Demodulation as Probabilistic Inference

Richard E. Turner*, *Member, IEEE*, and Maneesh Sahani, *Member, IEEE*

***Abstract*—Demodulation is an ill-posed problem whenever both carrier and envelope signals are broadband and unknown. Here, we approach this problem using the methods of probabilistic inference. The new approach, called Probabilistic Amplitude Demodulation (PAD), is computationally challenging but improves on existing methods in a number of ways. By contrast to previous approaches to demodulation, it satisfies five key desiderata: PAD has soft constraints because it is probabilistic; PAD is able to automatically adjust to the signal because it learns parameters; PAD is user-steerable because the solution can be shaped by user-specific prior information; PAD is robust to broad-band noise because this is modelled explicitly; and PAD's solution is self-consistent, empirically satisfying a Carrier Identity property. Furthermore, the probabilistic view naturally encompasses noise and uncertainty, allowing PAD to cope with missing data and return error bars on carrier and envelope estimates. Finally, we show that when PAD is applied to a bandpass-filtered signal, the stop-band energy of the inferred carrier is minimal, making PAD well-suited to sub-band demodulation.**

***Index Terms*—Carrier, demodulation, envelope, inference, learning.**

I. Introduction

DEMODULATION is the process by which a signal (y_t) is decomposed into the product of two component signals: a slowly varying envelope or modulator component (m_t) and a quickly varying carrier component (c_t). That is,

$$\mathrm{y}_t = \mathrm{m}_t \mathrm{c}_t. \tag{1}$$

Demodulation was originally developed for radio communications where the carrier is a sinusoid of known frequency, but it has since been applied to a range of audio processing problems including voice coding [1], [2], speech recognition [3], [4], music retrieval [5], speech enhancement [6] and source separation [7], [8], and it is used in hearing devices [6], [9]. In most of these applications, the underlying signal representation is derived by demodulating the sub-bands of the recorded signal. Indeed, the time-frequency spectrogram, a very widely used tool of signal processing, can be viewed as yielding just such a representation [10] further highlighting the importance of demodulation. Demodulation methods have also been used to investigate the relative importance of the sub-band envelopes and sub-band carriers (known collectively as the fine-structure) in the perception of sounds [9], [11]–[16]. However, the conclusions that can be drawn from these studies are limited by several well-known problems with the demodulation methods employed [6], [17]–[19], although a recently proposed approach [20], very similar to that advocated here, may help to address these limitations.

The central problem for any demodulation algorithm is that, in its most general form, the demodulation problem is ill-posed [21]; any modulator that is non-zero wherever the signal is non-zero can be associated with a valid carrier, and *vice versa*. Thus to achieve repeatable results, an algorithm must impose implicit or explicit assumptions about the form of carrier and envelope, often embodied by a set of constraints. For instance, in an amplitude-modulated radio signal the carrier is a sinusoid of known frequency very much higher than the pass-band of the modulator. Imposing this knowledge makes the demodulation problem well-posed and straightforward. Unfortunately, in applications involving natural audio signals, there is no clear separation between the carrier and modulator bands and so a more sophisticated approach to designing constraints is needed.

Arguably, a general approach to demodulation should impose constraints on the component signals that are soft (that is, violations incur penalties but do not necessarily rule out a candidate decomposition), and that can adapt automatically to the signal, but which are still steerable if required. Both softness and adaptability are needed to handle the variability and potential non-stationarity of the components of the signal. They allow the algorithm to identify suitable bandlimits for the carrier and envelope signals from the measured sound, and to permit temporary, or otherwise minor violations of these limits if the resulting solution is better in an overall sense. At the same time, specific knowledge about the properties of the signal generators or desired decomposition may provide partial or approximate information about component properties in some applications. In such cases it would be valuable if this knowledge could be used to steer the outcome of the demodulation algorithm.

At least two further properties seem desirable in the context of natural audio demodulation. The first is robustness to additive noise. Natural signals are often corrupted by broadband noise, and one might wish for this noise to have minimal impact on the recovered modulator [22]. Indeed, joint demodulation and denoising is essential for many practical applications. The second is self-consistency. There are many types of consistency property [21], [23], but of particular interest here is a criterion we call Carrier Identity, which requires that demodulating a recovered carrier yield a constant envelope signal. This is similar to the Modulator Identity property introduced in the preceding references, which requires that demodulating a recovered modulator yield an envelope equal to the original modulator (possibly rescaled), and a constant carrier. Both of these criteria enforce consistency—

R. E. Turner is with the Computational and Biological Learning Lab, University of Cambridge, Cambridge, CB2 1PZ, England (e-mail ret26@cam.ac.uk).

M. Sahani is with the Gatsby Computational Neuroscience Unit, University College London, London, WC1N 3AR, England (e-mail maneesh@gatsby.ucl.ac.uk)

Manuscript received July 1, 2010

that demodulation should remove all modulation information from the carrier, and *vice versa*—but the Modulator Identity property may be more difficult to satisfy as it demands that the second demodulation stage produce a modulator that varies more quickly than the (constant) carrier.

The discussion above proposes five desiderata for a demodulation algorithm in the context of natural signals:

1) soft constraints,
2) automatic adaptation to the signal,
3) user steerability,
4) robustness to noise,
5) self-consistency.

We argue in section II that no single existing approach to demodulation meets all of these desiderata. This observation motivates our development in section III of a new approach called Probabilistic Amplitude Demodulation (PAD) that naturally satisfies them all.

PAD is a new framework which views demodulation as a problem of *inference* and *learning*, where we adopt the usage of these terms from Machine Learning. In our context, inference means the estimation of the modulator and carrier signals from the recorded samples, given fixed parametric distributional constraints for both. Learning means the estimation of the parameters that describe these distributional constraints, such as the expected time-scale of variation of the modulator, and the modulation depth. In other words, inference corresponds to demodulation, whilst learning corresponds to adaptation of the algorithm to the signal.

The starting point for PAD is to articulate a probabilistic forward model (see section III-A) which is a statistical description of the carrier, modulator, and the way in which they combine to form the signal. Bayesian probabilistic calculus is then used to invert the forward model and thereby estimate the carrier and the modulator from the signal (see section III-B). The PAD solution is shaped by the assumptions specified in the forward model, but these constraints are imposed in a soft fashion because they are probabilistic. Furthermore, we show in section III-C that the parameters of the model, like the time-scale of the modulator, can be learned from the signal by maximum-likelihood or similar techniques. This enables the algorithm to automatically adapt to novel signals. Additionally, any knowledge the user has about these parameters can be incorporated in prior distributions, thereby enabling the algorithm to be steered. In sections IV-A and IV-B we show that PAD is robust to noise. In fact, it is simple to incorporate the noise explicitly in the forward model and, if not known *a priori*, learn its level from the signal. Finally, sections IV-A and IV-B demonstrate that PAD is self consistent in the sense that it approximately satisfies the important Carrier Identity property.

Thus PAD meets all the desiderata we have laid out, but this comes at the price of increased computational cost. PAD uses a range of well-established, but computationally-demanding methods for probabilistic inference. For instance, demodulation requires the iterative optimisation of a non-linear cost function. A main focus of this research has been to accelerate the algorithm and currently signals with a sampling rate of 16KHz can be demodulated in real time on a modern laptop computer.

Although the new approach is computationally challenging it does bring with it several advantages over and above the desiderata mentioned earlier. Unlike in most existing approaches, the unavoidable assumptions that determine the solution are stated explicitly in the specification of the forward model. This makes PAD easy to understand, critique, and improve. Moreover, PAD can return error-bars on the estimated modulators and carriers. These are especially relevant if signals are noisy or contain quickly varying modulators, because there can be considerable uncertainty in the carrier and modulator estimates in such cases. The ability to handle uncertainties also enables the range of demodulation tasks to be generalised, for example to signals containing missing regions in which the modulator must be filled-in (see sections IV-A and IV-B). This is an interesting application as signals with missing segments can arise in many ways: from device drop-out, damage to physical media, as a consequence of the removal of impulsive noise, or from loss of network packets. Restoration of such signals requires reconstruction of the missing sections. The modulator information in such reconstructions will often be a perceptually important component [24].

II. Background

There are many existing demodulation algorithms. Here we argue that no single earlier algorithm satisfies all of the desiderata introduced in the previous section. We limit our discussion to methods which return positive envelope signals as this is the focus of the paper.

Two classic techniques are the Square and Low-Pass (SLP) method [25] and the Hilbert Envelope (HE) [26] approach. The SLP method squares the signal to move modulator energy to low frequencies, where it is then picked off by low-pass filtering. The method is exact when the signal is composed of a high-frequency narrow-band carrier, and a low frequency modulator. When applied to more complex signals, a reasonable modulator can be extracted by judicious choice of the low-pass filter cut-off, although this parameter must be set by hand. Even then, the recovered carrier is often poor. This is because the filtered envelope may be small, or even zero, in regions where the signal is non-zero. This results in an associated carrier which is very large, possibly unbounded. Overall, the method fails desiderata 1, 2, and 5.

The failure of the SLP method to return bounded carrier estimates, and the need to set the low-pass filter, are both issues that are addressed by the HE demodulation approach. The HE, given by the magnitude of the analytic signal formed from the measurements, is guaranteed to return a bounded carrier and requires no hand-tuning. Like the SLP method, the HE is invariant to amplitude scale changes, and returns a constant envelope signal for pure sinusoidal input, both useful theoretical properties. However, it still suffers from several problems. Practically, the method performs poorly when the carriers are not single tones. For example, if the signal is a pair of harmonically-related sinusoids that undergo slow modulation, the HE will contain a contribution at the fundamental

frequency no matter how slowly the true envelope varies. Consequently, the HEs extracted from natural sounds often contain pitch information, even though many applications seek to separate this from modulation content [22]. Similarly, the HE is sensitive to noise in the signal, and therefore it is not robust. The HE has theoretical problems too. For example, the Hilbert carrier (formed by dividing the signal by the HE) can be discontinuous for continuous signals. Furthermore, the Hilbert carrier is not limited to the same frequency region as the signal—which leads to reconstruction problems when using demodulated sub-bands [17]. With regard to the desiderata enumerated here, the method fails all but number 2.

An alternative approach to demodulation is to focus on estimating the carrier and then to recover the modulator by division. Coherent approaches to demodulation [27] assume that the carrier is a frequency-modulated (FM) sinusoid, and estimate the instantaneous frequency of this carrier by, for instance, finding the spectral centre of gravity of the windowed signal. Coherent demodulation is usually applied in the complex domain, with the FM structure being grouped with the amplitude modulation by the inclusion of a time-varying complex phase. For this reason, the approach is not directly comparable to the one taken here which assumes a real-valued positive modulator. Coherent methods work well when the carrier is well-approximated as a single sinusoid (e.g. when operating on a narrow sub-band of the signal), but like the HE method, fail when the ideal carrier is more complicated. Moreover, parameters like the window time-scale have significant impact, and the method would benefit from an automatic procedure. Thus, while parameters are available to steer the results of coherent demodulation, and the method does exhibit consistency by some definitions, the approach generally fails the other desiderata, including the Carrier Identity property.

Here, we view demodulation as a Bayesian inference problem. We first introduced this perspective some years ago, using a simple version of PAD [28]. The main goal of that original paper was to extend PAD to handle cascades of modulators with different time-scales. That line of work was then generalised to multi-band PAD [29]. In the current paper, we return to consider single-band demodulation in more depth and extend PAD in several new directions. First we consider a more sophisticated and flexible model than that used in the original work. Second, we provide methods for learning all of the free-parameters of the model, which enable the model to automatically adapt to the signal. Third, we present methods for accelerating inference. The utility of these new methods is then demonstrated on synthetic and natural signals, in complete-data, noisy-data and missing-data tasks.

Probabilistic amplitude demodulation proceeds by optimising a non-linear cost function. This is potentially problematic as the optimisation can be slow and there can be multiple (local) optima. Recently, in an elegant paper, Sell and Slaney [22] develop a more computationally efficient demodulation algorithm that optimises a convex cost function [30] and therefore ensures the problem has a unique solution. In section III-D we show that Sell and Slaney's linear-demodulation algorithm can be viewed as a version of PAD. This perspective is important as it reveals the assumptions implicit in their method and it means that the machinery developed in this paper for learning the free-parameters of PAD models can also be applied in the convex case, allowing it to automatically adapt to the signal.

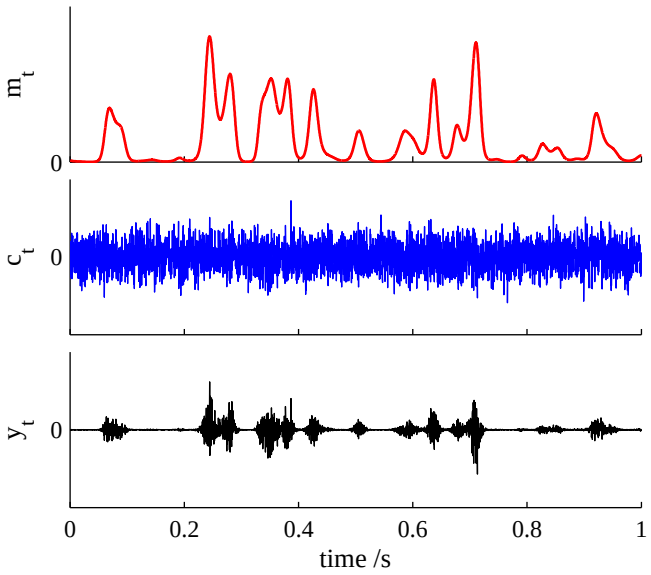

Fig. 1. A sample from the GP-PAD forward model produced using parameter values learned from a segment of speech. The top panel shows the slowly varying envelopes. The middle panel shows the quickly varying carriers. The bottom panel shows the generated signal which is formed by the product of the signals above.

III. Probabilistic Amplitude Demodulation

This section covers the theoretical development of PAD. In order to prevent the main ideas from being obscured by technical detail, we begin this section with a high-level roadmap, highlighting the relationship with the desiderata.

The starting point is the forward model [31], which is a description of the process by which we assume the signal is generated. In the present context, the forward model assumes that, (1) the observed signal is formed from a product of an unknown modulator signal with an unknown carrier, (2) the carrier is quickly varying, and (3) the modulator is slowly varying and positive. This information is encoded probabilistically in, (1) the likelihood $p(\mathrm{y}_{1:T}|\mathrm{c}_{1:T}, \mathrm{m}_{1:T}, \theta)$, (2) the prior distribution over the carrier, $p(\mathrm{c}_{1:T}|\theta)$, and (3) the prior distribution over the modulators, $p(\mathrm{m}_{1:T}|\theta)$. (The notation $\mathrm{x}_{1:T}$ represents all the samples of the signal x, running from 1 to a maximum value T). Each of these distributions depends on a set of parameters, θ, which controls factors such as the typical time-scale of variation of the modulator or the frequency content of the carrier. A specific set of modelling assumptions, in the form of specific choices for the distributions above, may be tested by drawing samples from the forward model (see Fig. 1). In general, a balance must be struck between the accuracy of the modelling assumptions and the tractability of inference.

The forward model specifies the parametrised joint proba-

bility of the signal, carrier and modulator

$$p(\mathrm{y}_{1:T}, \mathrm{c}_{1:T}, \mathrm{m}_{1:T}|\theta) = p(\mathrm{y}_{1:T}|\mathrm{c}_{1:T}, \mathrm{m}_{1:T}, \theta)p(\mathrm{c}_{1:T}|\theta)p(\mathrm{m}_{1:T}|\theta). \quad (2)$$

Inference proceeds by using Bayes' theorem to invert the forward model and form the posterior distribution over the modulators and the carriers, given the data,

$$p(\mathrm{c}_{1:T}, \mathrm{m}_{1:T}|\mathrm{y}_{1:T}, \theta) = \frac{p(\mathrm{y}_{1:T}, \mathrm{c}_{1:T}, \mathrm{m}_{1:T}|\theta)}{p(\mathrm{y}_{1:T}|\theta)}. \quad (3)$$

The full solution to PAD is therefore a distribution over possible modulators and carriers, and not a single modulator-carrier pair. This reflects the fact that there is not sufficient information to solve ill-posed problems unambiguously. For practical applications the posterior distribution must be summarised, and one approach is to return the most probable modulator and carrier given the signal,

$$\mathrm{c}^*_{1:T},\ \mathrm{m}^*_{1:T} = \underset{\mathrm{c}_{1:T}, \mathrm{m}_{1:T}}{\arg\max}\ p(\mathrm{c}_{1:T}, \mathrm{m}_{1:T}|\mathrm{y}_{1:T}, \theta), \quad (4)$$

together with error-bars which indicate the uncertainty around this best-estimate. Demodulation therefore reduces to optimisation of a cost-function which specifies how the various constraints trade off with one another in a soft manner (desideratum 1).

The parameters of the model, θ, control how the constraints trade off, and therefore determine the PAD solution. One way to set the parameters is to choose some general purpose values (e.g. [22]), possibly determined by inspecting samples from the forward model. However, mismatch between such modelling assumptions and the signal can lead to undesirable results. In general we would like the model to have a fairly large number of parameters which automatically adjust to the signal (desideratum 2). Fortunately, a number of methods are available to learn the parameters of probabilistic model from a signal alone. Perhaps the simplest is to use the maximum-likelihood (ML) value of the parameters.

$$\theta^{\mathrm{ML}} = \arg\max_{\theta}\ p(\mathrm{y}_{1:T}|\theta), \quad (5)$$

$$= \arg\max_{\theta} \int p(\mathrm{y}_{1:T}, \mathrm{c}_{1:T}, \mathrm{m}_{1:T}|\theta)\, d\mathrm{c}_{1:T}\, d\mathrm{m}_{1:T}. \quad (6)$$

Unfortunately the integral which this demands is often analytically intractable and so numerical approximation methods have to be used. The art is to find accurate, but fast approximations.

If the user has some prior knowledge of the parameters of the model then an alternative to the ML estimate is the *maximum a posteriori* (MAP) estimate. Here, the user's knowledge is incorporated into the prior over parameters, $p(\theta)$ which shapes the solution (desideratum 3),

$$\theta^* = \arg\max_{\theta}\ p(\theta|\mathrm{y}_{1:T}) = \arg\max_{\theta}\ p(\mathrm{y}_{1:T}|\theta)p(\theta), \quad (7)$$

and the ML method is often recovered when the prior over parameters is uniform, $p(\theta) = \textit{constant}$.

The next three sections follow the path described in this section, beginning with a mathematical description of the forward model, then considering inference, and ending with learning.

A. Forward Model

The defining feature of a probabilistic forward model for amplitude modulation is that the signal arises from a product of a slowly varying modulator and a quickly varying carrier (see equation 1). However, as real data are often noisy, the forward model developed here explicitly incorporates additive uncorrelated Gaussian noise around the value of this product, thus improving the noise-robustness of the method. For generality, the Gaussian noise is taken to have non-stationary variance. In other words, given a particular modulator and carrier, the signal is assumed to be a Gaussian-distributed random variable with a mean given by the product of the modulator and carrier, and a time-varying variance denoted by $\sigma^2_{\mathrm{y,t}}$,

$$p(\mathrm{y}_t|\mathrm{m}_t, \mathrm{c}_t, \sigma^2_{\mathrm{y,t}}) = \mathrm{Norm}(\mathrm{y}_t; \mathrm{m}_t\mathrm{c}_t, \sigma^2_{\mathrm{y,t}}). \quad (8)$$

We use the notation $\mathrm{Norm}(x; \mu, \Sigma)$ throughout to indicate a Gaussian or Normal density on the variable x with mean μ and (co)variance Σ.

The prior distribution for the carrier is assumed to be Gaussian and uncorrelated in time, so that a typical sample from the prior would be white noise. Actual carriers in natural sounds will frequently be more structured, as in speech where the carrier may contain pitch and formant information, but the details of this structure are difficult to anticipate in the prior. In practice, the broad spectral assumption tends to separate the inferred time-scales of the carrier and modulator, facilitating accurate inference. To avoid an amplitude scale degeneracy between the carrier and modulator, the carrier scale (or equivalently its variance) is set to unity,

$$p(\mathrm{c}_t) = \mathrm{Norm}(\mathrm{c}_t; 0, 1). \quad (9)$$

The slowly varying modulator process is constrained to be positive in the current approach. It is generated by the application of a pointwise non-linear function to a slowly varying real-valued (positive and negative) function—henceforth called the transformed-modulator (x_t)—drawn from a stationary Gaussian process (GP; see [32] or [33] for an introduction). In the sampled context, this simply means that the distribution over all the transformed-modulator samples ($\mathrm{x}_{1:T}$) taken jointly is a multivariate Gaussian distribution with mean $\mu_{1:T}$ and covariance matrix $\Gamma_{1:T,1:T}$,

$$p(\mathrm{x}_{1:T}|\mu, \Gamma) = \mathrm{Norm}(\mathrm{x}_{1:T}; \mu_{1:T}, \Gamma_{1:T,1:T}). \quad (10)$$

Stationarity requires that the mean of the Gaussian be constant over time, $\mu_t = \mu$, and that the covariance between the transformed-modulator sample at time t and that at time t' be a function of their temporal separation, $\Delta t = |t - t'|$, alone:

$$\Gamma_{t,t'} = \langle \mathrm{x}_t \mathrm{x}_{t'} \rangle - \langle \mathrm{x}_t \rangle \langle \mathrm{x}_{t'} \rangle = \gamma_{|t-t'|} = \gamma_{\Delta t}. \quad (11)$$

For signals from a stationary GP, the covariance (or autocorrelation) function determines the expected spectrum of the signal according to the Wiener-Khintchine theorem. Intuitively, if the covariance falls off quickly with Δt, then the spectrum of the transformed-modulators will contain appreciable higher frequency power, and the signals will tend to vary quickly over time. By contrast, if the autocorrelation falls off slowly, the transformed-modulator will vary relatively slowly.

A convenient choice for the transformed-modulator covariance function is the standard squared-exponential kernel,

$$\gamma_{\Delta t} = \sigma_{\mathrm{x}}^2 \exp\left(-\frac{1}{2\tau_{\mathrm{eff}}^2}\Delta t^2\right). \quad (12)$$

Of the two parameters, τ_{eff} determines how quickly the autocorrelation falls off and therefore fixes the time-scale of variation of a typical sample drawn from the GP, whilst $\sigma_{\mathrm{x}}^2 = \gamma_0$ determines its amplitude.

The non-negative modulator signal is derived deterministically from the transformed-modulator by a 'soft threshold-linear' function:

$$\mathrm{m}_t = \mathrm{m}(\mathrm{x}_t) = \sigma_{\mathrm{m}} \log(1 + \exp(\mathrm{x}_t)). \quad (13)$$

This non-linearity is dominated by its exponential part for large negative values of x, yielding small modulator signals m. For large positive values of x the mapping is approximately linear. Thus, the Gaussian marginal distribution of the transformed-modulators is modified into a sparse distribution over the modulator variables. A sparse distribution is often a good match to the modulator histogram derived from natural sounds [34].

Equations 8-13 define a specific subclass of possible models—and therefore possible algorithms—for probabilistic amplitude demodulation. We refer to algorithms derived from this subclass as Gaussian Process PAD (GP-PAD).

B. Inference

Full distributional inference is intractable in GP-PAD because of the two non-linearities of the forward model (equations 8 and 13), and so some form of approximation is necessary. The use of the joint mode (equation 4) is one option; however, the structure of GP-PAD allows for a slightly more sophisticated approach. The Gaussian prior on the carrier makes it possible to integrate over the unknown carrier values, leaving a marginal probability on the transformed-modulator alone. The inferred modulator signal is then derived from the mode of this marginal probability:

$$\begin{aligned}\mathrm{x}^*_{1:T} &= \arg\max_{\mathrm{x}_{1:T}} p(\mathrm{x}_{1:T}|\mathrm{y}_{1:T}, \theta), \quad (14)\\ &= \arg\max_{\mathrm{x}_{1:T}} \log p(\mathrm{y}_{1:T}, \mathrm{x}_{1:T}|\theta) = \arg\max_{\mathrm{x}_{1:T}} \mathcal{L}(\mathrm{x}_{1:T}),\end{aligned}$$

where the final equation defines the objective function $\mathcal{L}$. Experiments indicate that this approach is both faster and more robust to over-fitting.

There is no closed-form solution for this optimum, but a gradient-based method can be used to find a local maximum. The objective-function and its gradients can be computed efficiently as follows. Note first that the objective splits into two terms: one derived from the likelihood and one from the prior,

$$\mathcal{L}(\mathrm{x}_{1:T}) = \sum_{t=1}^{T} \log p(\mathrm{y}_t|\mathrm{x}_t, \theta) + \log p(\mathrm{x}_{1:T}|\theta).$$

The likelihood term is simple and fast to compute as the probability of the signal given the transformed-modulator at that time-step is a zero mean Gaussian distribution with a variance determined by the modulator, $p(\mathrm{y}_t|\mathrm{x}_t, \theta) = \mathrm{Norm}(\mathrm{y}_t; 0, \mathrm{m}^2(\mathrm{x}_t) + \sigma^2_{\mathrm{y,t}})$. The component from the prior is more challenging as it involves inverting the $T \times T$ covariance matrix of the GP which is intractable for time-series of even modest length ($T > 1000$).

One way around this obstacle is to introduce a new set of unobserved variables, $\mathrm{x}_{T+1:T'}$ where $T' > T$, such that the augmented set of variables $\mathrm{x}_{1:T'}$ is circularly correlated (so, for example, x_1 and $\mathrm{x}_{T'}$ are neighbours). This places the augmented latent variables on a ring and the new covariance matrix, $\Gamma_{1:T',1:T'}$, is circulant, $\Gamma_{t,t'} = \Gamma_{\mathrm{mod}(t-t',T')}$. The objective can now be computed efficiently using the Fast Fourier Transform (FFT):

$$\begin{aligned}\mathcal{L}(\mathrm{x}_{1:T'}) = c &- \frac{1}{2}\sum_{t=1}^{T} \log(\mathrm{m}_t^2 + \sigma^2_{\mathrm{y,t}})\\ &- \frac{1}{2}\sum_{t=1}^{T} \frac{y_t^2}{\mathrm{m}_t^2 + \sigma^2_{\mathrm{y,t}}} - \frac{1}{2T'}\sum_{k=1}^{T'} \frac{|\Delta\tilde{\mathrm{x}}_k|^2}{\tilde{\gamma}_k}, \quad (15)\end{aligned}$$

where $\Delta\tilde{\mathrm{x}}_k$ is the Discrete Fourier Transform (DFT) of the mean-shifted transformed-modulators $\Delta\mathrm{x}_t = \mathrm{x}_t - \mu$, and $\tilde{\gamma}_k$ is the DFT of the covariance function, which is the spectrum of the Gaussian Process. The derivatives can be computed using the expressions above and are omitted for brevity (see [34] for the details). The conjugate gradient method can be used for optimisation [35]. There is some freedom in setting T', but two useful rules of thumb are that it should be a power of two to accelerate the FFT, and it should be larger than T by several times the decay-time of the modulator covariance function so as to avoid wrap-around artifacts that might otherwise arise by the introduction of correlation between the beginning and end of the signal. Alternatively, these potential wrap-around artifacts might be avoided by zero-padding the signal.

One of the advantages of probabilistic demodulation is that it becomes possible to estimate the uncertainty in the recovered modulators. However, although the uncertainty is mathematically well-defined, computational tractability remains an issue. Here, we employ an approximate version of Laplace's method, itself an approximation, to compute it. Laplace's method [31] approximates the posterior distribution over the transformed-modulators by a Gaussian centred at the posterior mode, with a covariance matrix given by the negative inverse of the Hessian matrix, H, of the log-joint (given in equation 15),

$$\begin{aligned}p(\mathrm{x}_{1:T'}|\mathrm{y}_{1:T}, \theta) &\approx p(\mathrm{y}_{1:T}, \mathrm{x}^*_{1:T'}|\theta)\\ &\times \exp\left(\frac{1}{2}(\mathrm{x}_{1:T'} - \mathrm{x}^*_{1:T'})^{\mathsf{T}}\mathrm{H}(\mathrm{x}_{1:T'} - \mathrm{x}^*_{1:T'})\right), \quad (16)\end{aligned}$$

where,

$$\mathrm{H}_{t,t'} = \left.\frac{d^2}{d\mathrm{x}_t d\mathrm{x}_{t'}} \log p(\mathrm{y}_{1:T}, \mathrm{x}_{1:T'}|\theta)\right|_{\mathrm{x}_{1:T'} = \mathrm{x}^*_{1:T'}}. \quad (17)$$

Thus the estimated posterior uncertainty is $\Sigma^{\mathrm{post}} = -\mathrm{H}^{-1}$. Unfortunately, the Hessian is a $T' \times T'$ matrix and so inversion is typically intractable. A further approximation exploits the simple structure of the Hessian: H is the sum of a diagonal

term from the likelihood (D with $\mathrm{D}_{tt} = \frac{d^2}{dx_t^2} \log p(\mathrm{y}_t|\mathrm{x}_t)$), and the inverse prior covariance matrix,

$$\mathrm{H}^{-1} = (\mathrm{D} + \Gamma^{-1})^{-1} = \Gamma^{1/2}(\Gamma^{1/2}\mathrm{D}\Gamma^{1/2} + I)^{-1}\Gamma^{1/2}. \quad (18)$$

In this form the difficult inversion is limited to the matrix $A = \Gamma^{1/2}\mathrm{D}\Gamma^{1/2}$, the other terms being simple to compute exactly. The matrix A inherits the concentrated eigenspectrum of the prior covariance Γ (recall that, after augmentation, both are circulant matrices whose eigenvalues correspond to the Fourier coefficients of any row). Consequently, A can be well approximated by a truncated eigenexpansion, $A \approx \sum_{k=1}^{K_{\mathrm{MAX}}} \lambda_k \mathbf{e}_k \mathbf{e}_k^T$, and the problem reduces to finding an efficient method to compute the top K_{MAX} eigenvectors, $\mathbf{e}_k$, and eigenvalues, λ_k, of A. The Lanczos algorithm provides one solution, requiring only multiplication of A by a vector [36], and these products can be computed rapidly using the FFT. The eigenvalues and vectors then approximate the posterior covariance,

$$\Sigma^{\mathrm{post}} \approx \Gamma^{1/2}\left(I - \sum_{k=1}^{K_{\mathrm{MAX}}} \frac{\lambda_k}{\lambda_k + 1}\mathbf{e}_k\mathbf{e}_k^{\mathsf{T}}\right)\Gamma^{1/2}. \quad (19)$$

This expression has an instructive interpretation: in order to compute the approximate posterior covariance (the posterior uncertainties), begin with the marginal covariance of the prior (the prior uncertainties) and subtract uncertainty corresponding to each eigenmode of the likelihood. In practice, the most useful elements of uncertainty are given by the marginal posterior variances (the diagonal entries in Σ^{post}) and memory limitations will often restrict calculations to only these values.

C. Learning

The parameters of the GP-PAD forward model reflect assumptions about the statistics of the signal such as the time-scale of variation in the modulator or its degree of sparsity; and the quality of modulator estimates depends on the match between these parameters and the true signal properties. Although the non-linear model complicates the manual specification of these parameters, the Bayesian approach makes it possible to determine appropriate values automatically.

To avoid over-fitting, parameter learning must respect the uncertainty in the values of the latent variables [34]. Above, this uncertainty was approximated by Laplace's method (equation 16); this Gaussian approximation is conveniently integrated with respect to the transformed-modulators, to yield a marginal likelihood function for the parameters alone:

$$p(\mathrm{y}_{1:T}|\theta) = \int p(\mathrm{y}_{1:T}, \mathrm{x}_{1:T'}|\theta) d\mathrm{x}_{1:T'}, \quad (20)$$

$$\approx p(\mathrm{y}_{1:T}, \mathrm{x}^*_{1:T'}|\theta) \frac{(2\pi)^{T'/2}}{\sqrt{\det(-\mathrm{H})}}. \quad (21)$$

This (approximate) marginal likelihood provides an objective function for learning the parameters. In practice, the modulator time-scale (τ_{eff}) is learned in this way.

Laplace's approximation is still computationally costly in realistic settings, and so a more efficient alternative is used to learn values of the remaining parameters (σ_{x}^2, σ_{m} and μ). This approach is motivated by observing that these parameters are well-constrained by the marginal distributions of each sample, which do not depend on the temporal structure:

$$\arg\max_{\sigma_{\mathrm{m}},\mu,\sigma_{\mathrm{x}}^2} p(\mathrm{y}_{1:T}|\sigma_{\mathrm{m}}^2, \mu, \sigma_{\mathrm{x}}^2) \approx \arg\max_{\sigma_{\mathrm{m}},\mu,\sigma_{\mathrm{x}}^2} \prod_t p(\mathrm{y}_t|\sigma_{\mathrm{m}}^2, \mu, \sigma_{\mathrm{x}}^2).$$

The marginal integrals are easy to evaluate numerically (e.g. by gridding the region of significant probability under the parameter priors). This leads to a two stage scheme in which the marginal distribution of the signal is used to learn σ_{x}^2, σ_{m} and μ, and then Laplace's approximation to the likelihood is used to learn τ_{eff}. Any available prior information about the parameters can be incorporated into either of these stages (equation 7). Using our implementation of GP-PAD, the first stage of learning typically takes a few seconds on a standard laptop, independent of the signal duration. The second stage of learning typically takes a few minutes for a second of sound at 16000Hz.

A matlab implementation of GP-PAD can be obtained from the authors' website (www.gatsby.ucl.ac.uk/resources/pad).

D. Relationship to existing methods

Before demonstrating PAD practically, we note several connections to existing demodulation methods. First, the GP-PAD assumptions imply that the expected value of the square of a noiseless ($\sigma_{\mathrm{y,t}}^2 = 0$) signal, conditioned on the modulator values, is equal to the square of the modulator: $\langle \mathrm{y}_t^2 | \mathrm{m}_t \rangle = \mathrm{m}_t^2$. Thus, in principle, the square of the signal provides an unbiased estimator for the (squared) modulator; but it has very large variance (equal to $2\mathrm{m}_t^4$). This variance can be reduced by exploiting the slow variation of the modulator and averaging the squared signal over a local region, although this reduction in variance comes at the cost of introducing some bias because the true modulator may vary slightly over the local region. This leads precisely to the SLP method described earlier. From this new perspective, choosing the low-pass filter cut-off in SLP amounts to selecting a point on a bias-variance tradeoff. Motivated by this connection, the SLP method is used to intialise the gradient optimisation for GP-PAD. It is worth reiterating, however, that although the SLP method often provides a reasonable estimate of the modulator (in a squared-error sense), the corresponding estimate for the carrier is often extremely inaccurate and so fine-tuning of the SLP solution is essential for many applications.

In a recent paper, Sell and Slaney introduce an elegant approach to demodulation [22] that defines the modulator as the solution to a convex optimisation problem which can be written,

$$\text{minimise} \sum_f \left(W_f \mathcal{F}_f(\mathrm{m}(t))\right)^2 \text{ subject to } \mathrm{m}(t) \geq |\mathrm{y}(t)|.$$

Here, $\mathcal{F}_f$ is the Fourier coefficient at freqency f and W_f is a window function used to penalise high-frequency energy in the modulator. The basic idea is that the constraints ensure the modulator is greater than the absolute value of the signal at each time-point, and the cost-function ensures the solution goes near to the rectified signal, in a slowly varying and smooth manner. This sensible scheme was motivated heuristically, but here we show that it can also be derived from a

probabilistic model. Consider a forward model for amplitude demodulation in which a positive modulator ($\mathrm{m}_t \geq 0$) is drawn from a truncated multivariate Gaussian and where the carrier is drawn from a uniform distribution on the range $[-1, 1]$,

$$p(\mathrm{m}_{1:T}|\Sigma_{1:T,1:T}) = \frac{1}{Z}\exp\left(-\frac{1}{2}\mathrm{m}_{1:T}^{\intercal}\Sigma_{1:T,1:T}^{-1}\mathrm{m}_{1:T}\right), \quad (22)$$

$$p(\mathrm{c}_t) = \mathrm{Uniform}(\mathrm{c}_t; -1, 1), \quad (23)$$

$$\mathrm{y}_t = \mathrm{m}_t\mathrm{c}_t. \quad (24)$$

The fact that the carriers in this model are bounded between $-1 \leq \mathrm{c}_t \leq 1$ can be motivated from a sinusoidal model, $\mathrm{c}_t = \sin(\phi_t)$.

The prior over the carriers enforces the constraint that $|\mathrm{c}_t| \leq 1$. The likelihood enforces the constraint that, $\mathrm{c}_t = \mathrm{y}_t/\mathrm{m}_t$. Both of these constraints are only satisfied when the modulator is greater than or equal to the data magnitude, $\mathrm{m}_t \geq |\mathrm{y}_t|$. The posterior distribution over modulators is therefore another truncated Gaussian where the constraints define the new truncation points. The MAP modulator is then given by,

$$\mathrm{m}_{1:T}^{*} = \arg\min_{\mathrm{m}_{1:T}} \mathrm{C}(\mathrm{m}_{1:T}) \quad \text{such that} \quad \mathrm{m}_t \geq |\mathrm{y}_t|, \quad (25)$$

where the cost function is the negative of the log prior probability of the modulators,

$$\mathrm{C}(\mathrm{m}_{1:T}) = \frac{1}{2}\mathrm{m}_{1:T}^{\intercal}\Sigma_{1:T,1:T}^{-1}\mathrm{m}_{1:T} \approx \frac{1}{2}\sum_{k=1}^{T}\frac{|\tilde{\mathrm{m}}_k|^2}{\tilde{\gamma}_k}. \quad (26)$$

This is the same cost function as used in Sell and Slaney's 'linear' demodulation algorithm, revealing the connection between their approach and the probabilistic one. This connection is valuable as it suggests methods to determine the free parameters of the convex approach such as the spectral weighting function, W_f (equivalent to the covariances $\tilde{\gamma}_k^{-1/2}$).

One contribution that emerges from previous research is a catalogue of properties that are desirable in a demodulation algorithm. Many properties identical or similar to these arise naturally when the rules of probability are used to invert the PAD generative model. For instance, the carrier and the modulator recovered by GP-PAD from a bounded signal will also be bounded [21] because the prior probability of an unbounded carrier or modulator vanishes. Similarly, the modulator will be smooth [26] because a realisation from a GP prior with a squared exponential kernel is (almost surely, almost everywhere) analytic [32]. PAD is also covariant with respect to the scale of the input signal (a generalisation of a constraint in [26]) because the ML modulator variance rescales to compensate for any change in the signal scale.

IV. Results

Here, we apply GP-PAD to various signals, demonstrating its compliance with the remaining desiderata and comparing it to methods which fail them, like the SLP and HE methods. One of the main challenges posed by the evaluation of demodulation algorithms on natural signals is that the ground truth is unknown. This means that a quantitative comparison of different schemes must take an indirect approach. We present several different comparisons of this sort. The first uses synthetic signals, for which the ground truth carriers and modulators are known (see section IV-A). The results suggest that PAD is more flexible than other methods. In particular, GP-PAD performs well even when the modulator and carrier bands overlap, which is often the case in natural signals, and also when the signal is stochastic. Although these results are suggestive, they cannot be seen as conclusive without knowing which synthetic signal class is a sensible approximation to natural sounds. In section IV-B GP-PAD is applied to speech and the estimated carriers and modulators are shown to be qualitatively superior to those recovered by other methods. The solutions are evaluated for consistency, for example by demodulating the carrier to test for the Carrier Identity property. Another test of consistency, related to robustness, estimates the modulators from noisy signals and measures how close these come to inferences based on the clean signal. Similarly, modulators estimated in missing-data regions are compared to the values obtained from the complete signal. These consistency tests are important criteria that a demodulation algorithm should meet, but are not sufficient to guarantee a good algorithm. For example, an algorithm which returned a constant modulator, independent of the signal, would pass the tests above, but would evidently not qualify as a good demodulation algorithm.

Finally, in section IV-C, PAD is applied to the sub-bands of a signal showing that the carrier remains reasonably band-limited, in contrast to those returned by many existing methods. This is critical for reconstruction [18].

A. Synthetic signals

The experiments described in this section test PAD on synthetic signals in three different settings; noise-free data, noisy data and missing data. In all of the experiments, the synthetic signal comprised a modulated carrier, possibly combined with additive Gaussian noise, $\mathrm{y}_t = \mathrm{m}_t\mathrm{c}_t + \sigma_{\mathrm{y}_t}\epsilon_t$. Three different carriers were used

1) A sinusoidal carrier, $\mathrm{c}_t^{(1)} = \sin(2\pi f^{(c)}t)$ where $f^{(c)} = 100.7$Hz.
2) A harmonic carrier, $\mathrm{c}_t^{(2)} = \sin(2\pi f_1^{(c)}t) + \sin(2\pi f_2^{(c)}t)$ where $f_1^{(c)} = 100.7$Hz and $f_2^{(c)} = 201.4$Hz
3) A white noise carrier, $\mathrm{c}_t^{(3)} \sim \mathrm{Norm}(0, 1)$

The properties of the carrier were generally found to have a more substantial effect on the performance of the demodulation algorithms than the properties of the envelope. For this reason the same envelope was used throughout all of the experiments shown; an exponentiated sum of three sinusoids,

$$\mathrm{m}_t = \exp\left(\sum_{k=1}^{3}\left(\alpha_k^{(1)}\sin(2\pi f_k^{(m)}t) + \alpha_k^{(2)}\cos(2\pi f_k^{(m)}t)\right)\right).$$

The coefficients of the sinusoids were drawn from a unit variance Gaussian $\alpha_k^{(i)} \sim \mathrm{Norm}(0, 1)$ and the frequencies of the sinusoids uniformly sampled between $0 - 2$Hz, $f_k^{(m)} \sim \mathrm{Uniform}(0, 2)$. The results reported below are robust to the seed of the pseudo-random number generator. A sample rate of 2000Hz was used for the experiments.

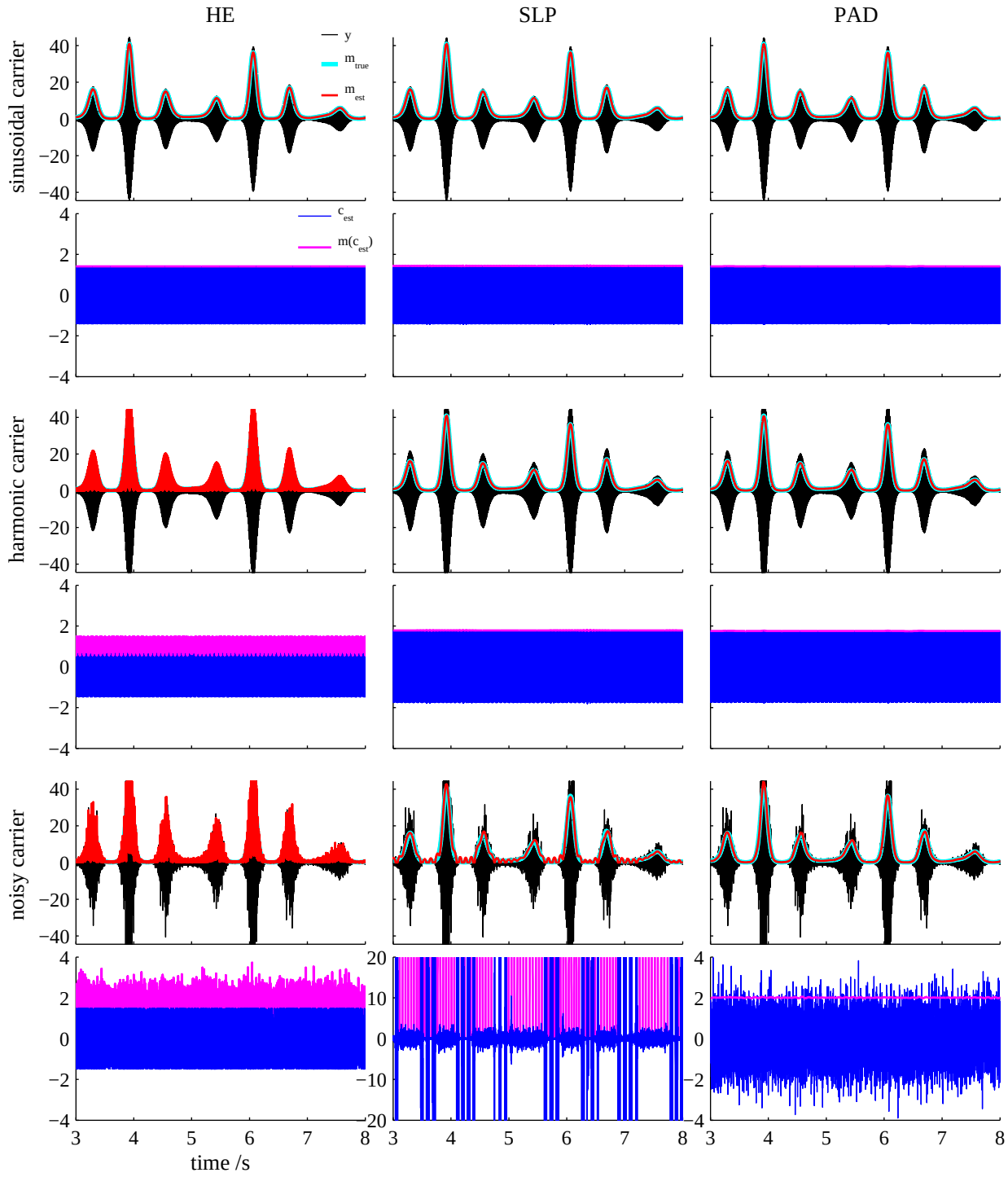

Fig. 2. Demodulating synthetic signals. Each pair of rows shows a different signal (y, black) decomposed into a modulator (m_{est}, thick line) and a carrier (c_{est}, lower panel) using three different methods; the Hilbert Envelope (column 1), the SLP method (column 2) and PAD (column 3). The true modulator is also shown for reference (m_{true}). The result of demodulating the carriers (an empirical test of the Carrier Identity property) is also shown ($\mathrm{m}(\mathrm{c}_{\text{est}})$).

For reference, PAD is compared to the HE and SLP methods below. One of the metrics used for comparison is the Signal to Noise Ratio (SNR) between the true envelopes (m_t) and the estimated envelopes ($\hat{\mathrm{m}}_t$),

$$\mathrm{SNR}_{\mathrm{m}} = 10\log_{10}\sum_{t=1}^{T}\mathrm{m}_t^2 - 10\log_{10}\sum_{t=1}^{T}(\mathrm{m}_t - \hat{\mathrm{m}}_t)^2 \quad (27)$$

A SNR can be defined analogously for the carriers.

Whilst the HE has no free parameters, and can therefore be applied directly to any signal, the SLP method must be adapted to the signal. We chose a low-pass filter with a logistic shape, $1/(1+\exp((f - f_{\text{cut-off}})/f_{\text{width}})$ and set the cut-off ($f_{\text{cut-off}}$) and width (f_{width}) to return the largest SNR for the estimated envelopes. This is an upper-bound on the

performance of the SLP method using the logistic filter shape because it exploits the ground-truth to optimise the parameters. One further correction is made to the usual SLP algorithm to address the fact that it can return an estimate for the square-envelope that is negative, as there is no positivity constraint on the output of the low-pass filter. We therefore thresholded the filter output, setting values that fell below 10^{-4} to 10^{-4}.

In order to make the tests as demanding as possible, all of the parameters in PAD were learned from the signal using maximum-likelihood; i.e., no prior knowledge was used. For the noisy signals, the noise-level was assumed to be known before hand and the parameters were set to those learned from the clean signal.

Some typical results for the three algorithms demodulating three different clean test signals are shown in Fig. 2 and the results on noisy signals are summarised in Fig. 3.

The HE is the best estimator when the carrier is a pure tone, although all three methods typically have high SNRs (~ 40dB) for signals of this type. However, when the carrier is more complex, the HE often contains components that are faster than desired. For example, when carriers contain harmonics the HE contains a component at the fundamental, and when the carriers are stochastic the HE becomes very noisy. For similar reasons, the HE degrades quickly as more noise is added to the signal.

The SLP envelope is more robust than the HE. It is often accurate (as measured by the SNR of the estimated envelopes) so long as it is known where to place the filter cut-off. For this reason, in normal applications it tends to perform well when the carrier and modulator bands are well separated, but less so when there is overlap. Here the oracular determination of the threshold overestimates SLP performance. However, whilst the estimate of the envelope is often accurate in a squared-error sense, the method can return very poor estimates for the carrier (e.g. see the noisy carrier in Fig. 2). This happens because in regions of low energy, the SLP envelope may become too small resulting in carrier estimates which are very large. As stated above, there is no constraint built into the method to ensure that the carrier remains well behaved.

GP-PAD out-performs the SLP method in every condition and the HE in every condition bar the simplest (a pure tone carrier in noise free conditions), both in terms of the SNR of the envelopes and the SNR of the carriers (data not shown). PAD approximately satisfies the Carrier Identity test for all of the synthetic signals, whilst the other methods do not (as shown in Fig. 2). The conclusion is that PAD is more robust, both to changes in the signal class and to additive Gaussian noise.

Finally, Fig. 4 demonstrates that PAD can be used to accurately fill in the envelopes in missing regions of the synthetic signals. Practically, the inference proceeds as for complete data, but the variance of the observation noise is set to infinity in the missing region ($\sigma^2_{\mathrm{y,t}} = \infty$). The estimated envelopes are very accurate for small gap sizes, but deteriorate for gaps comparable to the time-scale of the envelope. The uncertainty in the estimated envelopes grows correspondingly with the gap size. Importantly, the true envelope tends to remain within the error-bars at all gap sizes. The conclusion is that PAD can accurately estimate both the envelopes and the uncertainty in regions of missing data.

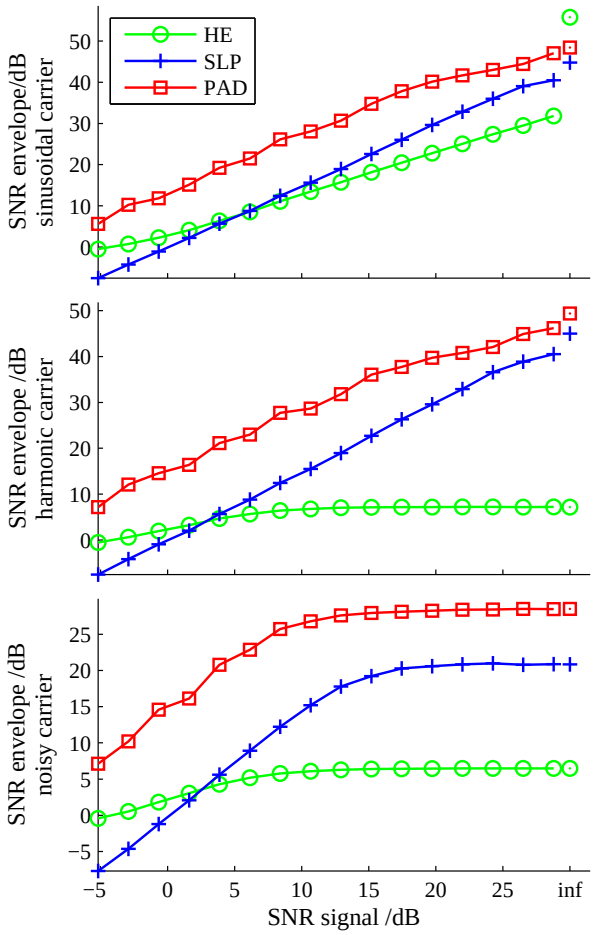

Fig. 3. Demodulating noisy synthetic signals. Noise was added to the signals shown in Fig. 2 and the envelopes estimated using the three methods. The panels show the SNR of the estimated envelopes as a function of the SNR of the signal. The three methods are; HE (circles), SLP (crosses), PAD (squares). PAD out-performs the other methods on the noisy data by about 3-12dB. The rightmost markers show the performance on the clean signal for reference.

B. Speech signals

This section applies PAD to a speech sound. Noise-free data, noisy-data and missing-data settings are considered. There is an important difference between the synthetic signals considered in the previous section and natural sounds like speech treated here. Whereas the synthetic sounds contained a single time-scale of modulation, natural sounds often contain modulation at multiple time-scales (see Fig. 5). Fortunately, PAD can be used to automatically diagnose when this is the case and to select between the various solutions. The key quantity in this process is the (approximate) likelihood of the time-scale, $p(\mathrm{y}|\tau_{\mathrm{eff}})$. This is found to have a single peak for the synthetic sounds considered in the last section indicating a single best solution. However, this quantity has three peaks for the speech sound considered here (see Fig. 5). Each of these peaks corresponds to modulation arising from a different physical process; the glottal pulse periods, the syllables, and

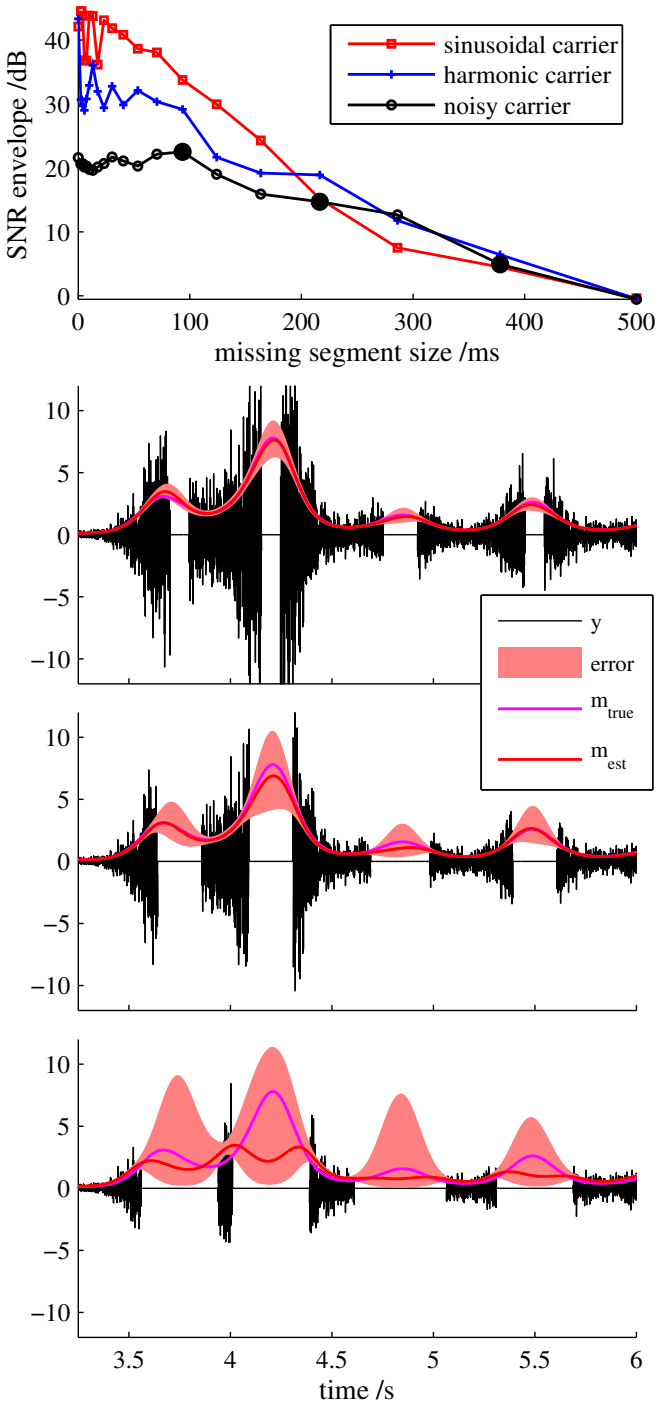

Fig. 4. Demodulating synthetic signals with missing data. Short sections of the three synthetic signals were removed from randomly chosen locations. PAD was used to fill in the missing regions and the SNR of the estimated envelopes in the missing regions was plotted as a function of gap size (top panel) for the three signals: with sinusoidal carrier (squares), harmonic carrier (crosses) and noisy carrier (circles). The larger filled black circles correspond to the examples plotted in the lower panels for the noisy carrier signal. The lower panels show a short section of the signal (y, thin black line), the true envelopes and the estimated envelopes (thick lines denoted $\mathrm{m}_{\mathrm{true}}$ and $\mathrm{m}_{\mathrm{est}}$ respectively) and the error-bars at three standard deviations (shaded region).

the sentences of speech. By appropriately choosing the prior (see equation 7), the user is able to select between these

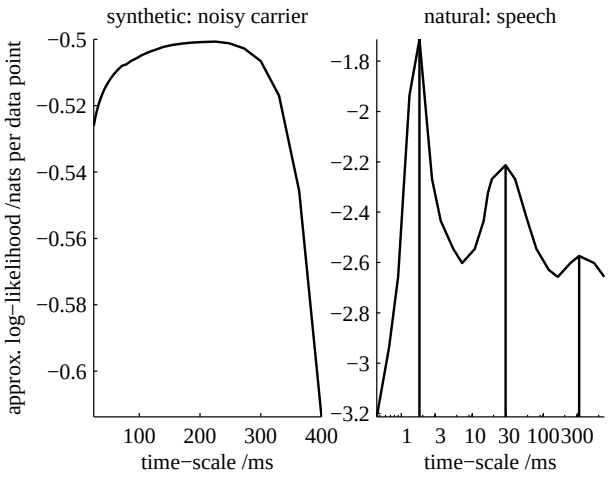

Fig. 5. Approximate likelihoods of the time-scale parameter, τ_{eff}. The left panel shows the log-likelihood for the noisy-carrier synthetic signal which is unimodal. The right panel shows the log-likelihood for the speech signal which has three modes originating from the pitch ($\tau_{\mathrm{eff}}^{-1} = 550$Hz), syllable ($\tau_{\mathrm{eff}}^{-1} = 34$Hz) and sentence ($\tau_{\mathrm{eff}}^{-1} = 3$Hz) structures in speech. These frequencies are higher than normally associated with these structures because each cycle corresponds to two or three time-scales.

different solutions. For example, a prior which favours the syllable time-scales is used in Fig. 6. The speech is demodulated effectively and the Carrier Identity property holds to a close approximation. For comparison, when the parameters of the SLP method are chosen to demodulate at this time-scale a reasonable looking modulator can be returned, but the carriers are typically ill-behaved. Similarly, when a prior is used which favours the sentence time-scales, PAD again demodulates the sound fairly effectively (see Fig. 7), especially in comparison with the result from SLP. Finally, when a prior is used which favours the glottal pulse time-scales, PAD recovers a modulator which resembles the solution provided by the HE. The HE method demodulates voiced phonemes at the time-scale of the vocal fold oscillations because this causes harmonic structure, but in unvoiced sections it becomes noisy.

Next we consider the performance of PAD on signals which are noisy and contain missing data. As ground truth is unknown we compare the envelope estimates derived from the noisy signals to those derived from the clean signals. This is an important consistency test. Fig. 8 indicates that the solution from PAD degrades less quickly in the presence of noise than that from the SLP or HE methods. Fig. 9 shows that PAD can reliably estimate the envelope of speech signals in missing regions up to 20ms long.

C. Sub-band demodulation

One of the more frequent applications of demodulation in the context of natural sounds is to the sub-bands of the signal. However, it is known that the carriers derived from band-limited signals using many demodulation methods, such as the SLP and HE methods, are not guaranteed to be spectrally limited to the pass-band of the filter (see Fig. 10 for one example). This can lead to artifacts, for example when reconstructing signals by recombining filtered versions

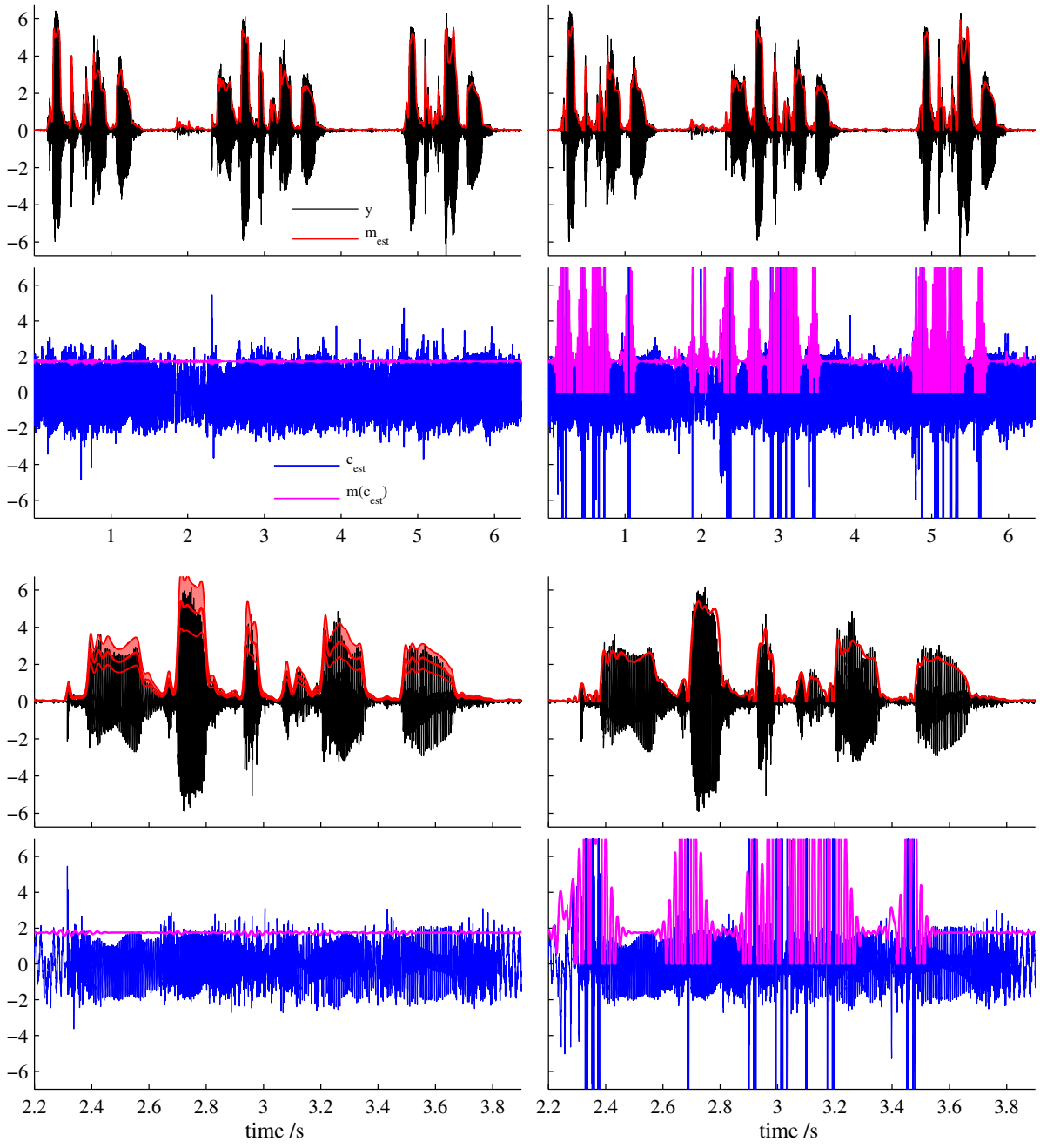

Fig. 6. Demodulating the syllables of speech. The top pair of rows shows a speech signal consisting of three spoken sentences (y, thin black line, upper panel) decomposed into a modulator (m_{est}, thick positive valued line, upper panel) and a carrier (c_{est} thin line, lower panel). The left hand column shows PAD and the right hand column the SLP method. The time-scale used for PAD was learned from the data using a flat prior, and the SLP cut-off was set to give similar results to PAD. The result of demodulating the carriers (an empirical test of the Carrier Identity property) is also shown ($m(c_{est})$, thick positive line, lower panel). The bottom pair of panels is a close up of the middle sentence. The three standard deviation error-bars on the PAD envelopes are also indicated by the shaded region.

of the sub-band modulators with the original carriers [17]. Fig. 10 demonstrates that the carriers derived from GP-PAD are substantially more band-limited than those derived from the HE or SLP methods. However, as there is no constraint on the carrier frequency content, carrier signal energy can still be found outside of the filter. In principle, PAD could be extended to add such a constraint to the spectral content of the carrier, which may improve performance in this application.

An alternative approach to sub-band demodulation, that suggests a different extension to PAD, is that of coherent demodulation [27]. Like the HE, this method assumes the carrier is a single frequency-modulated sinusoid, but it is constrained to be limited to the pass-band of the filter. The method performs well when the filters are narrow, but it can perform poorly for broad filters that contain harmonic or noisy carriers which violate the assumptions of the algorithm.

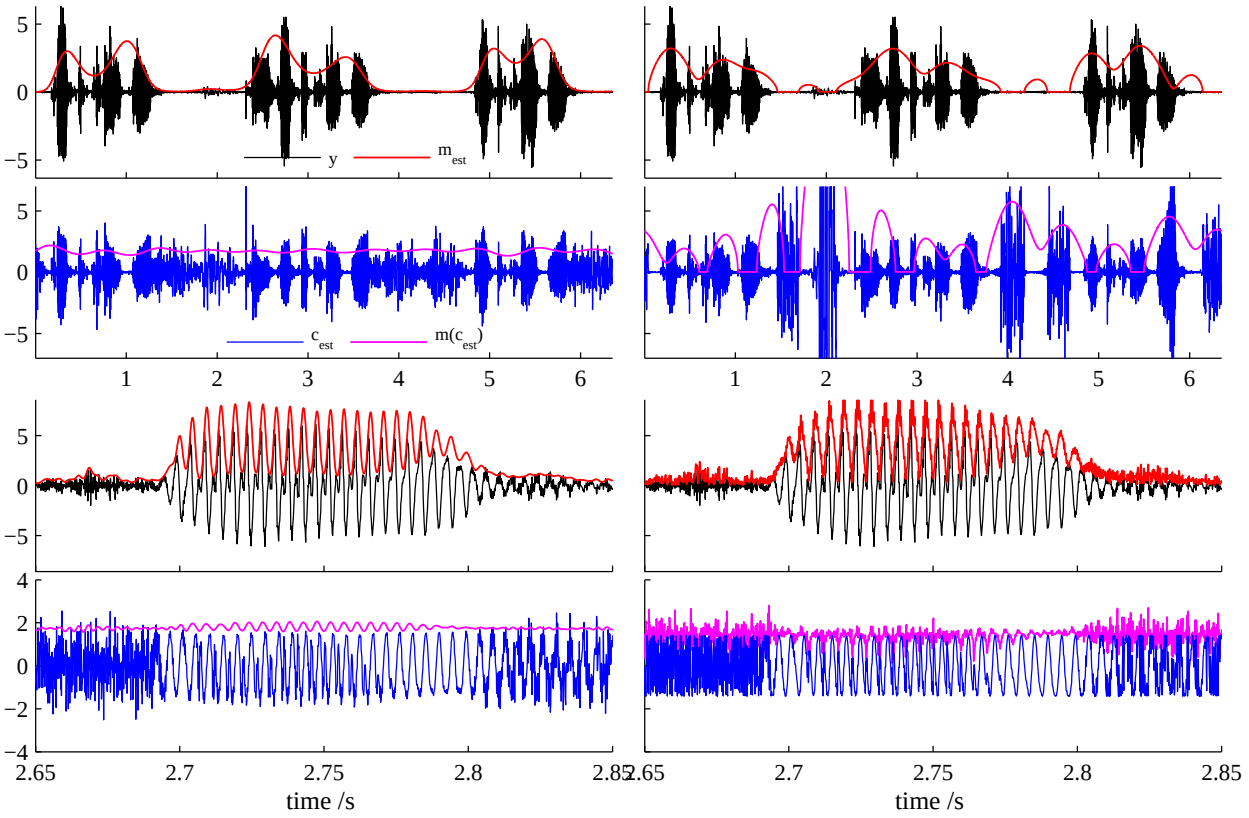

Fig. 7. Demodulating speech at sentence time-scales and glottal-pulse time-scales. The top pair of rows shows a speech signal consisting of three spoken sentences (y, black, upper panel) decomposed into a modulator (m_{est}, thick positive line, upper panel) and a carrier (c_{est}, thin line, lower panel). The left hand panels show PAD and the right hand panels the SLP method. The time-scale used for PAD was learned from the data using a prior favouring long time-scales, and the SLP cut-off was set to give similar results to PAD. The result of demodulating the carriers (an empirical test of the Carrier Identity property) is also shown ($m(c_{est})$, thick line, lower panel). The bottom pair of panels shows the first syllable of the middle sentence. Here the left hand panels show PAD with a time-scale learned from the data using a prior favouring short time-scales. The right hand panels show the results from using the HE.

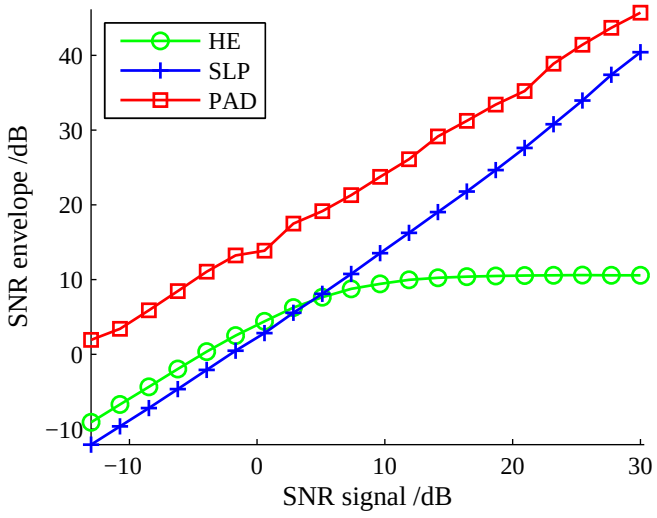

Fig. 8. Demodulating a noisy speech signal. Noise was added to the speech signal shown in Fig. 6 and the envelopes were estimated using the three methods. The panels show the SNR of the estimated envelopes as a function of the SNR of the signal. The three methods are; HE (circles), SLP (crosses), PAD (squares). PAD out-performs the other methods on the noisy data by about 5-11dB.

Furthermore, there is no positivity constraint on the envelope in this approach and so it is not directly comparable to the approach taken here. In fact, the envelope is completely unconstrained in coherent demodulation (although it may inherit bandwidth constraints from a band-limited input signal and a band-limited carrier estimate). It is interesting to contrast this to PAD where it is the carrier which is unconstrained (beyond having zero mean and unit variance). It appears that a method which imposes direct constraints on both the carrier and envelope would combine the benefits of both approaches. For instance, one approach in this direction would be to extend PAD so that the carrier is modelled as a frequency modulated sinusoid, and to place a prior over the frequency modulation.

One potential application of a version of probabilistic subband demodulation is to missing signal reconstruction. In fact, the approach developed for reconstructing the envelopes in PAD (see sections IV-A and IV-B) can be used for this purpose if the carriers in each subband are reconstructed along with the modulators. Two possible schemes are to either approximate the carriers as fixed frequency sinusoids or as Gaussian noise, which has been passed though the filter bank. This reconstruction approach, which could be improved by explicitly modelling the carrier, is a probabilistic version of that used by Clark and Atlas [24].

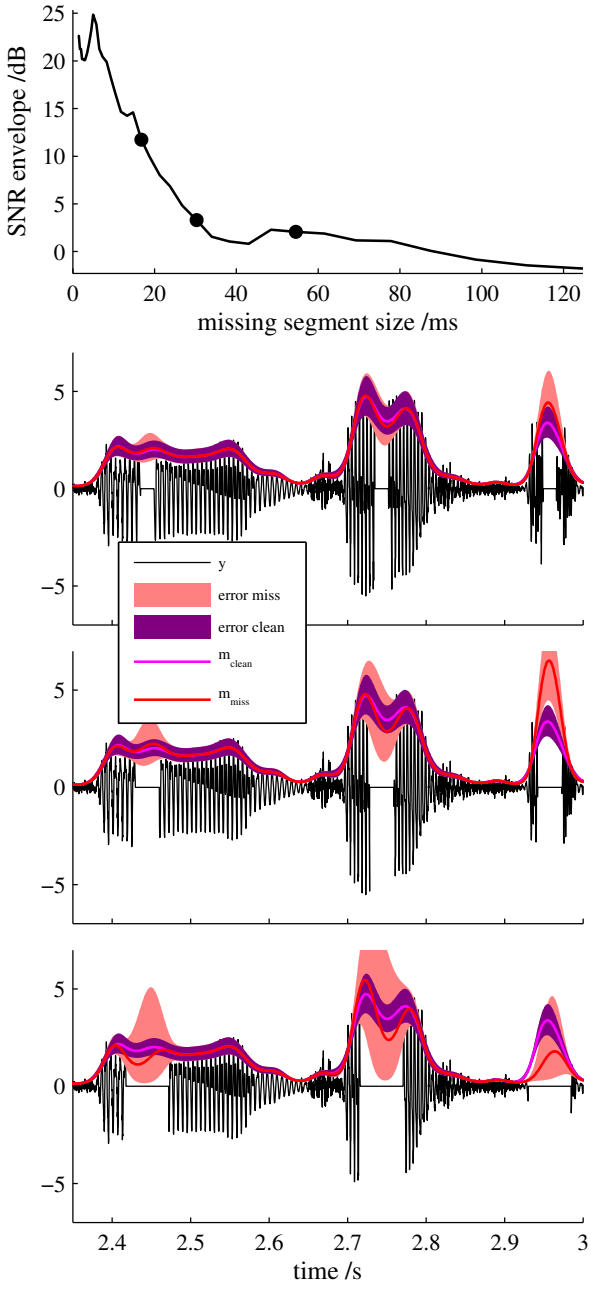

Fig. 9. Demodulating a speech signal with missing data. Short sections of the speech signal were removed from randomly chosen locations. PAD was used to fill in the missing regions and the SNR of the estimated envelopes in the missing regions was plotted as a function of gap size (top panel). The open black circles correspond to the examples plotted in the lower panels. The lower panels show short sections of the signals (y, black), the envelopes estimated from the clean signal ($\mathrm{m}_{\mathrm{clean}}$) with three standard deviation error-bars (dark shaded region), and the estimated envelopes ($\mathrm{m}_{\mathrm{est}}$) with error-bars (light shaded region).

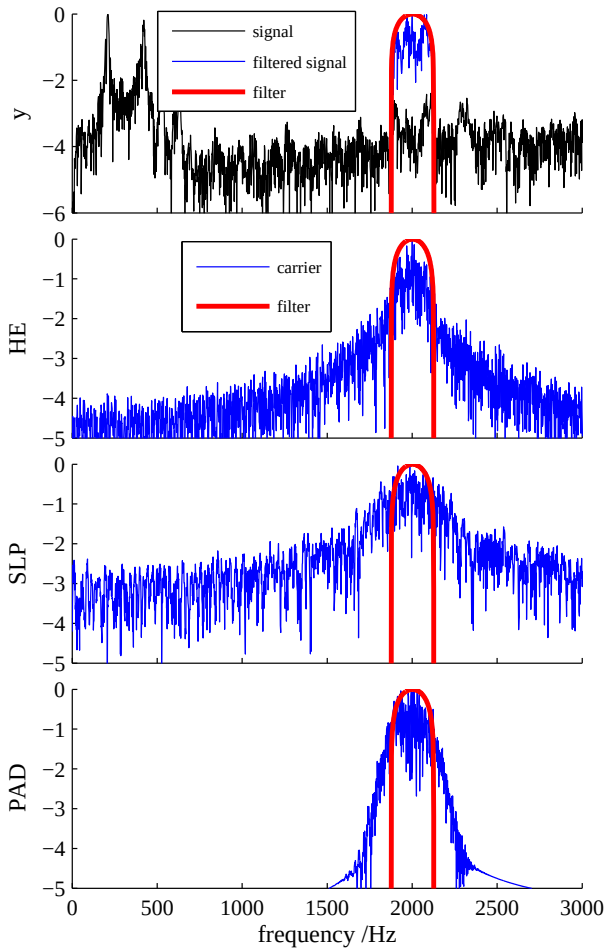

Fig. 10. Demodulating a filtered speech signal. A speech signal (top panel, normalised $\log_{10}$-spectrum shown in black) was filtered using a cosine shaped filter (centred at 2000Hz with full-width 250Hz, smooth line) and normalised to unit variance to produce a band-limited signal (top panel, blue line within the filter pass-band). This was demodulated using the HE, SLP and PAD methods and the the lower three panels show the log-spectra of the resulting carriers. The carrier derived using PAD is closer to being limited to the pass-band of the filter than the other methods, but energy still leaks outside.

V. Conclusion

This paper has introduced a new perspective on demodulation, viewing it as a probabilistic inference problem. This perspective led directly to the development of an algorithm called Probabilistic Amplitude Demodulation which proceeds via an optimisation of a non-linear cost function. The Fast Fourier Transform was used to accelerate inference allowing PAD to run in real-time on current hardware, and further approximations based on Laplace's method were introduced to make learning tractable. However, despite these improvements, PAD remains computationally intensive when compared to existing approaches to demodulation. Nevertheless, PAD has several advantages. For instance, we have highlighted five desiderata which previous demodulation algorithms fail to satisfy, but which are fulfilled by PAD. The first is that the method have soft constraints, which is naturally met by PAD because of the probabilistic calculus upon which it is based. Second, we demonstrated that the method can automatically adjust to the signal by learning important parameters, like the time-scale of variation in the modulator and the sparsity of the signal. Third, we have shown that the method can be steered by the user, for example on a speech signal where user-specific priors were used to select between modulation solutions of differing time-scales. Fourth, we demonstrated that the method was robust to broadband noise added to both synthetic and natural data. And fifth, that the PAD solution was consistent,

in the sense that PAD removes almost all of the modulator information from the carrier (the Carrier Identity property).

PAD not only returns an estimate of the modulator in a signal, it also returns an estimate of the uncertainty in the modulator. The fact that PAD handles uncertainties correctly, means that it can be naturally extended to missing-data tasks. The probabilistic framework also lends itself naturally to changes in the assumptions about carrier and modulation content, and so the current algorithm may form the basis of useful extensions. One hope, in particular, is that this approach can be extended to simultaneous amplitude and frequency demodulation.

ACKNOWLEDGMENT

The authors would like to thank the Gatsby Charitable Foundation and the EPSRC for funding.

REFERENCES

[1] H. Dudley, "Remaking speech," *The Journal of the Acoustical Society of America*, vol. 11, no. 2, pp. 169–177, 1939.
[2] J. L. Flanagan and R. M. Golden, "Phase vocoder," *Bell System Technical Journal*, pp. 1493–1509, 1966.
[3] B. E. D. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Communication*, vol. 25, pp. 117–132, 1998.
[4] D. Ellis, "An introduction to signal processing for speech," in *The Handbook of Phonetic Science*, 2nd ed., W. Hardcastle and J. Laver, Eds. Blackwell Handbooks in Linguistics, 2008.
[5] N. Orio, *Music Retrieval: A Tutorial and Review*. Now Publishers Inc., 2006.
[6] S. M. Schimmel, "Theory of modulation frequency analysis and modulation filtering, with applications to hearing devices," Ph.D. dissertation, University of Washington, 2007.
[7] G. Hu, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Transactions on Neural Networks*, vol. 15, pp. 1135–1150, 2004.
[8] L. E. Atlas and C. Janssen, "Coherent modulation spectral filtering for single-channel music source separation," in *Proceedings of the IEEE Conference on Acoustics Speech and Signal Processing*, 2005.
[9] P. C. Loizou, M. Dorman, and Z. Tu, "On the number of channels needed to understand speech," *Journal of the Acoustical Society of America*, vol. 104, no. 4, pp. 2097–2103, 1999.
[10] J. L. Flanagan, "Parametric coding of speech spectra," *Journal of the Acoustical Society of America*, vol. 68, pp. 412–419, 1980.
[11] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *The Journal of the Acoustical Society of America*, vol. 95, no. 2, pp. 1053–1064, 1994.
[12] R. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, pp. 303–304, 1995.
[13] M. F. Dorman, P. C. Loizou, and D. Rainey, "Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs," *Journal of the Acoustical Society of America*, vol. 102, no. 4, pp. 2403–2411, 1997.
[14] L. Xu and B. E. Pfingst, "Relative importance of temporal envelope and fine structure in lexical-tone perception (L)," *The Journal of the Acoustical Society of America*, vol. 114, no. 6, pp. 3024–3027, 2003.
[15] Z. M. Smith, B. Delgutte, and A. J. Oxenham, "Chimaeric sounds reveal dichotomies in auditory perception," *Nature*, vol. 416, no. 6876, pp. 87–90, 2002.
[16] M. G. Heinz and J. Swaminathan, "Quantifying envelope and fine-structure coding in auditory nerve responses to chimaeric speech," *Journal of the Association for Research in Otolaryngology*, vol. 10, no. 3, pp. 407–23, 2009.
[17] J. Dugundji, "Envelopes and pre-envelopes of real waveforms," *IEEE Transactions on Information Theory*, vol. 4, pp. 53–57, 1958.
[18] O. Ghitza, "On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception," *The Journal of the Acoustical Society of America*, vol. 110, no. 3, pp. 1628–1640, 2001.
[19] F. G. Zeng, K. Nie, S. Liu, G. Stickney, E. Del Rio, Y. Y. Kong, and H. Chen, "On the dichotomy in auditory perception between temporal envelope and fine structure cues (L)," *The Journal of the Acoustical Society of America*, vol. 116, no. 3, pp. 1351–1354, 2004.
[20] G. Sell and M. Slaney, "The information content of demodulated speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010, pp. 5470–5473.
[21] P. J. Loughlin and B. Tacer, "On the amplitude- and frequency-modulation decomposition of signals," *The Journal of the Acoustical Society of America*, vol. 100, no. 3, pp. 1594–1601, 1996.
[22] G. Sell and M. Slaney, "Solving demodulation as an optimization problem," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 8, pp. 2051–2066, 2010.
[23] P. Clark and L. E. Atlas, "Time-frequency coherent modulation filtering of nonstationary signals," *IEEE Transactions on Signal Processing*, vol. 57, no. 11, pp. 4323–4332, 2009.
[24] ——, "Modulation decompositions for the interpolation of long gaps in acoustic signals," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2008, pp. 3741–3744.
[25] R. Libbey, *Signal and image processing sourcebook*. Springer, 1994.
[26] D. Vakman, "On the analytic signal, the Teager-Kaiser energy algorithm, and other methods for defining amplitude and frequency," *IEEE Journal of Signal Processing*, vol. 44, no. 4, pp. 791–797, 1996.
[27] L. E. Atlas, Q. Li, and J. Thompson, "Homomorphic modulation spectra," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, 2004, pp. 17–21.
[28] R. E. Turner and M. Sahani, "Probabilistic amplitude demodulation," in *Independent Component Analysis and Signal Separation*, 2007, pp. 544–551.
[29] ——, "Statistical inference for single- and multi-band probabilistic amplitude demodulation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010, pp. 5466–5469.
[30] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
[31] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.
[32] P. Abrahamsen, "A review of Gaussian random fields and correlation functions," Norweigan Computing Centre, Oslo, Tech. Rep. 917, 1997.
[33] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. The MIT Press, 2006.
[34] R. E. Turner, "Statistical models for natural sounds," Ph.D. dissertation, Gatsby Computational Neuroscience Unit, UCL, 2010.
[35] K. Atkinson, *An Introduction to Numerical Analysis*. John Wiley and Sons, 1988.
[36] A. Bultheel and M. V. Barel, "Lanczos algorithm," in *Linear Algebra, Rational Approximation and Orthogonal Polynomials*. Elsevier, 1997, vol. 6, pp. 99–133.

Richard E. Turner received the M. Sci. degree in Physics from the University of Cambridge, UK and the Ph. D. degree in Computational Neuroscience and Machine Learning from the Gatsby Computational Neuroscience Unit, UCL, UK. He is now an EPSRC Postdoctoral research fellow at the Computational and Biological Learning Lab, University of Cambridge, UK and the Laboratory for Computational Vision, NYU, NY, USA. His research interests include machine learning for signal processing and probabilistic models of perception.

Maneesh Sahani received the B. S. degree in Physics and the Ph. D. degree in Computation and Neural Systems from the California Institute of Technology, Pasadena, CA.

He holds a Readership in Theoretical Neuroscience and Machine Learning at the Gatsby Computational Neuroscience Unit, University College London, London, UK. His work explores the roles and uses of probabilistic inference in perception, neural processing and machine learning.