Modeling Natural Sounds with Gaussian Modulation Cascade Processes

Anonymous Author(s)
Affiliation
Address
City, State/Province, Postal Code, Country
email

Abstract

The computational principles underpinning auditory processing are not well understood. This fact stands in stark contrast to early visual processing for which computational theories, and especially those built on statistical models, have recently enjoyed great success. We believe one of the reasons for this disparity is the paucity of rich, learnable generative models for natural scenes with an explicit temporal dimension. To that end we introduce a new generative model for the dynamic Fourier components of sounds. This comprises a cascade of modulatory processes which evolve over a wide range of time-scales. We show the model is capable of capturing both the sparse marginal distribution and the prevelance of amplitude modulation in natural sounds, to which the auditory system appears to listen so attentively. Moreover, we demonstrate that it is relatively easy to learn and to do inference in the Gaussian Modulation Cascade Process, due to the structure of its non-linearity. We hope that this provides a first step toward furthering our understanding of auditory computations.

1 Introduction

Natural sounds have a very rich temporal structure, that spans a wide range of time scales. For example, a typical section of speech (Fig. 1.) might contain formants at the finest temporal granularity (sub-milli second), pitch information (milli-second), phonemes (tens of milli-seconds), syllables (hundreds of milli-seconds), and finally the longest components: words and sentences (seconds). Shortly, we will argue that both the statistics of sound and the architecture of the auditory system suggest that a number of these aspects of speech, and other natural sounds, can be interpreted as a cascade of modulatory processes, operating over a wide range of time scales. Motivated by this insight, and guided by recent efforts in the statistical modeling of natural scenes, we introduce a new generative model - the Gaussian Modulation Cascade Process (GMCP) - for the dynamic Fourier components of natural sounds. One reason for working with such a representation is that it is similar to what the first layer of neurons in auditory system are faced with when sounds emerge at the basilar membrane. As such it is hoped in the future, that learning and inference in this model might shed a computational light on the hitherto murky world of auditory processing, just as indpendent component analysis (ICA) and sparse coding (SC) models have for the visual system.

Statistics of AM in sounds. A number of studies have probed the statistical stucture of natural sounds. A predominant characteristic is the extreme sparsity of the marginal amplitude distribution, which is considerably more kurtotic than its visual counterpart [1]. Two prevelent features of natural sound ensembles seem to be responsible: First, there is an abundance of soft sounds in natural ensembles [2]; for example, the relatively long pauses found between utterances in speech sounds (see Fig. 1). Second, there are rare, highly structured, localised events that carry substantial parts

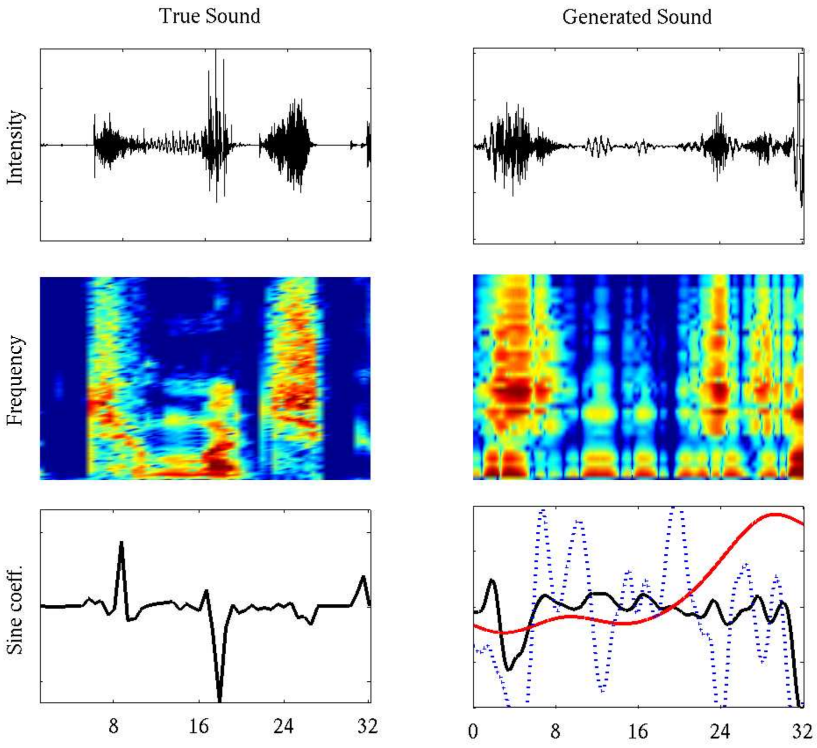

Figure 1: A comparison of a real sound (left column) and one generated from the forward model (right column). Top row: Sound pressure waveform. Middle row: Spectrogram. Bottom row: A typical Fourier sine coefficient, for the real sound, and for the simulated sound. The slowest and next to slowest modulator in the cascade are also shown.

of the sound energy [1]. Taken together, frequent low-energy sounds and infrequent high-energy events result in the sparse marginal statistics of sound energy.

The temporal structure of sounds is particularly rich, and is exemplified by the low-frequency amplitude modulation (AM) that is prominent in natural environments [3]. This modulation often carries important information: for example interaural envelope time differences are a salient cue for source localisation. Attias and Schreiner's systematic analysis of the statistics of AM [2] showed that there are extensive correlations in modulation, both across carrier frequencies, and over a wide range of time-scales up to $\sim 100ms$. Interestingly, they discovered a translation invariance in these statistics: in a sense each point on the cochlea 'sees' the same amplitude modulation statistics.

These studies suggest that a good generative model of sounds should capture both the highly kurtotic maginal distribution of sound amplitudes, and the rich, translationally-invariant amplitude-modulated structure, that spans a wide range of time scales.

AM in the Auditory system. AM is a feature of many natural signals and, as one might expect, is important to the auditory system. Psychoacoustically, AM impacts many tasks, over a wide range of time-scales. One example is the well-known phenomenon of "comodulation masking release", in which a tone masked by noise with a bandwidth greater than an auditory filter becomes audible if the noise masker is amplitude modulated (see [4] for a review). This suggests that envelope information is processed and analysed across frequency channels in the auditory system.

Electrophysiological data on the encoding of AM, also point to an important role in auditory processing. Although the data are still patchy, it is known that envelope information is abundant at the first stage of the auditory system: Type-I auditory nerve fibres phase-lock envelope of sounds (as well as their fine structure) and each nerve fibre transmits information over a stereotypical range of modulation frequencies, carrier frequencies and intensities [5]. Moving along the neuraxis to the cochlear nucleus and then to the inferior colliculus (IC) the tuning to AM typically shows larger gain, smaller bandwidth (200-300Hz) and the tuning changes from low pass, to more band pass. Interestingly, there is evidence for a tonotopic mapping of modulation frequency sensitivity in the IC, running perpendicularly to the carrier frequency tonotopy [6], although this finding is still debated. Little is known about cortical processing of AM, but temporal coding of AM seems to be limited to modualtions lower than 30Hz. Interestingly, and unlike lower levels of auditory processing, the bandwidth of this tuning appears to be independent of the centre frequency of the cell, suggesting that there is now independent processing of modulation frequency in each spectral band. This has led some authors to propose that cortex carries out a type of modulation filter bank analysis.

Although there is a skeptical perspective that the electrophysiological results are epiphenomena, there is an opposing view that amplitude modulation is a fundamental organising principle of the auditory system [5]. This suggests that a good generative model that incorporates AM structure could benefit our understanding of computations performed in the auditory system.

2 Previous statistical models of natural scenes

Principal components analysis was the first statistical model applied to natural scenes. Essentially a linear Gaussian model, it is clear that it fails to capture the highly kurtotic, non-Gaussian structure present in natural stimuli. This was one of the motivations behind the development of the Independent components analysis (ICA, [7]; table 1, column 1) and sparse coding (SC, [8] table 1, column 2) algorithms. These two related methods improved upon PCA, modeling the latent causes as sparse and independent and have had great success as computational models for cortical processing of visual stimuli. Lewicki [9] applied ICA to natural sounds showing the resulting filters have a time-frequency tiling resembling that of auditory nerve filters. However there is no temporal dimension in ICA, and as a consequence short overlapping segments of sound are treated as if they are independent ‘images’. In such a form, ICA is not a true generative model for sounds. Clearly this is undesirable and a good model for movies and sounds should have an explicit temporal dimension.

As well as neglecting correlations through time, ICA and SC also tend to recover latent variables that are decorrelated, but not entirely independent [10]. This statistical dependancy takes the form of correlations in the power of the latent variables extracted from natural scenes. Wainwright et al [11] showed how to derive a new prior distribution for ICA that takes account of these correlations: As a first step, the recognition distribution of the new model is chosen to be the same as for ICA $p(\mathbf{x}|\mathbf{y}) = \delta(\mathbf{x} - R\mathbf{y})$. For the complete case this fixes the generative distribution $p(\mathbf{y}|\mathbf{x}) = \delta(\mathbf{y} - G\mathbf{x})$, where $G = R^{-1}$. A prior is then chosen to match the statistics of images: $p(\mathbf{x}) = \int d\mathbf{y} p(\mathbf{y}) p(\mathbf{x}|\mathbf{y})$, which we can approximate by taking lots of samples from images and running them through the recognition distribution: $p(\mathbf{x}) \propto \sum_{\mathbf{y}} \delta(\mathbf{x} - R\mathbf{y})$. Wainwright et al show that the histograms that result are found to be well approximated by (infinite) mixtures of Gaussians with differing variances, so called Gaussian scale mixture (GSM) priors [1]. One way of generating a Gaussian scale mixture prior is by multiplying a Gausian random variable by an independent positive scalar random variable called a multipier. Let us take one moment to generalise this to the temporal setting. Regarding the modulators as slowly varing envelopes and the Gaussian as quickly varying fine structure, we see that this feature of natural scenes is consistent with a prevelance of strong AM.

Karklin and Lewicki [12] showed how to generalise the original GSM framework (table 1, column 3), placing a generalised log-normal prior on the variances of the coefficients with a variance matrix that allows sharing of multipliers. Essentially, the top layer of latent multiplier variables in their model represent contrast patterns in the variances of the coefficients of the lower-order representation. They show how to learn the sharing of the multipliers and the lower level weights, discovering that for images, power is shared between latents with broadly similar basis functions. In passing we note the similarity to the auditory system in which amplitude modulations are highly correlated across quite broad frequency regions.

Meaningful components can be extracted from from natural scenes using sparseness as a heuristic. However, in a parallel avenue of research, slowness has also been shown to be a useful heuristc for extracting meaningful components [13]. Hyvarinen ([14], see table 1, column 4) takes this as a sign that latent causes of natural images are slow and sparse and this is why both methods perform similarly. To this end he devised a proof-of-concept generative model for movies, and equivalently

[1] More recent work has assumed the priors are largely unaffected when moving to the overcomplete and stochastic case

Table 1: Generative models for natural stimuli

	ICA	**SC**	**GSMs**	**Bubbles**
$p(x^{(2)})$			$\mathrm{Norm}(\mathbf{0}, I)$	point process like
$p(x^{(1)} \vert x^{(2)})$	sparse	sparse	$\mathrm{Norm}(\mathbf{0}, \lambda_i^2)$	$\mathrm{Norm}(\mathbf{0}, \lambda_{i,t}^2)$
	e.g. 1/cosh	e.g. cauchy	$\lambda_i^2 = \exp(\mathbf{h}_i^T \mathbf{x}^{(2)})$	$\lambda_{i,t}^2 = f(\mathbf{h}_i^T \mathbf{x}_t^{(2)} \otimes \phi_t)$
$p(y \vert x^{(1)})$	$\delta(\mathbf{y} - G\mathbf{x}^{(1)})$	$\mathrm{Norm}(G\mathbf{x}^{(1)}, \sigma_y^2 I)$	$\mathrm{Norm}(G\mathbf{x}^{(1)}, \sigma_y^2 I)$	$\delta(\mathbf{y} - G\mathbf{x}^{(1)})$

sounds, which has a temporally smooth, sparse prior of the GSM flavour. He shows how to learn the bottom set of weights using the likelihood as a guide for the sort of terms that should be present in a suitable cost function. Unlike Karklin and Lewicki however, the neigbourhood of dependence on the multiplier-latents and the dynamics of the bubble is hard-wired.

In summary, there has been significant progress in the statistical modeling of natural scenes over the past ten years. For auditory signals most work has treated spectrograms like images, but clearly a more suitable generative model would have an explicit temporal dimension. The exception to this rule is the proof-of-concept bubbles framework. However, in this case the complicated temporal prior makes learning problematic. In contrast recent work on Gaussian scale mixture models has centred on learning this heirarchical prior. It seems plausible that this prior is a signature of the marginal distribution of AM processes, collapsed into a non-temporal setting.

In many of these methods it is common to learn the parameters using zero-temperature EM. Essentially, the recognition distribution is approximated as a delta function and consequently uncertainty and correlational information is lost. It is not known to what extent this effects learning. Furthermore, to compare the results of inference in the probabilistic model to neural data we have to specify a mapping from the recognition distribution to neural responses. As most work has a delta function recognition disribution (either due to simple models or simple approximations) the mode of the posterior has been exclusively used as a regressor. In general we have more freedom; indeed we believe neural populations will represent uncertainty and correlations in latent variables. As such it is useful to retain variance and correlational information, both for learning and for comparison to biology.

3 Motivating the Gaussian Modulation Cascade Process Model

The aim of this paper is to produce a generative model for natural sounds, which exhibits at least four desirable properties:

1. The output of the model should be sparse, mimicking the kurtotic distribution of Fourier coefficients in natural scenes.
2. Unlike ICA and GSMs, the model should have an explicit temporal dimension, and the latent variables (and therefore the output) should vary smoothly over time.
3. The model should have a hierarchical prior that captures the AM statistics of sounds at different time scales. The hierarchy will form a cascade of modulatory processes, with slowly varying processes at the top of the hierarchy modulating more rapidly varying signals towards the bottom.
4. The model should be learnable; and we would like to preserve information about the uncertainty, and possibly correlations, in our inferences.

Our approach is to multiply samples drawn from a set of Gaussian processes with varying correlation lengths. The details of the model are presented in the next section; here we make two core observations that motivate this choice. First, the product of two or more Gaussian random variables is sparsely distributed. For a pair of random variables, this is just a GSM distribution with a multiplier drawn from a rectified Gaussian. Second, it is relatively simple to build smooth temporal priors for Gaussian distributed latent variables using Gaussian processes and linear dynamical systems. A hyper-prior can be placed over the time-scales of these processes enabling us to build a cascade. In the next secation we will see that these two observations enable us to build a generative model capable of capturing rich statistical structure present in sounds. Importantly, due to the Gaussian structure of the model, we can derive a family of fast variational learning algorithms.

4 Gaussian Modulation Cascade Processes

In this section we introduce the Gaussian Modulation Cascade Process (GMCP) model. We first state the model mathematically, before providing more detail on the emission distribution, and then the temporal dynamics. The output of the model is a dynamic Fourier representation of the sampled sound waveform, in terms of the D (real) sine and cosine components obtained from each D-sample-long window. These coefficients ($\mathbf{x}_t$) are generated from a multilinear combination of M multi-dimensional latents $\mathbf{y}_t^{(m)}$, with additive Gaussian noise. The individual latents $y_{k_m,t}^{(m)}$ evolve independently over time according to smooth linear Gaussian dynamics[2]:

$$p(\mathbf{y}_t|\mathbf{x}_t^{(1:M)}, \mathbf{g}_{k_1:k_M}, \sigma_y^2) \quad = \quad \text{Norm}\left(\sum_{k_1:k_M} \mathbf{g}_{k_1:k_M} \prod_{m=1}^{M} x_{k_m,t}^{(m)}, \sigma_y^2 I\right) \qquad (1)$$

$$p(x_{k,t}^{(m)}|x_{k,t-1:t-\tau_m}^{(m)}, \lambda_{k,1:\tau_m}^{m}, \sigma_{m,k}^2) \quad = \quad \text{Norm}\left(\sum_{t'=1}^{\tau_m} \lambda_{k,t'}^{(m)} x_{k,t-t'}^{(m)}, \sigma_{m,k}^2\right) \qquad (2)$$

For completeness, $\forall \quad t \in \{(2-\tau_m) : 1\}$, the latent variables are drawn from unit Gaussians.

The emission distribution. The non-linear emission distribution is best understood through a simple example: imagine we have a cascade of two process $M = 2$, one of which is two dimensional $K_1 = 2$, and the other one dimensional $K_2 = 1$, then the mean of the output is given by:$\langle \mathbf{y}_t \rangle = \left(\mathbf{g}_{11} x_{1,t}^{(1)} + \mathbf{g}_{21} x_{2,t}^{(1)}\right) x_{1,t}^{(2)}$. In this case $\mathbf{g}_{11}$ and $\mathbf{g}_{21}$ define directions in the output space - collections of frequency channels - that participate in a common fine structure feature. The strength of these features are controlled independently by $x_1^{(1)}$ and $x_2^{(1)}$. The power in the features is correlated due to the common modulator $x_1^{(2)}$. Fig 1. shows a draw from the forward model for a more complicated example where $M = 3$ and $K_m = [9, 3, 1]$. Promisingly, for a large region of parameter space, samples from these models share many features with natural sounds.

This multi-linear emission distribution is similar to those used by Tenenbaum and Freeman [15] and Grimes and Rao [16] when $M = 2$, and Vasilescu and Terzopoulos for general M [17]. Importantly, these non-temporal settings cannot disambiguate the effects of the various latent variables without using data sets which are essentially labeled by the identity of the component in the product that has changed. The hope is that the temporal slowness prior can break these degeneracies and facilitate completely unsupervised learning.

The temporal dynamics. As stated earlier, we expect the latent variables to vary smoothly over time. Moreover, latent variables higher in the hierarchy (corresponding to larger values of m) should vary progressively more slowly. As stated above the model does not require either of these. One way to remedy this, is to specify a distribution over the power spectrum of each of the latents. A flexible parameterisation for the power spectrum is a sum of Gaussians:

$$P(\omega) \quad = \quad \sum_p \gamma_p \left(\exp\left[-\frac{1}{2\sigma_p^2}(\omega - \mu_p)^2\right] + \exp\left[-\frac{1}{2\sigma_p^2}(\omega + \mu_p)^2\right]\right) \qquad (3)$$

Rather than learning λ, which can contain rather large numbers of parameters, the parameters of the power spectrum can be learned directly. Furthermore, a prior distribution over power-spectra can be induced by specifying a prior distribution over the peak heights, centres and widths $[p(\gamma_{1:P}, \mu_{1:P}, \sigma_{1:P}^2)]$. For example, a sensible prior might tend to assign lower values of m higher values of σ_p $[\mu_p]$, thereby making them faster variables. Finally we have to relate this new parameterisation of the joint distribution to the conditional distribution above. This can be achieved by forming the inverse Fourier transform of the power spectrum, and from this autocorrelation the conditional can be computed.

[2] To make the notation more compact, throughout the following: $a : b \overset{\text{def}}{=} a, a+1, ..., b-1, b$ or $z_a : z_b \overset{\text{def}}{=} z_a, z_{a+1}, ..., z_{b-1}, z_b$

5 Variational Learning in the GMCP model

One of the criteria for our model was that algorithms for parameter identification be tractable. A key observation towards this end is that latent variables at level m in the hierarchcy, $x^{(m)}_{1:K_m,1:T}$, are Gaussian distributed when conditioned on the observations and on the other $M-1$ latent variables. This means it is relatively straightforward to learn the parameters of the model using variational expectation maximisation (vEM) which is a fast approximate learning algorithm [18]. Briefly, vEM optimises a lower bound on the log-likelihood, $\log p(Y|\theta)$, which is formed by approximating the recognition distribution $p(X|Y,\theta)$ by a simpler distribution $q(X)$. This bound, called the variational free-energy, is optimised sequentially with respect to the parameters (the M-Step) and then with respect to the distribution over the hidden variables (the E-Step). A common approach is to make a structural approximation, which assumes some factorisation of the posterior; $q(X) = \prod_i q(x_i)$ and then derives the best parametric form for this factorisation, yielding the new E-Step updates:

$$q(x_i) \quad = \quad \frac{1}{Z_i} \exp\left[\langle \log p(Y,X)\rangle_{\prod_{i'\neq i} q(x_{i'})}\right]. \tag{4}$$

In fact, there are a family of choices for the factorisation of the approximate distribution over latent variables in the GMCP. The key feature that makes learning simple is that component of the factored posterior will be a Gaussian distribution. The first and simplest approximation is fully factored, the second factorises across time and levels in the hierarchy, the third is factored across levels within a hierarchy and the fourth, and richest, is factored across levels in the hierarchy alone.

	Factored over k	**Unfactored over k**
Factored over t	$q_1(X) = \prod_{k,t,m} q(x^{(m)}_{k,t})$	$q_3(X) = \prod_{t,m} q(x^{(m)}_{1:K(m),t})$
Unfactored over t	$q_2(X) = \prod_{k,m} q(x^{(m)}_{k,2-\tau(m):T})$	$q_4(X) = \prod_{m} q(x^{(m)}_{1:K(m),2-\tau(m):T})$

The M-Step is identical for each of these approximations (see below); the methods differ only in the way the sufficient statistics are calculated. A useful way to understand the variational approximation is to think of the approximate conditionals above, as the *exact* conditional to a different joint. As an example, consider approximation 4, for which the update for the distributions is:

$$\begin{aligned} Q_m(\mathbf{x}^{(m)}_{1:T}) \quad &= \quad \frac{1}{Z_m}\prod_k \left[p(x^{(m)}_{k,1}) \prod_{t=2}^{T} p(x^{(m)}_{k,t}|x^{(m)}_{k,t-1:t-\tau})\right] \\ &\quad \times \prod_{t=1}^{T} \exp\langle \log p(\mathbf{y}_t|\mathbf{x}^{(1:M)}_t, \mathbf{g}_{k_1:k_M})\rangle_{\prod_{n\neq m} Q_n}. \end{aligned} \tag{5}$$

We interpret this as having the form: $Q_m(\mathbf{x}^{(m)}_{1:T}) = \frac{1}{Z_m} Q_m(\mathbf{x}^{(m)}_{1:T}, \tilde{\mathbf{y}}^{(m)}_{1:T})$, which upon comparison with eq. 5 means the pseudo-joint distribution $Q_m(\mathbf{x}^{(m)}_{1:T}, \tilde{\mathbf{y}}^{(m)}_{1:T})$ takes a relatively simple form: The dynamics of the chains in the variational approximation are equivalent to those in the true system and the remaining terms can be interpreted as a new distribution: $P(\tilde{\mathbf{y}}^{(m)}_t|\mathbf{x}^{(m)}_t) = \mathrm{Norm}(\mathbf{x}^{(m)}_t, \Gamma^{(m)}_t)$. Therefore, the pseudo-joint can be recognised as a regular linear Gaussian state space model where the $\tilde{\mathbf{y}}_t$ have the useful interpretation as pseudo-observations and there is time-varing noise given by:

$$\left[\Gamma^{(n)}_{k_n,k'_n,t}\right]^{-1} \quad = \quad \sum_{\substack{k_1:k_M\neq n \\ k'_1:k'_M\neq n}} \frac{1}{\sigma^2_y} \mathbf{g}^T_{k_1,\ldots,k_n,\ldots,k_M} \mathbf{g}_{k'_1,\ldots,k'_n,\ldots,k'_M} \prod_{m\neq n} \langle x_{k_m,t} x_{k'_m,t}\rangle \tag{6}$$

$$\tilde{y}^{(n)}_{k'_n,t} \quad = \quad \frac{1}{\sigma^2_y} \sum_{k_1:k_M} \Gamma^{(n)}_{k'_n,k_n} \mathbf{y}^T_t \mathbf{g}_{k_1:k_M} \prod_{m\neq n} \langle x^{(m)}_{k_m,t}\rangle \tag{7}$$

The advantage of this perspective is that we can use the machinery developed for regular linear dynamical systems (the Kalman smoother and lag-minus-one covariance smoother) to calculate the sufficient statistics required for the M-Step. The caveat being that they have to be suitably modified to deal with non-stationary observation noise.

Finally we sketch the M-Step updates for the parameters, firstly those of the emission distribution, $\frac{d\langle \log P(X,Y|\theta)\rangle}{d\mathbf{g}_{k_1:k_M}} = 0$ yields:

$$\sum_{t=1}^{T} \mathbf{y}_t \prod_{m=1}^{M} \langle x^{(m)}_{k_m,t}\rangle \quad = \quad \sum_{k'_1:k'_M} \mathbf{g}_{k'_1:k'_M} \sum_{t=1}^{T}\prod_{m=1}^{M} \langle x^{(m)}_{k_m,t} x^{(m)}_{k'_m,t}\rangle \tag{8}$$

The above is a linear equation and can therefore be solved by re-writting as a matrix multiplication. Alternative methods will be needed when $\prod_m K_m \geq 1000$. The ouput noise update is very similar to that for a regular linear dynamical, however in contrast, the updates for the dynamical parameters have no closed form and have to be made using a gradient-based method, like conjugate-gradients. The derivarives of the free-energy with respect to the new parameters are found using the chain rule:

$$\frac{dF}{d\log\theta} = \sum_{t'} \frac{dF_{old}}{d\lambda_{t'}} \frac{d\lambda'_t}{d\theta} \frac{d\theta}{d\log\theta} + \frac{dF_{old}}{d\sigma^2} \frac{d\sigma^2}{d\theta} \frac{d\theta}{d\log\theta} + \frac{d\log p(\theta)}{d\log\theta} \tag{9}$$

Evaluation of variational learning The objective of learning is to maximise the likelihood of the parameters, and the previous section introduced a family of variational approximations to do this, approximately. It would be useful to know how different members of this family perform. This is not a simple problem, for although we might expect the bounds to be ordered: $L(\theta) \geq F(\theta, q_4) \geq F(\theta, q_3), F(\theta, q_2) \geq F(\theta, q_1)$, this says little about where the peaks of the free-energies fall relative to the peak of the likelihood. Furthermore, for real world applications with time constraints, it is entirely possible that the speed of the first approximation would bring it closer to the summit of the free-energy in limited time.

One way to evaluate these methods is to move to a simpler system in which the true likelihood can be computed, and to use this to evaluate the various methods. By setting $M = 1$ we can do exactly that (as this results in a normal linear dynamical system). Clearly this system does not have the troublesome output non-linearity of the full model, but we will return to this issue later. Fig 2A. shows a comparison of the 4 methods in this simplified setting, with different settings of the observation noise. The observation noise is a critical parameter, as it controls how influential the prior is for inference. Similarly, the strength of the dynamical parameters is also important, because as the temporal correlations vanish approximation 3 contains the true model class. In the experiments below, fairly long temporal correlations were used. Approximation 4 is exact for this model and is thus a tight bound to the likelihood, but it takes slow steps and so it finds parameters with a low likelihood. Approximation 2 gives the next tightest bound to the likelihood (data not shown), however in low output noise conditions it performs worse than 3 and 1 in terms of the likelihood of the parameters it discovers. This is for two reasons, first because it is slower (also requiring Kalman smoothing) and second because it does not capture the correlational structure between chains which is a feature of the posterior in low output noise conditions. It therefore has a maximum in the wrong place. However, in low noise conditions it describes dynamical information more exactly than the other two approximate methods and hence performs better. The converse is true for approximation 3. Suprisingly, the simple fully factored approximation works fairly well over a wide range, often with a performance intermediate between methods 2 and 3.

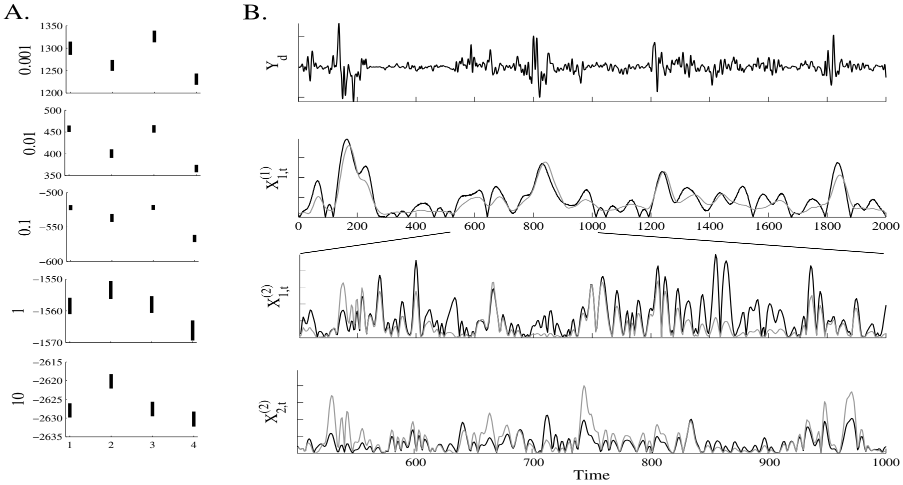

Figure 2: A) Comparison of the four different variational approximations, showing the likelihood of the parameters after learning under five different settings of the true observation noise (shown on the left of the plots). The length of the black bar represents the uncertainty in the likelihood and the centre, the mean value. B) The results of vEM on a toy example. The top panel shows the first dimension of the traing data. The second, the modulus of the true value of the slowest modulatory process (black) and the inference (gray). Similarly, the third and fourth panels show the inferences for the faster processes.

These experiments indicate that the variational approximations work reasonably well in a wide range of parameter settings. Notably the simple approximations can outperform the exact model, because they are faster (scaling like $K\tau$ or $K\tau^2$ and not $K^2\tau^2$). An interesting area of research is how to choose the variational approximation on the fly using criteria such as the free-energy. Importantly these experiments do not address the question of how reliable the variational approximation will be when the output non-linearity is present. This non-linearity contributes a hyperbolic likelihood function which is combined with a temporal prior to produce the posterior. If the temporal prior is broad then the variational distribution must approximate the resulting bannana using an axis aligned Gaussian. Clearly this approximation will break down when the curvature of the posterior is too great.

Results Finally we provide computational simulations on a toy example that indicate that learning and inference is viable in the GMCP. In these simulations we used the simplest, fully factored variational approximation. A data set of $T = 2000$ samples of dimension $D = 40$ were drawn from the forward model with $M = 2$ levels. Within this, there were $K_1 = 2$ fast processes (with length scales ≈ 3 time steps) and one slow modulation $K_2 = 1$ (with a length scale ≈ 30 time steps). The dynamics of these chains were drawn from the priors, which were log-normal and had an effective width of about 20% of the means. Experiments showed that the free-energy landscape has many small local maxima, but a large peak at the true parameters. Initialisation is thus critical and a successful heuristic was to initialise the sufficient statistics of the slow process first, for example using a Hilbert transform of the input which recovers an approximation to the modulation envelope. In turn, the fast processes can be initialised using the observations, normalised by the slow process.

The results of inference after learning are shown in Fig. 2B. The algorithm recovers a close approximation to the true latent variables, although there is ambiguity up to a sign. For instance, it is fairly common in the generative process for two modulatory processes to cross zero at approximately the same time. From the persepective of inference this is ambiguous and an equally plausible explanation of the data might be that both processes approached zero, but subsequently smoothly diverged without changing sign. For this reason the absolute value of the inferences is compared to the absolute value of the true latent variables, and there is a good correspondence between the two.

6 Conclusion

We have argued that the statistics of sound and the behaviour of the auditory system suggest that AM is an important dimension of natural sounds. Indeed, current non-temporal models of natural scenes appear to bare the hallmarks of AM, and this suggests they should be extended into the temporal domain. To that end, we propose a new generative model of the dynamic Fourier coefficients of sounds, which consists of a cascade of modulatory processes, operating over a wide range of timescales. Due to the structure of the non-linearity in the model, it is ameable to variational EM learning and a family of algorithms have been presented of that type. A simple example verifies the worth of these methods, and it is hoped, provides a first step along the road to using the model to further our understand of auditory processing.

References

[1] Iordanov, L. G. & Penev P.S. (1999) *Adv in Neural Info Processing Sys 11*. MIT Press.

[2] Attias H. & Schreiner C. E. (1997) *Adv in Neural Info Processing Sys 9*. MIT Press.

[3] Nelken I, Rotman Y, and Bar Yosef O. (1999) *Nature* 397:154-157.

[4] Moore, B.C.J. (2003) An Introduction to the Psychology of Hearing. San Diego, CA: Academic.

[5] Joris P.X., Schreiner, C.E. & Rees, A. (2004) *Physiol Rev* 84:541-577

[6] Schreiner, C.E. & Langner G. (1988) *J Neurophysiol* 60:1823-1840.

[7] Bell, A.J. & Sejnowski, T.J. (1997). *Vision Res*, 37(23):3327-3338.

[8] Olshausen, B.A., & Field, D.J. (1996). *Nature*, 381(6583):607-609.

[9] Lewicki, M.S. (2002) *Nature Neurosci* 5(4):356-363.

[10] Buccigrossi, R.W., & Simoncelli, E.P. (1999) *IEEE Trans Image Proc* 8(12):1688-1701.

[11] Wainwright, M., Simoncelli, E.P., & Willsky, A. (2001) *Appl Comput Harmonic Anal* 11(1):89-123.

[12] Karklin, Y. & Lewicki, M.S. (2005) *Neural Comput* 17(2):397-423.

[13] Berkes, P., & Wiskott, L. (2005) *J Vis*, 5(6)579-602.

[14] Hyvärinen, A., Hurri, J. & Väyrynen J. (2003) *J Opt Soc America* 20(7):1237-1252

[15] Tenenbaum, J.B., & Freeman, W.T. (2000) *Neural Comput*, 12(6):1247-1283.

[16] Grimes, D.B., & Rao, R.P.N. (2005) *Neural Comp.* 17(1):47-73.

[17] Vasilescu, M.O.A., & Terzopoulos D. (2002) *Proc of ECCV*, Copenhagen, Danmark,. 447-460.

[18] Jordan, M.I., Ghahramani, Z., Jaakkola, T., & Saul, L.K. (1999) *Mach Learning*, 37(2):183-233.