above the CF) while a damping field is present in the CRF (bottom right). The frequency separation from the center in the context corresponds to the relative distance between the HF field and CF in the PRF. The average CRF for the present population had damping fields surrounding the center frequency and very little activation outside a range of 1 ms and 0.2 octaves (Fig. 9(E)). Correspondingly, the strongest reductions are observed for the HF (−41%) and CF(H) fields (−44%, Fig. 9(F)), followed by the LF field (−31%). These reductions were highly significant (Wilcoxon Rank Sum test, $p < 10^{-9}$), while the reduction of the CF(L) field by 10% was only barely significant ($p = 0.041$). This mild reduction could be a consequence of using a CRF which only extends towards the past.

Further examples of changes in the PRF when adding the CRF are shown for four individual cells in Fig. 7 (right side, see caption for details).

Discussion

We investigated the stimulus representation in the MNTB on a quantitative level using an array of phenomenological, multilinear models. Best overall performance was achieved by the context model with nonlinear input scaling based on a cochlea-like stimulus representation and at sampling rates around 2 kHz. Further, static nonlinear scaling of either stimulus level or predicted rate led to significant improvements in performance, bringing the overall model to capture 75% of the explainable variance.

Comparison with other modeling approaches

A quantitative comparison with other modeling approaches is difficult, since responses of the MNTB have to our knowledge not been modeled before.

However, MNTB responses share many properties with cells from the cochlear nucleus (especially globular bushy cells of the AVCN) and auditory nerve fibers (ANF). We therefore provide a broader perspective on some of the recent modeling approaches.

What is the purpose of models for neuronal responses? An accurate model can be used to simulate the time-varying response of a given brain area to other complex stimuli, which can be useful for predicting the effect of this brain area in overall processing. At the same time, a model can provide insight into structural aspects of the investigated system. Different models will allow different degrees of insight. Predictive power and structural understanding are the dimensions around which the following comparison is developed.

Several types of phenomenological models have been used for modeling auditory responses. The earliest models are probably the *revcor* models of ANFs by De Boer and De Jongh (1978), closely followed by the introduction of the STRF by Aertsen and colleagues (Aertsen et al., 1981a, 1980, 1981b). Generally, these models differ in the stimulus representation: *revcor* models are based on the sound pressure waveform and particularly effective in modeling low-frequency units which phase-lock to the fine structure of the sound pressure wave. However, for high-frequency units *revcors* are typically flat. In recent studies Recio-Spinoso et al. (2005), Temchin et al. (2005), and Lewis et al. (2002) have extended *revcor* models to second order Wiener series and demonstrated their ability to provide good predictions for white noise stimuli Temchin et al. (2005). Quantitatively these predictions are of similar quality as the present predictions. Temchin et al. (2005) measured model performance by the fraction of explained variance. The difference to the predictive power should, however, be small since the average was obtained from a large number of repetitions. Structurally, sound pressure based models present both insights and hurdles. The *revcor* can be attributed to the impulse response of the basilar membrane for low CFs. The two-point temporal correlations in the second order Wiener kernel are harder to interpret. Usually, further transformations are required to extract characteristic properties. For example the CF can be extracted from the first singular value of the second order kernel and provides an estimator over the whole frequency range (Recio-Spinoso et al., 2005; Temchin et al., 2005).

STRFs are based on a stimulus spectrogram. They have been used in studies of ANFs (Kim and Young, 1994) and different cell types of the cochlear nucleus (Backoff and Clopton, 1991; Clopton and Backoff, 1991). Predictive performance was not addressed in these studies. The present results indicate that STRFs and especially their multilinear extensions provide reliable quantitative predictions for principal cells in the MNTB. Based on the assumption of increasing complexity in stimulus transformation, one might speculate these models would exhibit equal or better performance for neurons in the AVCN or ANFs. The similarity in performance with the second order Wiener models discussed above is partly implied by the correspondence between the STRF and the Fourier transform of the second order Wiener kernel (Klein et al., 2000). Due to the phaseless representation of the stimulus, the performance of spectrotemporal models is usually reduced for low CFs (for the few low CF units in the present study such a trend was present; data not shown). Structurally, spectrogram based models are easier to interpret, yet harder to link to underlying structures. STRFs usually exhibit a dominant excitatory field with neighboring suppressive fields. Their interpretation is guided by the easily extractable spectrotemporal properties. While these allow the selection of candidate mechanisms, their link to actual biological structures is weaker than for the *revcor*.

Yu and Young (2000) introduced another variant of a second-order Wiener model which aims to predict the average response over several hundred milliseconds. Correspondingly, the stimuli are defined by their spectrum over this period. Their spectrum is normally distributed around a chosen reference level. By ignoring the temporal dimension these so-called Random Spectral Shape (RSS) stimuli are only one dimensional. This approach is motivated by the separability of the corresponding STRFs. It therefore becomes possible to fit a second order, memoryless Wiener model which predicts the average rate from the stimulus spectrum (Yu and Young (2000) call it quadratic model). It has been applied to ANFs (Young and Calhoun, 2005) and neurons from the ventral and dorsal cochlear nucleus (Yu and Young, 2000; Bandyopadhyay et al., 2007). Quantitatively, model performance is in a similar range as for the present model. However, excluding the temporal dimension reduces the complexity of the problem. Conversely, a time-resolved model could also be used to predict the average rate by suitable averaging. Based on the present data and models we would predict that average rate predictions from time-resolved models should be more precise than predictions of the quadratic model at least for the ANF-AVCN-MNTB pathway. Structurally, RSS based models are probably hardest to interpret. Leaving out temporal aspects strongly reduces the possibility to attribute model properties to underlying structures.

Recently, Bandyopadhyay et al. (2007) introduced another RSS based model which explicitly accounts for level-dependencies in each frequency channel (level-dependent weighting model, LDWM). For stimuli with larger spectral contrast (12 dB) the LDWM outperformed the quadratic model which in comparison seems to be better matched for interactions between frequencies at low contrast. Interestingly, the LDWM is contained within the presently considered class of multilinear models if the time dimension is removed in a context model with the frequency-level grouped and a minimal (just the frequency bin itself), linear CRF. From this perspective both a transition to a temporally resolved model and the use of the multilinear estimation and advanced regularization methods would be a natural step.

For both the spectrogram and RSS based approaches, a complete model will always need to include a representation of the transformation from the sound pressure wave to the spectral or spectrotemporal representation. From this perspective, the present comparison of stimulus representations should be attributed to the model rather than the stimulus.

While we demonstrate that about 75% of the variance is explainable based on the stimulus spectrogram, we have noticed certain influences of the stimulus fine-structure on the response. Presenting the same TORC profile with different carrier phases led to characteristic differences in the responses. These appeared to be local modulations of the spike timing, possibly induced by spurious, brief correlations between carrier frequencies. Due to the importance of spike-timing for sound localization, a focus of future models could be on the interplay between spectrum and fine structure. Further, it would be of interest to know whether the assumption of a random process characterized by the firing rate is valid or whether the trial-to-trial variability can be explained by neuronal models which include the spiking history (Ahrens et al., 2008b).

To achieve this fine-structure match, it would probably be wise to integrate structure based models of early processing, e.g. cochlear processing as in Zhang et al. (2001) and add phenomenological models only for subsequent stages (the γ-tone stimulus representation could be interpreted along these lines). Recio-Spinoso et al. (2005) provided a simple, yet convincing, structure based account for the shape of their second order Wiener kernels. Besides a more direct correspondence between the model and modeled system, this approach usually results in a reduction in the number of free parameters. However, phenomenological models are likely to keep their importance as the numerous and recurrent stages of the auditory system might resist a completely structural model.

Physiological interpretation of model parameters

The orderly organization of the kernel parameters along stimulus dimensions facilitates extracting meaningful properties of the encoding machinery. The spectrotemporal location of excitatory and inhibitory fields can guide the delay structure in a parametric model. Adding the multiplicative context offers further insight into the underlying structure: if a nonlinear system is modeled by a linear system, the estimated linear kernels can be confounded by correlations with nonlinear terms (Christianson et al., 2008). Introducing nonlinear terms into the model can help to identify the natural constituents of the system. The quadratic terms of the context model form an example of such nonlinearities. Including them in the model enables to estimate their contribution which otherwise would have erroneously been attributed to the linear parts.

We focus here on the origin of the LF/HF inhibitory fields. It would be desirable to distinguish cochlear suppression from neuronal inhibition which could be introduced in the AVCN or presynaptically at the MNTB. Postsynaptic inhibition in the MNTB of the gerbil is unlikely given recent results (Mc Laughlin et al., 2008; Englitz et al., 2009). Cochlear suppression is a nonlinear interaction between neighboring frequency channels (Robles and Ruggero, 2001; Zhang et al., 2001) and could thus account for the presently observed low and high frequency inhibitory fields (relative to CF, see Fig. 8(D)). Their strong reduction in the case of the context model suggests a nonlinear, e.g. multiplicative underlying source, akin to the response properties obtained in the classical two-tone paradigm used for probing cochlear suppression in the auditory nerve (Sachs and Kiang, 1968). This hypothesis is further supported by the tight temporal relationship between the LF/HF fields and CF (Fig. 9(C)), possibly reflecting the mechanical basis of cochlear suppression. Further evidence for suppression comes from a study at ANFs (Kim and Young, 1994), where inhibitory fields were found in STRFs in similar spectrotemporal relation to the CF (Note that STRFs were estimated in Kim and Young (1994) based on the Wigner transform of the signal without correction for stimulus correlations.). To substantiate this possibility, we estimated the present models for the output of an auditory nerve model (Zhang et al. 2001), obtaining similar spectrotemporal relationships between LF/HF fields and CF (data not shown). If on the other hand neuronal inhibition accounted for the LF/HF inhibitory fields, it would have to be rapid (thus excluding recurrent interactions) and probably divisive rather than additive (e.g. in the absence of an excitatory stimulus, spontaneous rates should not be reducible).

Context models of auditory cortex neurons

The context model was first applied to responses from neurons of primary auditory areas of the cortex (primary auditory cortex (A1), of rats and mice, and anterior auditory field of mice).

Predicting cortical responses is clearly a more challenging task which is also reflected in the greater predictive power obtained in the MNTB (75%) than in cortex ((Ahrens et al. (2008a): 32–52%, although further improvements were recently achieved, MBA, MS, Jennifer F. Linden, unpublished results). On a structural level, it would be desirable to compare the kernels obtained from the cortical areas and the present nucleus. However, this task is complicated by the different time-scales (20 ms vs. 0.5 ms) and stimuli (Dynamic random chord vs. TORC) used in estimating the models. Despite these differences an interesting agreement exists for the nature of the suppressive fields: inhibitory fields were also found to have multiplicative origin (Ahrens et al., 2008a). Whether this observation has the same basis is debatable as the spectrotemporal profiles in cortex are quite different. Especially the attribution to cochlear suppression can be made in the MNTB with greater confidence due to the higher temporal resolution and the smaller set of alternatives. On the other hand the observed model structure may reflect general auditory processing mechanisms, reflecting the time-scale invariance found in the structure of auditory signals (Attias and Schreiner, 1997).

In future studies it would be desirable to compare the kernels obtained in different areas in detail to understand the intermediate transformations. To this end it would be advantageous to use the same stimulus in different areas, e.g. the inferior colliculus could be compared to the cortical data by stimulating with the dynamic random chord or to lower brainstem data by stimulating with TORCs.

Acknowledgements

BE was funded by the Graduate College InterNeuro at the University of Leipzig (1097). ST and RR acknowledge funding from the Deutsche Forschungsgemeinschaft (Ru 390–18/1). MBA and MS acknowledge funding by the Gatsby Charitable Foundation. Further, we thank Rainer Stollhoff and two anonymous reviewers for comments on an earlier version of the manuscript.

Note

[1] For simpler notation we here compare different trials for a given stimulus rather than different portions of a stimulus.

References

Aertsen A, Johannesma PIM, Hermes DJ. 1980. Spectro-temporal receptive-fields of auditory neurons in the grassfrog II. Analysis of the stimulus-event relation for tonal stimuli. Biol Cyb 38(4):235–248.

Aertsen A, Olders JHJ, Johannesma PIM. 1981a. Spectro-temporal receptive fields of auditory neurons in the grassfrog. I. Characterization of Tonal an Natural Stimuli. Biol Cyb 39(3):195–209.

Aertsen A, Olders JHJ, Johannesma PIM. 1981b. Spectro-temporal receptive-fields of auditory neurons in the grassfrog III. Analysis of the stimulus-event relation for natural stimuli. Biol Cyb 39(3):195–209.

Ahrens MB, Linden JF, Sahani M. 2008a. Nonlinearities and contextual influences in auditory cortical responses modeled with multilinear spectrotemporal methods. J Neurosci 28(8):1929–42.

Ahrens MB, Paninski L, Sahani M. 2008b. Inferring input nonlinearities in neural encoding models. Network 19(1):35–67.

Attias H, Schreiner C. 1997. Temporal low-order statistics of natural sounds. Adv in Neural Inf Proc Syst 27–33.

Awatramani GB, Turecek R, Trussell LO. 2004. Inhibitory control at a synaptic relay. J Neurosci 24(11):2643–7.

Backoff PM, Clopton BM. 1991. A spectrotemporal analysis of DCN single unit responses to wideband noise in guinea pig. Hear Res 53(1):28–40.

Bandyopadhyay S, Reiss LAJ, Young ED. 2007. Receptive field for dorsal cochlear nucleus neurons at multiple sound levels. J Neurophysiol 98(6):3505–15.

Bernstein LR. 2001. Auditory processing of interaural timing information: New insights. J Neurosci Res 66(6):1035–46.

Bernstein LR, Trahiotis C. 2002. Enhancing sensitivity to interaural delays at high frequencies by using "transposed stimuli". J Acoust Soc Am 112(3 Pt 1):1026–36.

Brand A, Behrend O, Marquardt T, McAlpine D, Grothe B. 2002. Precise inhibition is essential for microsecond interaural time difference coding. Nature 417(6888):543–7.

Carney LH. 1993. A model for the responses of low-frequency auditory-nerve fibers in cat. J Acoust Soc Am 93(1):401–17.

Christianson GB, Sahani M, Linden JF. 2008. The consequences of response nonlinearities for interpretation of spectrotemporal receptive fields. J Neurosci 28(2):446–55.

Clopton BM, Backoff PM. 1991. Spectrotemporal receptive fields of neurons in cochlear nucleus of guinea pig. Hear Res 52(2):329–44.

De Boer E, De Jongh HR. 1978. Cochlear encoding – potentialities and limitations of reverse-correlation technique. J Acoust Soc Am 63(1):115–135.

Depireux DA, Simon JZ, Klein DJ, Shamma SA. 2001. Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. J Neurophysiol 85(3):1220–34.

Efron B, Tibshirani R. 1993. An Introduction to the Bootstrap. New York: Chapman & Hall.

Elhilali M, Fritz JB, Klein DJ, Simon JZ, Shamma SA. 2004. Dynamics of precise spike timing in primary auditory cortex. J Neurosci 24(5):1159–72.

Englitz B, Tolnai S, Typlt M, Jost J, Rübsamen R. 2009. Reliability of synaptic transmission at the synapses of Held *in vivo* under acoustic stimulation. PLoS ONE 4(10):e7014.

Escabi MA, Miller LM, Read HL, Schreiner CE. 2003. Naturalistic auditory contrast improves spectrotemporal coding in the cat inferior colliculus. J Neurosci 23(37):11489–504.

Gill P, Zhang J, Woolley SMN, Fremouw TE, Theunissen FE. 2006. Sound representation methods for spectro-temporal receptive field estimation. J Comput Neurosci 21(1):5–20.

Guinan JJ, Li RY. 1990. Signal processing in brainstem auditory neurons which receive giant endings (calyces of Held) in the medial nucleus of the trapezoid body of the cat. Hear Res 49(1–3):321–34.

Haefner R, Cumming B. 2008. An improved estimator of Variance Explained in the presence of noise. Adv in Neural Inf Proc Syst 1–8.

Joris PX, Yin TCT. 1998. Envelope coding in the lateral superior olive. III. Comparison with afferent pathways. J Neurophysiol 79(1):253–69.

Kim PJ, Young ED. 1994. Comparative analysis of spectro-temporal receptive fields, reverse correlation functions, and frequency tuning curves of auditory-nerve fibers. J Acoust Soc Am 95(1):410–22.

Klein DJ, Depireux DA, Simon JZ, Shamma SA. 2000. Robust spectrotemporal reverse correlation for the auditory system: Optimizing stimulus design. J Comput Neurosci 9(1):85–111.

Kopp-Scheinpflug C, Dehmel S, Dörrscheidt GJ, Rübsamen R. 2002. Interaction of excitation and inhibition in anteroventral cochlear nucleus neurons that receive large endbulb synaptic endings. J Neurosci 22(24):11004–18.

Lewis ER, Henry KR, Yamada WM. 2002. Tuning and timing in the gerbil ear: Wiener-kernel analysis. Hear Res 174(1–2):206–21.

Lyon, R. 1982. A computational model of filtering, detection, and compression in the cochlea. Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '82 7:1282–1285.

Machens CK, Wehr MS, Zador AM. 2004. Linearity of cortical receptive fields measured with natural sounds. J Neurosci 24(5):1089–100.

Mc Laughlin M, van der Heijden M, Joris PX. 2008. How secure is *in vivo* synaptic transmission at the calyx of Held?. J Neurosci 28(41):10206–19.

Moore MJ, Caspary DM. 1983. Strychnine blocks binaural inhibition in lateral superior olivary neurons. J Neurosci 3(1):237–42.

Pecka M, Brand A, Behrend O, Grothe B. 2008. Interaural time difference processing in the mammalian medial superior olive: the role of glycinergic inhibition. J Neurosci 28(27):6914–25.

Recio-Spinoso A, Temchin AN, van Dijk P, Fan Y.-H, Ruggero MA. 2005. Wiener-kernel analysis of responses to noise of chinchilla auditory-nerve fibers. J Neurophysiol 93(6):3615–34.

Reiss LAJ, Bandyopadhyay S, Young ED. 2007. Effects of stimulus spectral contrast on receptive fields of dorsal cochlear nucleus neurons. J Neurophysiol 98(4):2133–43.

Robles L, Ruggero MA. 2001. Mechanics of the mammalian cochlea. Physiol Rev 81(3):1305–52.

Sachs MB, Kiang NY. 1968. Two-tone inhibition in auditory-nerve fibers. J Acoust Soc Am 43(5):1120–8.

Sahani M, Linden JF. 2003a. Evidence Optimization Techniques for Estimating Stimulus-Response Functions. Adv in Neural Inf Proc Syst :109–116.

Sahani M, Linden JF. 2003b. How Linear are Auditory Cortical Responses?. Adv in Neural Inf Proc Syst 15:301–308.

Temchin AN, Recio-Spinoso A, van Dijk P, Ruggero MA. 2005. Wiener kernels of chinchilla auditory-nerve fibers: verification using responses to tones, clicks, and noise and comparison with basilar-membrane vibrations. J Neurophysiol 93(6):3635–48.

Tolnai S, Englitz B, Kopp-Scheinpflug C, Dehmel S, Jost J, Rübsamen R. 2008. Dynamic coupling of excitatory and inhibitory responses in the medial nucleus of the trapezoid body. Eur J Neurosci 27(12):3191–204.

Tolnai S, Englitz B, Scholbach J, Jost J, Rübsamen R. 2009. Spike transmission delay at the calyx of Held *in vivo*: rate dependence, phenomenological modeling, and relevance for sound localization. J Neurophysiol 102(2):1206–17.

Young ED, Calhoun BM. 2005. Nonlinear modeling of auditory-nerve rate responses to wideband stimuli. J Neurophysiol 94(6):4441–54.

Yu JJ, Young ED. 2000. Linear and nonlinear pathways of spectral information transmission in the cochlear nucleus. Proc Natl Acad Sci USA 97(22):11780–6.

Zhang X, Heinz MG, Bruce IC, Carney LH. 2001. A phenomenological model for the responses of auditory-nerve fibers: I. Nonlinear tuning with compression and suppression. J Acoust Soc Am 109(2):648–70.

Appendix A: Estimating fraction of explained variance from predictive power

As outlined in Methods (and detailed in Sahani and Linden (2003b)) the predictive power provides an improved estimate over the explained variance, but has a bias which varies as a function of noise power and model complexity. This bias can be (approximately) avoided by considering the average between the insample and crossvalidation predictive powers. The validity of this approximation will be demonstrated in the following using analytic and simulation tools.

Firstly, we derive that the bias depends linearly on the the noise power, if the underlying and estimated models are both linear. Secondly, we demonstrate that under these conditions the bias also depends linearly on model complexity with the same slope for both insample and crossvalidation. Finally, we illustrate these dependencies in a simple model system and compare it to an analytical correction proposed by Haefner and Cumming (2008). While this analytical correction would in principle be preferrable over the present approach it is only valid for unregularized estimates.

A.1. β as a function of $P(\eta)$

For population analysis Sahani and Linden (2003b) and Ahrens et al. (2008a) used polynomial fits to extrapolate the predictive power of linear and multilinear models to the noise free condition. We here provide a justification for linear extrapolation in the case of a linear system.

The underlying system is assumed to be given by $\boldsymbol{r}=\boldsymbol{X}'\boldsymbol{w}'+\boldsymbol{\eta}$ with normally distributed errors $\boldsymbol{\eta}$, stimulus representation $\boldsymbol{X}'$, and kernel $\boldsymbol{w}'$. $\boldsymbol{w}'$ is of dimension $M'\times 1$, $\boldsymbol{X}$ is $T\times M'$, and $\boldsymbol{r}$ and $\boldsymbol{\eta}$ are $T\times 1$, with T the number of time steps and M' the dimension of the linear system.

To estimate $\boldsymbol{w}'$, we use an extended linear model $\boldsymbol{X}\,\boldsymbol{w}$, where $\boldsymbol{X}$ is $T\times M$ and $\boldsymbol{w}$ is $M\times 1$. For the present considerations we will assume $M\geq M'$ and $\boldsymbol{X}'$ to be contained in $\boldsymbol{X}$. These conditions mean that our model dimension and stimulus representation are adequately chosen to estimate the underlying system. Hence, we can formally express the system by this model via $\boldsymbol{r}=\boldsymbol{X}'\boldsymbol{w}'+\boldsymbol{\eta}=\boldsymbol{X}\;\boldsymbol{w}+\boldsymbol{\eta}$, if $\boldsymbol{w}$ equals $\boldsymbol{w}'$ in the corresponding dimensions of $\boldsymbol{X}$ and $\boldsymbol{X}'$ and is 0 otherwise. Distinguishing the dimensions of system and model is necessary to study the dependence of β on the dimension of the model.

The maximum likelihood estimate of the kernel $\hat{\boldsymbol{w}}$ is then given by the normal equation

$$\hat{\boldsymbol{w}}=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{r}=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{w}+\boldsymbol{\eta})=\boldsymbol{w}+(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\eta} \tag{2}$$

To obtain the dependence of β on $P(\eta)$ for the insample estimate, we insert $\hat{\boldsymbol{w}}$ into the definition of β

$$\begin{aligned}\beta_{\text{insample}}&=\frac{P(\boldsymbol{r})-P(\boldsymbol{e})}{\hat{P}(\boldsymbol{\mu})}=\frac{P(\boldsymbol{X}\boldsymbol{w}+\boldsymbol{\eta})-P(\boldsymbol{r}-\boldsymbol{X}\hat{\boldsymbol{w}})}{\hat{P}(\boldsymbol{\mu})}\\&\overset{(2)}{=}\frac{P(\boldsymbol{X}\boldsymbol{w})+P(\boldsymbol{\eta})-P(\boldsymbol{X}\boldsymbol{w}+\boldsymbol{\eta}-\boldsymbol{X}\boldsymbol{w}-\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\eta})}{\hat{P}(\boldsymbol{\mu})}\\&=\frac{P(\boldsymbol{\mu})+P(\boldsymbol{\eta})-P(\boldsymbol{\eta}-\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\eta})}{\hat{P}(\boldsymbol{\mu})}\\&=C_1+\frac{P(\boldsymbol{\eta})-P\big(\boldsymbol{\eta}-\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\eta}\big)}{\hat{P}(\boldsymbol{\mu})}\end{aligned} \tag{3}$$

with $C_1\approx 1$, if $\hat{P}(\boldsymbol{\mu})$ is accurate. Analogously, we obtain for the crossvalidation predictive power (with $\boldsymbol{\eta}_k$ and $\boldsymbol{\eta}_u$, denoting the noise in the known training and unknown test data[1], respectively)

$$\begin{aligned}\beta_{\text{crossval}} &= \frac{P(\boldsymbol{r}) - P(\boldsymbol{e})}{\hat{P}(\boldsymbol{\mu})}\\ &= \frac{P(\boldsymbol{Xw} + \boldsymbol{\eta}_u) - P(\boldsymbol{Xw} + \boldsymbol{\eta}_u - \boldsymbol{Xw} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\eta}_k)}{\hat{P}(\boldsymbol{\mu})}\\ &= \frac{P(\boldsymbol{\mu}) + P(\boldsymbol{\eta}_u) - P(\boldsymbol{\eta}_u) - P(\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\eta}_k)}{\hat{P}(\boldsymbol{\mu})}\\ &= C_1 - \frac{P(\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\eta}_k)}{\hat{P}(\boldsymbol{\mu})} \qquad (4)\end{aligned}$$

In both relationships the last term is linear in $P(\boldsymbol{\eta})$ since a k-fold scaling of $\boldsymbol{\eta}$ leads to k^2 scaling of $P(\boldsymbol{\eta})$ and $P(A\boldsymbol{\eta})$ for a $T \times T$ matrix A. For a given $P(\boldsymbol{\mu})$ the predictive power will hence also be linear in $P_{\boldsymbol{\mu}}(\eta)$.

A.2. β as a function of model complexity

In the insample and the crossvalidation expression the last term quantifies the model-dependent misestimate: $\boldsymbol{w}_\eta := (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\eta}$ is the normal equation if one takes $\boldsymbol{\eta}$ as the response. The multiplication $\boldsymbol{Xw}_\eta$ then generates the linear model prediction. Hence, in the crossvalidation case, the slope is determined by the power of the prediction based on the training noise. In the insample case, the slope is determined by the power of the difference between noise and noise-based prediction.

With $\boldsymbol{X}$ an $T \times M$ matrix, $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ has maximally rank M, the dimensionality of the model. Low dimensional models will have low rank and a shallow slope of overestimation for insample or underestimation for crossvalidation. High dimensional models correspondingly have high slopes. This hypothesis is supported in simulations for $\boldsymbol{X}$ and $\boldsymbol{\eta}$ drawn from independent, Gaussian distributions and $T = 1000$ (Fig. 10(A)).

We can extend these results to regularized linear regression. Denoting the covariance matrix as $\boldsymbol{C}$, the maximum likelihood estimate of the parameters is given by

$$\hat{\boldsymbol{w}} = (\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{C}^{-1})^{-1}\boldsymbol{X}^T\boldsymbol{r} = (\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{C}^{-1})^{-1}\boldsymbol{X}^T(\boldsymbol{Xw} + \boldsymbol{\eta})$$

For the insample case one similarly obtains

$$\beta_{\text{insample}} = C_1 + \frac{P(\boldsymbol{\eta}) - P\big(\boldsymbol{\mu} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{C}^{-1})^{-1}\boldsymbol{X}^T\boldsymbol{\mu}\big) - P\big(\boldsymbol{\eta} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{C}^{-1})^{-1}\boldsymbol{X}^T\boldsymbol{\eta}\big)}{\hat{P}(\boldsymbol{\mu})} \qquad (5)$$

and for the crossvalidation case

$$\beta_{\text{crossval}} = C_1 - \frac{P\big(\boldsymbol{\mu} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{C}^{-1})^{-1}\boldsymbol{X}^T\boldsymbol{\mu}\big) + P\big(\boldsymbol{\eta} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{C}^{-1})^{-1}\boldsymbol{X}^T\boldsymbol{\eta}\big)}{\hat{P}(\boldsymbol{\mu})} \qquad (6)$$

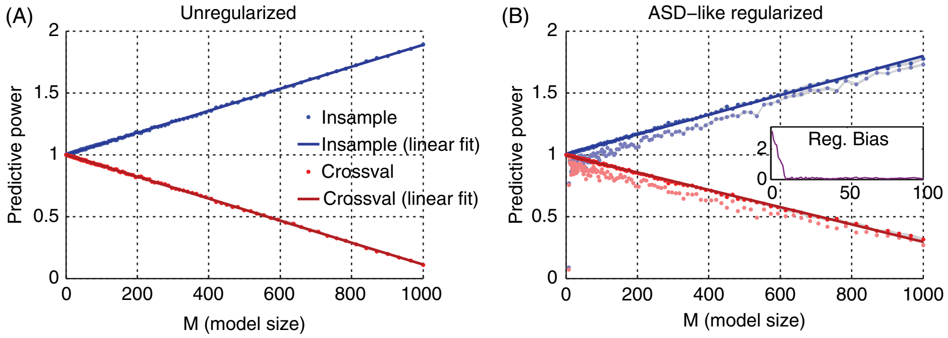

Figure 10. Predictive power as a function of model dimension. (A) For unregularized estimation the insample and crossvalidation predictive powers depend linearly on model dimension. (B) For ASD-like regularized estimates the dependence is influenced by two factors. The $\boldsymbol{\eta}$ related term in Equations 5 and 6 again has a linear relationship with reduced slopes due to the regularisation (solid red and blue dots and lines). Taking into account the $\boldsymbol{\mu}$ related term, leads to a negative bias on both curves (light red and blue dots). The value of this bias is shown in the inset. As long as $M \geq M' = 10$, the bias remains small. For the simulations it was assumed $P(\boldsymbol{\mu}) = 1$, $\boldsymbol{\eta} \sim N(0, 1)$, and $\boldsymbol{X} \sim N(0, 1)$ for all dimensions and points in time.

The stimulus dependent power terms again quantify the residuals of the linear model estimate with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\eta}$. The term containing $\boldsymbol{\mu}$ occurs as a consequence of the regularization, accounting for the possibility that the assumed covariance on the parameters conflicts with the actual covariance. It will always reduce the predictive power for both insample and crossvalidation results, especially for low model dimensions. The term containing $\boldsymbol{\eta}$ will now exhibit a weaker dependence on M, consistent with the aim of regularisation.

Fig. 10(B) shows the results of numerical simulations for the case of ASD-type regularization. The values of the scale and smoothing parameters were set to fixed values rather than being data dependent. The nonlinear dependence observed in Sahani and Linden (2003b) might be explained by the necessity to estimate the ASD parameters from the data.

A.3. A simple example

The following example is meant to illustrate the behavior of the various quantities considered above. An artificial response of length L was created by convolving a unimodal, positive Gaussian kernel w' of length N_w with a Gaussian white noise stimulus S. To simulate the presence of systematic but stimulus unrelated aspects of the response, a simple sinusoidal modulation was added. The noise model was chosen to be Gaussian white noise with variance 1 drawn independently between trials, i.e. $\eta \sim N(0, 1)$. Altogether, sets of N responses were given as

$$\{r_n\}_{n=1\ldots N} = w' * S + a\sin(2\pi f\, t) + \sigma\eta_n$$

where $*$ denotes convolution. We then used standard linear regression to estimate models w of different complexity N_w (i.e. corresponding to the length of the kernel in this one-dimensional setting) under a range of noise strengths (matched to the

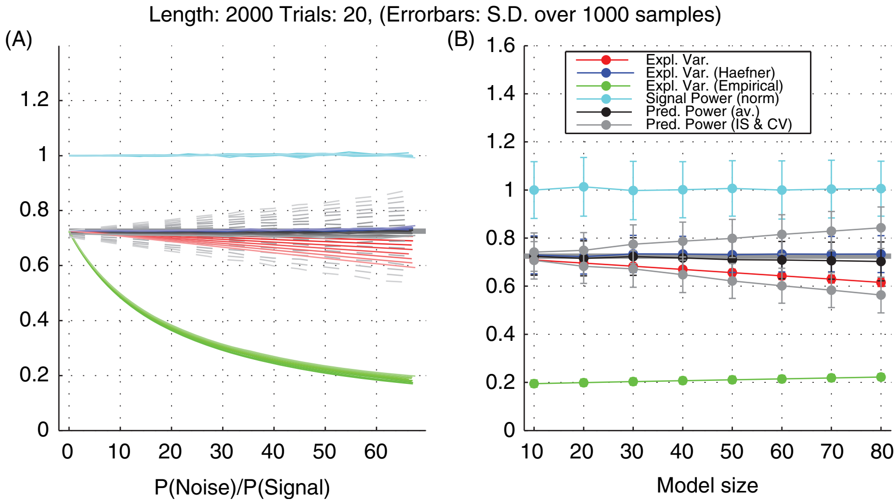

Figure 11. Estimators as a function of noise power and model size. (A) The actual fraction of explained variance (red), the predictive power (grey, insample above and crossvalidation below 0.7), and the corrected fraction of explained variance (blue) change quite linearly as a function of noise power with the slope regulated by the model size and almost no offset. The estimated fraction of explained variance (green) decreases rapidly with noise power and only weakly depends on the model size. Different shades of a color indicate the different sizes of the estimated model (saturation decreases with model size; errorbars omitted for clarity, color legend as in B). (B) The dependence on model size is also approximately linear for the above mentioned (linear) quantities (here at normalized noise power ≈50). Both the analytic and the average (between insample and crossvalidation predictive power) estimators stay remarkably flat as a function of noise power, rendering both a good estimator for the actual fraction of explained variance in the noise free case. Further, the estimate of the signal power (turquoise) is quite accurate in expectation. The actual fraction of explained variance in the noise free case is ≈0.7 due to the additional, stimulus independent signal component.

experimentally observed range) and for a number of available data-points L and repetitions N.

Figure 11 shows the estimates of the relevant quantities both as a function of normalized noise power (A) and as a function of model size (B). The depicted quantities are the fraction of explained variance estimated directly from response and prediction (green), the fraction of explained variance estimated against the noisefree response (red), the predictive power (for insample & crossvalidation (gray) and their average (black)) and the corrected fraction of explained variance described in Haefner and Cumming (2008) (blue). Haefner and Cumming (2008) proposed a correction which takes both noise and model complexity into account. We implemented this correction according to Eq. 8 from Haefner and Cumming (2008). For reference the estimate of the signal power from Sahani and Linden (2003b) is also shown (cyan). Errorbars indicate S.D. (to indicate the variability we chose not to use SEM) over 1000 samples.

Both estimation methods provide a good estimate of the true fraction of explained variance, the analytical estimator with a close, yet slight overestimate, while the average estimator underestimates the true value by a few percent. As noted above, the analytical estimator was not applicable to the regularized (multi)linear models in the main text as it only applies to unregularized linear models.

We have performed the same estimates for data lengths ranging from 1000 to 10000 points and up to 40 repetitions with qualitatively similar results. If the ratio between the model size (number of parameters) and available data points/samples is taken as the relevant quantity for convergence, then the numbers depicted in Fig. 11 are approximately in the correct range, since we have more data points (≈11000) but also larger models (100–800 parameters).

Appendix B: Tensor products

The generalized tensor product can be defined as

$$\mathbf{M}^{\mathbf{i_1,\ldots,i_n,j_1,\ldots,j_m}} = \mathbf{M}^{\mathbf{i_1,\ldots,i_n}} \otimes \mathbf{M}^{\mathbf{j_1,\ldots,j_m}},$$

$$\text{with } M^{\mathbf{i_1,\ldots,i_n,j_1,\ldots,j_m}}_{i_1,\ldots,i_n,j_1,\ldots,j_m} := M^{\mathbf{i_1,\ldots,i_n}}_{i_1,\ldots,i_n} M^{\mathbf{j_1,\ldots,j_m}}_{j_1,\ldots,j_m}$$

for example a 2D kernel can be constructed from two 1D kernels

$$\mathbf{w}^{\mathbf{tf}} = \mathbf{w}^{\mathbf{t}} \otimes \mathbf{w}^{\mathbf{f}} \text{ with } w^{\mathbf{tf}}_{jk} = w^{\mathbf{t}}_j w^{\mathbf{f}}_k$$

Further, the generalized inner product is defined by

$$\mathbf{M}^{\mathbf{i_1,\ldots,i_{k_1},i_{k_1+m+1},\ldots,i_{k_2}}} = \mathbf{M}^{\mathbf{i_1,\ldots,i_{k_1},\ldots,i_{k_1+m}}} \bullet \mathbf{M}^{\mathbf{i_{k_1+1},\ldots,i_{k_1+m},\ldots,i_{k_2}}},$$

$$\text{with } M^{\mathbf{i_1,\ldots,i_{k_1},i_{k_1+m+1},\ldots,i_{k_2}}}_{i_1,\ldots,i_{k_1},i_{k_1+m+1},\ldots,i_{k_2}} := \sum_{i_{k_1+1}} \cdots \sum_{i_{k_1+m}} M^{\mathbf{i_1,\ldots,i_{k_1},\ldots,i_{k_1+m}}}_{i_1,\ldots,i_{k_1},\ldots,i_{k_1+m}} M^{\mathbf{i_{k_1+1},\ldots,i_{k_1+m},\ldots,i_{k_2}}}_{i_{k_1+1},\ldots,i_{k_1+m},\ldots,i_{k_2}}$$

for example a 2D matrix can be contracted with a 1D kernel to another 1D kernel

$$\mathbf{Q}^{\mathbf{f}} = \mathbf{w}^{\mathbf{t}} \bullet \mathbf{M}^{\mathbf{tf}} \text{ with } Q^{\mathbf{f}}_k = \sum_j w^{\mathbf{t}}_j M^{\mathbf{tf}}_{jk}$$