STATISTICAL INFERENCE FOR SINGLE- AND MULTI-BAND PROBABILISTIC AMPLITUDE DEMODULATION

Richard E. Turner and Maneesh Sahani

University College London, Gatsby Computational Neuroscience Unit,
Alexandra House, 17 Queen Square, London. WC1N 3AR

ABSTRACT

Amplitude demodulation is an ill-posed problem and so it is natural to treat it from a Bayesian viewpoint, inferring the most likely carrier and envelope under probabilistic constraints. One such treatment is Probabilistic Amplitude Demodulation (PAD), which, whilst computationally more intensive than traditional approaches, offers several advantages. Here we provide methods for estimating the uncertainty in the PAD-derived envelopes and carriers, and for learning free-parameters like the time-scale of the envelope. We show how the probabilistic approach can naturally handle noisy and missing data. Finally, we indicate how to extend the model to signals which contain multiple modulators and carriers.

Index Terms— Amplitude estimation, Bayes procedures

1. INTRODUCTION

Amplitude demodulation is the task of decomposing a signal into the product of a slowly varying, positive, envelope and a quickly varying (positive and negative) carrier. Demodulation is fundamentally ill-posed; any positive modulator defines a valid carrier, via division of the signal. As such, prior information such as smoothness in the envelope must be leveraged in order to select one of the infinity of valid decompositions. Traditional approaches to demodulation often make these prior assumptions implicit, making it difficult to understand and improve the methods or to adapt them to the particular demands of specific problems. Consequently, these traditional approaches often yield undesirable results when applied to natural sounds like speech. They also suffer from several important theoretical drawbacks, potentially yielding unbounded modulators or carriers, or demodulating band-limited data to yield carriers which are not band-limited (see [1] for a review). Motivated by these deficiencies, we have developed an inferential approach called Probabilistic Amplitude Demodulation (PAD; [2]). By incorporating explicit priors on envelope and carrier, this approach serves to lay out clearly the unavoidable assumptions that determine the solution. It also allows us to tap the powerful machinery of probabilistic inference, thus providing a natural way to describe uncertainty in the estimated quantities (which may then be propagated to provide uncertainty in decisions based on the audio signal, e.g. in speaker-recognition); facilitating the data-driven estimation of various crucial parameters, such as the most natural timescale of envelope modulation; and allowing generalisation of the method to handle noisy or missing data, and thus to restore damaged audio. However, because the required estimation steps often involve iterative refinement of a non-linear cost function, the methods are considerably slower than traditional feed-forward approaches to demodulation.

This paper provides specific algorithms for each of these potential applications using a novel version of PAD described below. It then extends this approach to handle multi-band modulation (where multiple carriers and envelopes combine to produce the signal) within a single inferential process.

1.1. The forward model

The defining feature of probabilistic forward models for amplitude modulation is that they comprise a positive, slowly varying amplitude, a_t, which multiplies a quickly varying real-valued, (positive and negative) carrier, c_t, to produce the data, y_t. Real data is often noisy and so the forward model also incorporates additive uncorrelated non-stationary Gaussian noise. In the single-band model we take the carrier to be white noise. This is often unrealistic (e.g. for speech where the carrier may contain pitch and formant information), but works surprisingly well in practice because a separation in the time-scales of the carrier and amplitude is sufficient to facilitate accurate inference. The positive amplitude process is produced by taking a slowly varying real-valued process—henceforth called the transformed amplitude (x_t)—and passing it through a static positive non-linearity. The complete forward model can therefore be written:

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

$$\mu_t = \mu, \quad \Gamma_{t,t'} = \gamma_{|t-t'|}, \quad (2)$$

$$\mathrm{a}_t = \mathrm{a}(\mathrm{x}_t) = \log(1 + \exp(\mathrm{x}_t)), \quad (3)$$

$$p(\mathrm{c}_t|\sigma_\mathrm{c}^2) = \mathrm{Norm}(\mathrm{c}_t; 0, \sigma_c^2), \quad (4)$$

$$p(\mathrm{y}_t|\mathrm{a}_t, \mathrm{c}_t, \sigma_{\mathrm{y},\mathrm{t}}^2) = \mathrm{Norm}(\mathrm{y}_t; \mathrm{a}_t\mathrm{c}_t, \sigma_{\mathrm{y},\mathrm{t}}^2). \quad (5)$$

Thanks to the Gatsby Charitable Foundation for funding.

The transformed amplitudes are produced from a stationary Gaussian process and so this form of PAD is called Gaussian Process PAD (GP-PAD). A standard choice for the transformed amplitude covariance function is the squared-exponential kernel,

$$\gamma_{|t-t'|} = \sigma_{\mathrm{x}}^2 \exp\left(-\frac{1}{2\tau_{\text{eff}}^2}(t-t')^2\right), \tag{6}$$

where the parameter τ_{eff} defines the timescale of a typical sample drawn from the Gaussian Process. The transformed amplitudes are passed through a 'soft threshold-linear' function to produce the amplitudes—the nonlinearity is exponential, and therefore small, for large negative values of x, and linear for large positive values. This modifies the Gaussian marginal distribution of the transformed amplitudes into a sparse distribution over envelope amplitudes, which is often a good match to the amplitude histogram of natural sounds.

1.2. Inference

The two non-linearities of GP-PAD (equations 5 and 5)make exact inference analytically intractable. The simplest approximation is to integrate out the carrier and find the most probable setting of the transformed amplitude variables given the data:

$$\begin{aligned}\mathrm{x}_{1:T}^{\text{MAP}} &= \arg\max_{\mathrm{x}_{1:T}} p(\mathrm{x}_{1:T}|\mathrm{y}_{1:T},\theta), \qquad (7)\\ &= \arg\max_{\mathrm{x}_{1:T}} \log p(\mathrm{y}_{1:T},\mathrm{x}_{1:T}|\theta) = \arg\max_{\mathrm{x}_{1:T}} \mathcal{L}(\mathrm{x}_{1:T}).\end{aligned}$$

There is no closed-form solution for this optimisation problem, but a gradient based method can be used to find a local maximum. The objective-function and the gradients of that function can be computed efficiently, by noting that the objective can be split into a component derived from the likelihood and a component from the prior,

$$\mathcal{L}(\mathrm{x}_{1:T}) = \sum_{t=1}^{T} \log p(\mathrm{y}_t|\mathrm{x}_t,\theta) + \log p(\mathrm{x}_{1:T}|\theta). \tag{8}$$

The likelihood component is simple and fast to compute as $p(\mathrm{y}_t|\mathrm{x}_t,\theta) = \mathrm{Norm}(\mathrm{y}_t; 0, \mathrm{a}_t^2\sigma_{\mathrm{c}}^2 + \sigma_{\mathrm{y,t}}^2)$. The component from the prior is more challenging as it involves inverting the $T \times T$ covariance matrix of the Gaussian Process 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 $T' = 2(T-1)$. These new variables are chosen so that the complete set of augmented variables, $\mathrm{x}_{1:T'}$ are circularly correlated. This places the augmented latent variables on a ring and so the new covariance matrix, $\Gamma_{1:T',1:T'}$, becomes circulant. This leads to efficient computation using the Fast Fourier Transform (FFT):

$$\mathcal{L}(\mathrm{x}_{1:T}) \approx c + \sum_{t=1}^{T} \log \mathrm{a}_t - \frac{1}{2\sigma_{\mathrm{c}}^2}\sum_{t=1}^{T}\frac{y_t^2}{\mathrm{a}_t^2} - \frac{1}{2T'}\sum_{k=1}^{T'}\frac{|\Delta\tilde{\mathrm{x}}_k|^2}{\tilde{\gamma}_k}$$

Where $\Delta\tilde{\mathrm{x}}_k$ is the Discrete Fourier Transform (DFT) of the mean shifted transformed-envelopes $\Delta\mathrm{x}_t = \mathrm{x}_t - \mu$, and $\tilde{\gamma}_k$ is the DFT of the covariance function, which is the spectrum of the Gaussian Process:

$$\Delta\tilde{\mathrm{x}}_k = \sum_{t=1}^{T'} \mathrm{FT}_{k,t}(\mathrm{x}_t - \mu), \quad \tilde{\gamma}_k = \sum_{t=1}^{T'} \mathrm{FT}_{k,t}\gamma_t, \tag{9}$$

$$\mathrm{FT}_{k,t} = \exp(-2\pi i(k-1)(t-1)/T'). \tag{10}$$

The derivatives can be computed using the expressions above and are omitted for brevity. The conjugate gradient method can be used for optimisation.

1.3. Error-bars and parameter learning

Two key advantages of framing demodulation as an inference problem are that it leads to methods for estimating the uncertainties in the recovered amplitudes and for learning the free-parameters in the model. This section describes how to use an approximate version of Laplace's method (itself an approximation) to do this. Laplace's method approximates the posterior distribution over transformed amplitudes by a Gaussian centred at the true posterior mode, and with a covariance matrix given by the negative inverse of the Hessian, H of the log-joint [3],

$$\begin{aligned}&p(\mathrm{x}_{1:T'}|\mathrm{y}_{1:T},\theta) \approx\\ &\qquad \exp\left(\frac{1}{2}(\mathrm{x}_{1:T'} - \mathrm{x}_{1:T'}^{\text{MAP}})^{\mathsf{T}}\mathrm{H}(\mathrm{x}_{1:T'} - \mathrm{x}_{1:T'}^{\text{MAP}})\right), \qquad (11)\end{aligned}$$

where,

$$\mathrm{H}_{t,t'} = \left.\frac{d^2}{d\mathrm{x}_t d\mathrm{x}_{t'}} \log p(\mathrm{y}_{1:T},\mathrm{x}_{1:T'}|\theta)\right|_{\mathrm{x}_{1:T'}=\mathrm{x}_{1:T'}^{\text{MAP}}}. \tag{12}$$

Laplace's approximation thus provides an estimate of the posterior uncertainty ($\Sigma^{\text{post}} = -\mathrm{H}^{-1}$) and it can also be used to perform an approximate integration of the transformed amplitudes,

$$\begin{aligned}p(\mathrm{y}_{1:T}|\theta) &= \int d\mathrm{x}_{1:T'} p(\mathrm{y}_{1:T},\mathrm{x}_{1:T'}|\theta), \qquad (13)\\ &\approx p(\mathrm{y}_{1:T},\mathrm{x}_{1:T'}^{\text{MAP}}|\theta)\frac{(2\pi)^{T-1}}{\sqrt{\det(-\mathrm{H})}}. \qquad (14)\end{aligned}$$

Unfortunately, the Hessian is a $2(T-1)\times 2(T-1)$ matrix and so exact inversion is typically intractable, necessitating a further approximation. Fortunately, the simple structure of the Hessian makes this easy. Specifically, H comprises a diagonal term from the likelihood (D), and a term from the prior, which is the inverse covariance matrix,

$$\begin{aligned}\mathrm{H}^{-1} &= -\Sigma^{\text{post}} = (\mathrm{D} + \Gamma^{-1})^{-1} = \Gamma(\mathrm{D}\Gamma + I)^{-1}, \qquad (15)\\ &= \Gamma^{1/2}(\Gamma^{1/2}\mathrm{D}\Gamma^{1/2} + I)^{-1}\Gamma^{1/2}. \qquad (16)\end{aligned}$$

This new form is helpful because 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 property from the prior covariance Γ that only the low-frequency components are strongly active. 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 and eigenvalues of A. Fortunately, the Lanczos algorithm can do just this, requiring only multiplications of A times a vector [4]. These multiplications can themselves be computed rapidly using the FFT.

2. RESULTS

In this section we validate the methods derived above by applying them to natural data. In the first experiment a fully observed spoken sentence sound was demodulated using GP-PAD. A squared-exponential covariance function was used to model the transformed amplitudes. The observation noise was set to zero, $\sigma^2_{\mathrm{y,t}} = 0$, and the remaining parameters, $\theta = \{\sigma^2_{\mathrm{c}}, \sigma^2_{\mathrm{x}}, \mu, \tau_{\mathrm{eff}}\}$, were learned from the approximate marginal likelihood using an iterative grid search. The results, shown in Fig. 1, indicate that GP-PAD discovers modulation content at the time-scale of the phonemes (the timescale learned from the signal was $\tau_{\mathrm{eff}} \approx 20$ms). Both the inferred amplitude and the carriers are well behaved, unlike those recovered from traditional approaches to demodulation. Importantly, when the carriers recovered from the speech sound are themselves demodulated using GP-PAD, the result is a amplitude which is almost constant and a carrier which is equal to a rescaled version of the original carrier. Many demodulation algorithms fail this simple consistency test catastrophically.

GP-PAD is able to estimate modulators in sections of a signal which are missing. This can be handled by setting the noise variance to be infinite in these regions as this means that the prior is used exclusively to fill-in the missing modulator. In order to test this ability on natural signals it is necessary to establish a measure of 'ground-truth'. A consistent approach is to estimate the amplitude of the complete signal using GP-PAD. This can then be compared to the estimates derived from the signals which have missing sections. The quality of the inferences in the missing sections is measured using the signal to noise ratio. The results, shown in Fig. 2, indicate that the envelope of missing sections can be accurately predicted in missing sections of speech up to about 50ms in length.

3. MULTIPLE MODULATORS AND CARRIERS

Many sounds contain multiple carriers and modulators. For example, the vowels of speech can be well approximated by a comodulated harmonic stack of sinusoids. This presents a problem for PAD because it contains just a single carrier and modulator. In this section we show how to generalise PAD to

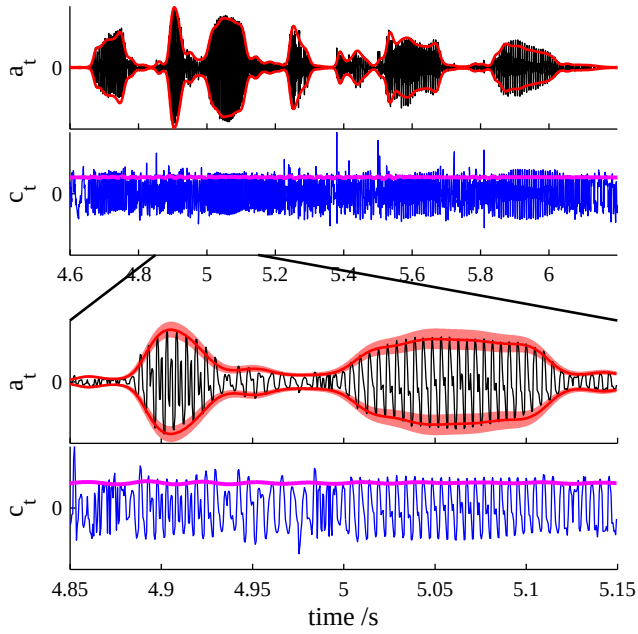

Fig. 1. GP-PAD of a spoken sentence sound shown at two different scales. The speech signal is shown in black. The envelopes are shown in red and the carriers in blue. The error-bars are 3 times the marginal uncertainty derived from Laplace's approximation. The magenta lines show the amplitudes derived from demodulating the carriers using the parameters learned from the original signal.

this new setting. We begin by extending the forward model to comprise a set of positive, slowly varying amplitudes ($\mathrm{a}_{d,t}$) which multiply a set of quickly-varying real-valued, (positive and negative) carriers ($\mathrm{c}_{d,t}$) which are summed, along with Gaussian noise, to produce the data (y_t). That is,

$$\mathrm{y}_t = \sum_{d=1}^{D} \mathrm{c}_{d,t}\mathrm{a}_{d,t} + \sigma_{\mathrm{y}}\epsilon_t \,. \tag{17}$$

The carrier processes are second order auto-regressive (AR(2)) Gaussian random variables,

$$p(\mathrm{c}_{d,t}|\mathrm{c}_{d,t-1:t-2}, \theta) = \mathrm{Norm}\left(\mathrm{c}_{d,t}; \sum_{t'=1}^{2} \lambda_{d,t'}\mathrm{c}_{d,t-t'}, \sigma_d^2\right).$$

The amplitude processes are formed from real-valued, independent transformed amplitudes ($\mathrm{x}_{d,t}$) which are linearly mixed, and then passed through the soft-threshold linear function,

$$\mathrm{a}_{d,t} = \mathrm{a}\left(\sum_{e=1}^{E} \mathrm{g}_{d,e}\mathrm{x}_{e,t} + \mu_d\right), \tag{18}$$

In the following, the transformed amplitudes will be generated from zero-mean stationary Gaussian process with

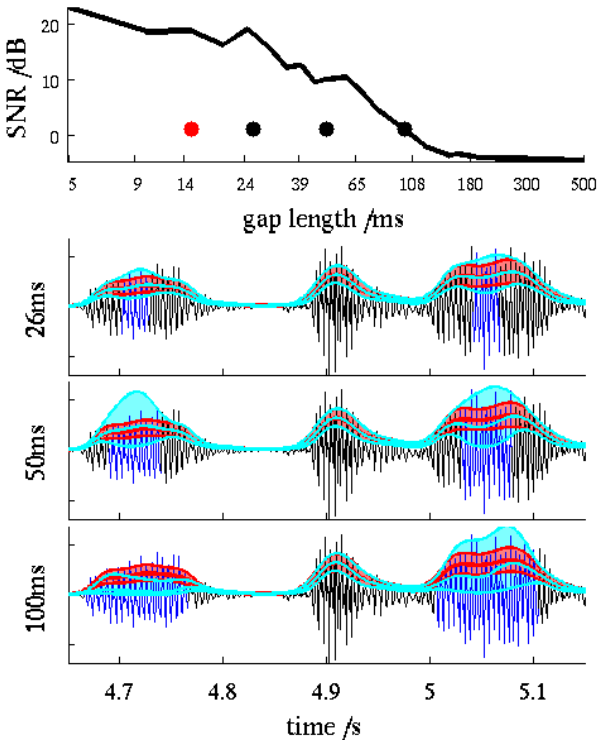

Fig. 2. Filling in the envelopes of missing sections of speech using GP-PAD. Top Panel: Signal to noise ratio (in decibels) of the inferred envelopes as a function of gap size. Bottom panels: A short section of the speech sound (black) with progressively longer missing sections (blue). The size of these gaps is shown for reference on the top plot by black circles. τ_{eff} is shown in red. The envelopes estimated using the complete signals are shown in red with associated error-bars at 3 standard deviations. The envelopes estimated on the missing data, with associated error-bars, are shown in cyan.

squared exponential kernels. Typically the parameters of the AR(2) processes and the transformed envelopes are chosen so that the carriers are expected to vary more quickly than the amplitudes.

3.1. Inference and Learning

Exact inference in this model is analytically intractable and so approximations are required for inference. One approach is to follow the scheme developed for PAD which is to find the most probable transformed amplitude, given the data,

$$\mathrm{X}^{\mathrm{MAP}} = \arg\max_{\mathrm{X}} p(\mathrm{X}|\mathrm{Y},\theta) = \arg\max_{\mathrm{X}} \log p(\mathrm{X},\mathrm{Y}|\theta).$$

The log-joint is complicated because it involves an integral over the carriers,

$$p(\mathrm{X},\mathrm{Y}|\theta) = p(\mathrm{X}|\theta) \int d\mathrm{C}\; p(\mathrm{Y},\mathrm{C}|\mathrm{X},\theta). \qquad (19)$$

However, when the amplitudes are fixed, the joint distribution of the carriers and the data, $p(\mathrm{Y},\mathrm{C}|\mathrm{X},\theta)$, is Gaussian and so it is possible to compute the integral exactly using the Kalman Smoother. The gradients can also be computed using the expectations returned by the Kalman Smoother (see [1] for more details). The parameters of the model, which include the centre-frequencies and bandwidths of the carriers, the time-scales, marginal variances, and means of the transformed modulators, and the weights, can be learned using a similar scheme to that described in section 1.3 (again we refer the reader to [1] for more details).

3.2. Results

The methods described in the previous section were used to learn the parameters of the model from training data which included running water, wind, rain, fire, and speech. Sample sounds generated from the forward model using these parameters indicate the aspects of the data which the model is capturing (see `http://tinyurl.com/archivesounds`). Realistic sounding running water, wind, rain and fire sounds are produced indicating that these acoustic-textures are defined by relatively low-level statistics. In contrast, the speech sound is too rich to be accurately captured.

4. CONCLUSIONS

This paper has introduced a new approach to Probabilistic Amplitude Demodulation. Methods have been provided for inferring envelopes, estimating the uncertainty in these inferences, and for learning the parameters of the model such as the time-scale of the modulation. The power of these new methods was illustrated on speech sounds where they were able to infer the modulation in missing sections up to 50ms in duration. Finally we indicated how to extend the framework to handle multiple carriers and amplitudes.

5. REFERENCES

[1] R. E. Turner, *Statistical Models for Natural Sounds*, Ph.D. thesis, Gatsby Computational Neuroscience Unit, UCL, 2009.

[2] R. E. Turner and M. Sahani, "Probabilistic amplitude demodulation," in *Independent Component Analysis and Signal Separation*, 2007, pp. 544–551.

[3] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, 2003.

[4] A. Bultheel and M. Van Barel, "Lanczos algorithm," in *Linear Algebra, Rational Approximation and Orthogonal Polynomials*, vol. 6 of *Studies in Computational Mathematics*, pp. 99 – 133. Elsevier, 1997.