Time-Frequency Analysis as Probabilistic Inference

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

***Abstract*—This paper proposes a new view of time-frequency analysis framed in terms of probabilistic inference. Natural signals are assumed to be formed by the superposition of distinct time-frequency components, with the analytic goal being to infer these components by application of Bayes' rule. The framework serves to unify various existing models for natural time-series; it relates to both the Wiener and Kalman filters, and with suitable assumptions yields inferential interpretations of the short-time Fourier transform, spectrogram, filter bank, and wavelet representations. Value is gained by placing time-frequency analysis on the same probabilistic basis as is often employed in applications such as denoising, source separation, or recognition. Uncertainty in the time-frequency representation can be propagated correctly to application-specific stages, improving the handing of noise and missing data. Probabilistic learning allows modules to be co-adapted; thus, the time-frequency representation can be adapted to both the demands of the application and the time-varying statistics of the signal at hand. Similarly, the application module can be adapted to fine properties of the signal propagated by the initial time-frequency processing. We demonstrate these benefits by combining probabilistic time-frequency representations with non-negative matrix factorization, finding benefits in audio denoising and inpainting tasks, albeit with higher computational cost than incurred by the standard approach.**

***Index Terms*—Audio signal processing, inference, machine-learning, time-frequency analysis.**

I. Introduction

MANY real-world signals are characterized by sparse frequency content that varies relatively slowly. Examples include spoken vowels and other animal vocalizations, in which harmonic structure remains relatively stationary for tens of milliseconds or more, and electroencephalographic and other physiological signals, which often contain slowly modulated oscillatory components laid over a broader-band background. Consequently, analyses that extract the time-dependent locally stationary frequency content of natural signals are ubiquitous, finding applications in areas such as speech recognition, audio retrieval and restoration, medical signal processing, and source separation and localization.

Time-frequency analysis is a mature field with a well-developed mathematical foundation that characterizes a variety of different time-frequency representations (e.g. [1]) and efficient digital implementations (e.g. [2]). Nonetheless, open issues remain. For one, different choices of window function in the short-time Fourier transform, transfer functions in a filter-bank, or mother wavelet in a wavelet transform yield quite different representations. Despite several proposals [3]–[8] there is no consensus on how to select the best time-frequency representation for a particular signal or task, nor are there robust algorithms for automatic (and potentially time-varying) signal-dependent adaptation of the representation. Similarly, corruption of a signal by noise or missing samples should introduce uncertainty into the values of the time-frequency representation; but again, no unified robust method exists for computing and handling such uncertainty.

These issues are sharpened when the time-frequency representation is not the goal in itself, but instead forms a pre-processing stage to an adaptive application module such as a classifier, recognizer, or source-separation algorithm. The conventional approach is to set the parameters of the time-frequency representation first, and then to select the parameters of the second-stage application based on the transformed signals. This step-wise approach has three limitations. First, it necessitates a cumbersome process of validation to find time-frequency parameters that improve performance at the second stage. In all but the simplest cases no more than a small set of parameters can feasibly be evaluated, thus limiting the capacity to identify the optimal values and making continuous adaptation impossible. Second, where time-frequency representations are overcomplete, representations derived from real signals are constrained to lie on a submanifold of the full representational space [1]. Learning [9] and prediction [10] in the application module should respect these constraints, but this is made complicated by the algorithmic separation. Similarly, while signals from different sources may be linearly superimposed in the waveform, their time-frequency representations combine in a more complicated way [11]. The separation of representation and application again makes it difficult to flexibly account for such combination rules in later processing. Third, probabilistic application modules—such as hidden Markov models (HMMs), non-negative matrix factorization (NMF) or independent component analysis—perform best with information about the reliability of input values, but this information about uncertainty is precisely that which is difficult to propagate to a conventional time-frequency representation.

Manuscript received May 04, 2014; revised August 14, 2014; accepted September 22, 2014. Date of publication October 08, 2014; date of current version November 05, 2014. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Joseph Tabrikian. Funding was provided by EPSRC (grant numbers EP/G050821/1 and EP/L000776/1) and Google (R.E.T.) and by the Gatsby Charitable Foundation (M.S.).

R. E. Turner is with the Department of Engineering, University of Cambridge, Cambridge CB2 1TN, U.K. (e-mail: ret26@cam.ac.uk).

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

This paper has supplementary downloadable multimedia material available at http://ieeexplore.ieee.org provided by the authors. This includes a technical report which provides further details on mathematical aspects of the paper. This material is 1–3 MB in size.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2014.2362100

The goal of this paper is to provide a new perspective on time-frequency analysis, extending the classical framework so as to compute and represent uncertainty, select parameters of the representation in a principled way, and provide straightforward integration with applications. The core idea has connections to the frameworks of both Wiener [12] and Kalman [13] filtering; it is to frame time-frequency analysis as a problem of estimating unknown underlying signals from (possibly noisy) observations. A natural signal is assumed to be formed by a superposition of time-varying sub-band elements, with varying degrees of time-frequency concentration. The analytic task then becomes to find these components.

It will be useful to review the general schema of probabilistic inference as it applies here [14]. Consider a signal observed at T samples, $\mathbf{y} = \mathrm{y}_{1:T}$ (the subscript notation $_{1:N}$ represents the sequence of indices $\{1, 2, \ldots N\}$; for compactness we use summary symbols such as $\mathbf{y}$ when we do not need to refer to specific samples). We assume that this signal depends on a set of D unobserved components $\mathrm{X} = \mathrm{x}_{1:D,1:T}$ according to a parametrized conditional probability distribution $p(\mathbf{y}|\mathrm{X}, \theta)$. For now we use the symbol θ generically to represent the union of the parameters of all relevant probability distributions; specific choices of distributions and their parameters are described in later sections. By assumption, each sequence $\mathrm{x}_{d,1:T}$ (for $d \in 1:D$) makes a spectrally local contribution to $\mathbf{y}$, with a spectral profile that is determined by the parameters. Thus, the set of sequences X constitutes a particular time-frequency representation for the signal $\mathbf{y}$, the form of which is determined by θ and the family of distributions $p(\cdot|\cdot, \theta)$.

The probabilistic dependence of $\mathbf{y}$ on X makes it possible to model interference, measurement noise, quantization, and other forms of signal corruption. This incorporation of an explicit "noise model" (where "noise" stands for all forms of corruption) is a hallmark of probabilistic methods. It removes the need for separate signal recovery or denoising, makes it possible to determine noise parameters adaptively, and to quantify the effects of noise on the estimated time-frequency representation. No generality is lost: if the signal is known to be uncorrupted, the conditional distribution simply picks out with probability 1 a single signal $\mathbf{y}$ for each representation X.

The goal of probabilistic time-frequency analysis is to estimate the representation that underlies a particular measured signal. It is evident from our construction that this problem is ill-posed. The dimensionality of X is D times larger than that of $\mathbf{y}$, and thus many different possible representations X will generally achieve the same likelihood for a given signal $\mathbf{y}$. Thus, a second crucial element of the probabilistic model is the *prior* distribution on X, $p(\mathrm{X}|\theta)$. This defines the values of X that are compatible with the time-frequency model. For instance, if each sequence $\mathrm{x}_{d,1:T}$ represents an amplitude modulated narrowband carrier signal, then it might be reasonable to assume that the spectrum of these sequences should be minimized outside the bandwidth of the corresponding carrier [15], [16].

The two distributions combine according to Bayes' rule to define the *posterior* distribution over X determined by the observed signal and the model parameters.

$$p(\mathrm{X}|\mathbf{y}, \theta) = \frac{p(\mathbf{y}|\mathrm{X}, \theta)p(\mathrm{X}|\theta)}{p(\mathbf{y}|\theta)}. \tag{1}$$

Unless X is severely constrained by the prior, this posterior will assign non-zero probability to many different time-frequency representations, thereby representing uncertainty about its value. The scale of the uncertainty, which may often be summarized by standard deviations for each $\mathrm{x}_{d,t}$ (or more briefly by the total variance or entropy of the posterior) will depend on the parameters of the noise model, the parameters of the prior, and sometimes (but not always) on the signal itself. In this way, the probabilistic formulation directly addresses the issue of uncertainty in the recovered representation.

The calculus of probabilities also provides a natural and principled scheme to select or adjust the representational parameters and thus tailor the time-frequency representation to the features of a particular signal. A fully Bayesian approach would, in fact, integrate over the unknown parameters, with the best-matched parameter values naturally dominating the value of the integral. However, practical considerations often dictate a two-step approach. First, a single optimal parameter value is identified on the basis of the model likelihood

$$\mathcal{L}(\theta) = p(\mathbf{y}|\theta) = \int d\mathrm{X} p(\mathbf{y}|\mathrm{X}, \theta)p(\mathrm{X}|\theta), \tag{2}$$

in which X has been integrated out. The representation is then found according to (1).

The calculation of a time-frequency representation is often only a precursor to a higher-level analysis algorithm. Many such applications are themselves based around a probabilistic model such as an HMM or probabilistic dictionary. Just as the time-frequency analysis model defines a distribution $p(\mathbf{y}|\mathrm{X}, \theta)$ on the signal given the time-frequency representation, the application model defines a distribution $p(\mathrm{X}|\mathrm{W}, \theta)$ on the time-frequency representation given a set of analytic variables W (for now W stands for all variables in the model; later it will refer specifically to the NMF spectral dictionary). Thus, when the time-frequency stage is probabilistic the two stages can be combined to give a marginal likelihood:

$$p(\mathbf{y}|\mathrm{W}, \theta) = \int d\mathrm{X} p(\mathbf{y}|\mathrm{X}, \theta)p(\mathrm{X}|\mathrm{W}, \theta). \tag{3}$$

In essence, the higher-level model specifies an application-specific prior on X.

If the integral of (3) is tractable, then the time-frequency representation becomes implicit, and the analytic variables and parameters can be found without need for an explicit representational step. More generally, however, direct calculation of the posterior on the variables W is not possible. Instead it is necessary to use exact or approximate message passing, computing or approximating the posterior $p(\mathrm{X}|\mathbf{y}, \theta)$ and propagating the implied estimate of X along with the associated uncertainty to obtain an estimate of W.

In this paper we first briefly review the basic properties of classical audio time-frequency analysis that will be important for the probabilistic development. The general inferential framework is introduced in Section III and connected to the classical representations. Subsequent sections consider learning (IV) and non-stationary noise and missing data (V). Finally, we illustrate the combination of probabilistic time-frequency

analysis with NMF (VI), demonstrating improved performance in audio denoising and restoration tasks (VII).

II. Classical Time-Frequency Representations

Time-frequency analysis takes many forms. Here we review some key properties of two simple and widely used approaches: sub-band filtering and demodulation, and the short-time Fourier transform (STFT) and spectrogram.

In sub-band filtering a bank of band-pass filters splits the signal into different frequency channels. In the time domain:

$$\mathrm{x}^{\mathrm{FB}}_{d,t} = \sum_{t'} \mathrm{V}_{d,t-t'} \mathrm{y}_{t'}. \tag{4}$$

where, in principle, the discrete impulse responses of the D filters, $\{\mathrm{V}_{d,t}\}_{d=1}^{D}$ may be infinitely long. The time-frequency representation is then formed either by the sub-band signals themselves, or by their amplitude envelopes (often found using the Hilbert transform) which together characterize the time-frequency distribution of energy in the signal.

The STFT is formed by repeated discrete Fourier transforms of the signal restricted to local windows (W_t),

$$\mathrm{s}^{\mathrm{STFT}}_{d,t} = \sum_{t'} e^{-i\omega_d t'} \mathrm{W}_{t-t'} \mathrm{y}_{t'}. \tag{5}$$

Here $\{\omega_d\}_{d=1}^{D}$ are the frequencies at which the STFT is evaluated. Again, this representation may be transformed to show only the energy density, yielding the short-time spectral density or *spectrogram* given by the magnitude of the STFT, $\mathcal{S}_{d,t} = |\mathrm{s}^{\mathrm{STFT}}_{d,t}|^2$.

The filter bank and STFT output are related. Consider filters constructed by centering a bandlimited transfer function at frequencies $\{\omega_d\}_{d=1}^{D}$. The impulse response functions of the filters are then $\mathrm{V}_{d,t} = \mathrm{W}_t \cos(\omega_d t)$, where W_t is the impulse response of the low-pass filter obtained when the desired transfer function is centered at 0. Substituting into (4) we find that the outputs of these filters correspond to waves centered at ω_d and modulated in amplitude and frequency by the STFT coefficients (with W_t as the window function) [17],

$$\mathrm{x}^{\mathrm{FB}}_{d,t} = \Re\left(e^{i\omega_d t} \sum_{t'} e^{-i\omega_d t'} \mathrm{W}_{t-t'} \mathrm{y}_{t'}\right) = \Re\left(e^{i\omega_d t} \mathrm{s}^{\mathrm{STFT}}_{d,t}\right). \tag{6}$$

Conversely, the STFT-modulated sub-band signals $(\mathrm{z}^{\mathrm{FB}}_{d,t} = e^{i\omega_d t}\mathrm{s}^{\mathrm{STFT}}_{d,t})$ are given by the output of a bank of quadrature filters $(\mathrm{V}_{d,t} = \mathrm{W}_t e^{i\omega_d t})$. It also follows that the Hilbert envelope of the filter bank output is identical to the square root of the spectrogram provided that the filter bandwidth is less than twice the smallest center-frequency, $\Delta\omega < 2\min_d \omega_d$ [17]. Thus, the spectrogram represents the amplitude of bands defined by the window function. Wavelet representations are analogously related to non-uniform filter banks in which window shape W_t scales with the center frequency ω_d. These connections between different time-frequency constructions also carry over to the probabilistic setting.

III. Gaussian Time-Frequency Representations

We begin by considering probabilistic approaches in which (the inferential expected value of) the time-frequency representation depends linearly on the input signal, as it does for the filter-bank and STFT representations. A probabilistic approach is defined by a generative model which specifies how the signal derives from the unobserved sequences [i.e. $p(\mathbf{y}|\mathrm{X}, \theta)$] and gives a prior distribution over those sequences $[p(\mathrm{X}|\theta)]$. A simple choice is to assume that $\mathbf{y}$ is formed by the (weighted) superposition of band-limited signals $\mathrm{x}_{d,1:T}$, possibly corrupted by noise (a similar assumption is made when a signal is re-synthesized from sub-band channels by summation). Then linearity of inference can be assured by setting the distributions of each $\mathrm{x}_{d,1:T}$ and the noise to be Gaussian. Although apparently simple, the Gaussian assumption proves to be of quite general value. It encompasses auto-regressive, moving-average and sinusoidal models [18] and it generalizes simply to nonregularly sampled data in which case the signal is considered to be a realization of a Gaussian process [19] (see section 2 of the supplementary material for further details). Indeed, if the only constraint on the signal $\mathbf{y}$ is that it have finite power, a Gaussian model retains full generality in a maximum entropy sense [20].

We are typically interested in representations that are time-invariant. That is, the representation of a time-shifted signal should be time-shifted by the same amount, but otherwise unchanged. In the probabilistic view, this requires that the generative model be statistically stationary. A Gaussian prior on $\mathrm{x}_{d,1:T}$ is defined by a mean signal $\xi_{d,1:T}$, and a covariance matrix $\Sigma_{d,1:T,1:T}$. Both are simplified by the assumption of statistical stationarity. The mean signal must be constant in time, and with no loss of generality can be taken to be zero (with any constant offset in the signal being added after the x_d are summed). Further, the covariance matrix must be shift-invariant $(\Sigma_{d,t,t'} = \Sigma_{d,|t-t'|})$ and therefore summarized by the expected auto-correlation function. Neglecting edge effects, the eigenvectors of the covariance matrix are sinusoidal functions with eigenvalues given by the power spectrum $(\gamma_{d,k})$,

$$\Sigma_{d,t,t'} \approx \sum_{k=1}^{T} \mathrm{FT}^{-1}_{t,k} \gamma_{d,k} \mathrm{FT}_{k,t'} \tag{7}$$

where we define the discrete Fourier transform matrix with frequency index k and order T as $\mathrm{FT}_{k,t} = e^{-2\pi i(k-1)(t-1)/T}$. The error in this approximation vanishes in the limit $T \to \infty$ (see [21], [22]) and is often negligible in real world applications in which the duration of the signal is much larger than the reciprocal of the bandwidth of the sub-band processes.

We refer to models specifying a stationary joint Gaussian distribution over the signal and its time-frequency representation as Gaussian time-frequency (GTF) models. As will be seen below, GTF models provide probabilistic analogues to both the filter-bank and STFT.

A. Probabilistic Filter Banks

In the basic GTF model, the sub-band processes are generated independently of each other from zero-mean stationary

Gaussian priors with covariance $\Sigma_{d,t,t'} = \Sigma_{d,|t-t'|}$, which determine their frequency profiles; and the signal is formed by their sum, possibly corrupted by uncorrelated Gaussian noise of amplitude σ_y (defined by $\epsilon_t \sim \mathcal{N}(0,1)$).

$$\mathrm{x}_{d,1:T} \sim \mathcal{N}(\mathbf{0}, \Sigma_d), \quad \mathrm{y}_t = \sum_{d=1}^{D} \mathrm{x}_{d,t} + \sigma_{\mathrm{y}} \epsilon_t. \tag{8}$$

Inference in this model yields the familiar sub-band filter-bank form. Since the model is linear and Gaussian, the posterior distribution over the sub-band processes will also be Gaussian,

$$p(\mathrm{x}_{1:D,1:T} | \mathrm{y}_{1:T}, \theta) = \mathcal{N}(\mathrm{x}_{1:D,1:T}; \mu_{1:D,1:T}, \Gamma_{\mathrm{pos}}). \tag{9}$$

In Appendix A we show that the posterior covariance (Γ_{pos}) does not depend on the signal. Thus, the uncertainty in a simple GTF representation where the level of noise in the signal is constant and known depends only on the parameters of the representation. The signal-dependent information is summarized by the posterior mean, $\mu_{1:D,1:T} = \langle \mathrm{x}_{1:D,1:T} | \mathrm{y}_{1:T} \rangle$. Appendix A shows that this mean can be found by passing the signal through a filter bank,

$$\langle \mathrm{x}_{d,t} | \mathrm{y}_{1:T} \rangle = \sum_{t'} \mathrm{V}_{d,t-t'} \mathrm{y}_{t'} = \mathrm{x}^{\mathrm{FB}}_{d,t}. \tag{10}$$

The filters are fixed and do not depend on time, a property which can be seen to follow from the signal-independence of the posterior variance.[1] The filters take the Wiener form [12], with frequency response given by the ratio of the model component power spectrum $(\gamma_{d,k})$ to the model total signal power spectrum $(\gamma_{\mathrm{y},k})$,

$$\sum_t \mathrm{FT}_{k,t} \mathrm{V}_{d,t} = \frac{\gamma_{d,k}}{\gamma_{\mathrm{y},k}} \text{ where } \gamma_{\mathrm{y},k} = \sum_d \gamma_{d,k} + T\sigma^2_{\mathrm{y}}. \tag{11}$$

That is, the coefficient estimates are recovered by weighting each frequency by the signal-to-noise ratio. Practically, this leads to efficient implementations of estimation using the fast Fourier transform (FFT).

That estimation in the GTF model is equivalent to Wiener filtering should not be surprising. The Wiener filter recovers a target signal with minimal squared error, rejecting stationary interference of known spectral density. In our case, the target signal is the dth component $(\mathrm{x}_{d,1:T})$, while the interference is the mixture of the other components $(\sum_{d' \neq d} \mathrm{x}_{d',1:T})$ and noise. Wiener's assumptions of a linear filter and squared-error measure, are equivalent to the assumed generative linearity, stationarity, and Gaussianity of the GTF framework. The generative view facilitates extensions, however—particularly by allowing adaptive estimation of the spectra of the relevant signals (see Section IV).

[1]As the posterior covariance is data-independent, the effective "window" of the time-frequency representation is not adaptive. In this regard we disagree with the analysis of [23], who argue that the effective window of a specific example of a probabilistic STFT (discussed later) *is* adaptive. In fact, it is edge effects which are causing the window to change in their application, and in central portions of a long signal these do not make a contribution. The cases of adaptive and hierarchical models with signal-dependent uncertainties are treated here in Sections V and VI.

B. The Probabilistic STFT and Spectrogram

The link between the classical filter bank and STFT analyses suggests that it should also be possible to construct a GTF model in which inference matches the STFT. This requires sub-band processes $(\mathrm{s}_{d,1:T})$ that are complex-valued, like the Fourier coefficients. In fact, it is sufficient to define separate and independent priors on the real and imaginary parts of $\mathrm{s}_{d,1:T}$. These are taken to be zero-mean stationary Gaussians, with a low-pass covariance structure. Following the equivalence between STFT-modulated waves and filter bank channels established in (6), the signal is formed by a sum of the real parts of complex-modulated waves and noise:

$$\Re(\mathrm{s}_{d,1:T}) \sim \mathcal{N}(\mathbf{0}, \Gamma_d), \quad \Im(\mathrm{s}_{d,1:T}) \sim \mathcal{N}(\mathbf{0}, \Gamma_d),$$
$$\mathrm{y}_t = \sum_{d=1}^{D} \Re(\mathrm{s}_{d,t} e^{i\omega_d t}) + \sigma_{\mathrm{y}} \epsilon_t. \tag{12}$$

This definition echoes the relationship between the classical filter bank and STFT by construction. Defining $\mathrm{z}_{d,t} = \mathrm{s}_{d,t} e^{i\omega_d t}$, we see that $\mathrm{y}_t = \sum_{d=1}^{D} \Re(\mathrm{z}_{d,t}) + \sigma_{\mathrm{y}} \epsilon_t$, and thus $\mathrm{x}_{d,t} = \Re(\mathrm{z}_{d,t})$ as in the classical case. Furthermore, the marginal distribution over $\mathrm{x}_{d,t}$ implied by this construction is identical to that of (8) provided that $\Gamma_{d,t,t'} \cos(\omega_d(t-t')) = \Sigma_{d,t,t'}$ (see Appendix C).

The linear-Gaussian construction again ensures that inference is linear and that the posterior covariance does not depend on the signal. Appendix B shows that the posterior mean is given by:

$$\langle \mathrm{s}_{d,t} | \mathrm{y}_{1:T} \rangle = \sum_{t'} e^{-i\omega_d t'} \mathrm{W}_{d,t'} \mathrm{y}_{t-t'} = \mathrm{s}^{\mathrm{STFT}}_{d,t}, \tag{13}$$

with, in general, a component-dependent window function given by the convolution of the frequency-shifted inverse signal correlation and the component prior correlation function:

$$\mathrm{W}_{d,t} = \sum_{t'} \Gamma_{d,t-t'} \Sigma^{-1}_{\mathrm{y},t'} e^{-i\omega_d t'}. \tag{14}$$

As $\Gamma_{d,t-t'}$ must be low-pass in structure to ensure frequency localization of the sub-band processes, this window function will also be low-pass. When the coefficients have the same prior low-pass covariance structure, $\Gamma_d = \Gamma$, the window becomes component-independent and we recover the standard STFT (5). If, instead, the windows are scaled linearly with frequency, the representation corresponds to a multiscale wavelet transform.

The posterior mean of the complex filter bank coefficients $z_{d,t}$ is obtained by frequency shifting the posterior mean of the STFT coefficients, or equivalently, by filtering using complex filters formed from the STFT window, $\mathrm{V}_{d,t} = e^{i\omega_d t} \mathrm{W}_{d,t}$. When the windows, $\mathrm{W}_{d,t}$, are bandpass, this corresponds to a quadrature filter bank in which the real and imaginary parts are related by the Hilbert Transform. Furthermore, the amplitudes of the posterior mean correspond to the spectrogram,

$$|\langle \mathrm{s}_{d,t} | \mathrm{y}_{1:T} \rangle|^2 = |\langle \mathrm{z}_{d,t} | \mathrm{y}_{1:T} \rangle|^2 = \mathcal{S}_{d,t}. \tag{15}$$

Thus, with appropriate parametrizations, GTF models provide probabilistic analogues to the traditional filter bank, STFT and spectrogram. This theoretical connection to an additive generative model justifies additive re-synthesis techniques. It

also leads to a number of generalizations. In the next section we see how the probabilistic framework can be used to adapt a representation to match the statistics of the signal. Then, in Section V, we will show how to incorporate non-stationary noise and missing data into the framework.

IV. Adaptation of Time-Frequency Representations

In the probabilistic framework, the parameters that dictate the form of the time-frequency representation—the filter functions and STFT window—are derived from the expected covariances of the sub-band processes. These covariances can themselves be fit to provide a statistically optimal description of the signal (or family of signals). This yields a principled approach for adapting the parameters of the representation.

One strategy is to maximize the likelihood $p(\mathrm{y}_{1:T}|\theta)$ (2). In the GTF model, the marginal distribution of the signal is itself Gaussian, with a covariance given by the sum of the sub-band covariances plus a constant diagonal contribution from noise,

$$p(\mathrm{y}_{1:T}|\theta) = \mathcal{N}\left(\mathrm{y}_{1:T}; \mathbf{0}, \sum_{d=1}^{D} \Sigma_{d,1:T,1:T} + \sigma_{\mathrm{y}}^2 \mathrm{I}\right). \quad (16)$$

Direct optimization in the time domain is cumbersome owing to the large covariance matrices involved. However, by transforming to the frequency domain using (7), we obtain a simple expression for the marginal distribution,

$$p(\mathrm{y}_{1:T}|\theta) \propto \prod_{k=1}^{T} \gamma_{\mathrm{y},k}^{-1/2}(\theta) \exp\left(-\frac{1}{2}|\tilde{\mathrm{y}}_k|^2/\gamma_{\mathrm{y},k}(\theta)\right). \quad (17)$$

Here, $|\tilde{\mathrm{y}}_k|^2 = |\sum_{t=1}^{T} \mathrm{FT}_{k,t}\mathrm{y}_t|^2$ is the signal power, and $\gamma_{\mathrm{y},k}(\theta)$ the model power defined in (11) with the parameter dependence made explicit. The parameters of the component spectra can now be adjusted numerically (for example, by gradient ascent) to minimize divergence between the model and signal spectra as measured by the likelihood (17). A similar approach is used in Bayesian spectrum analysis [18] where the components are spectral lines. We find that this frequency-domain fitting approach generally leads to more efficient and more accurate parameter estimation than some other methods suggested in the literature (e.g. [24], [25]).

V. Non-Stationary Models

The explicit generative model of the probabilistic framework can be modified to embody known structure in the signal. An example is a signal corrupted by time-varying noise, or with missing samples. Classical time-frequency methods have no natural way to account for such non-stationarity, and the incorporation of non-stationary noise into the GTF model breaks its simple equivalence to classical approaches. Unfortunately, non-stationarity also makes inference in models defined by explicit covariance structure, as in (8) or (12), computationally burdensome. In this section we develop a different GTF specification, which allows efficient inference using a Kalman filter approach.

The simplest non-stationarity involves time-dependent noise. Formulating a GTF model in terms of complex sub-band processes $\mathrm{z}_{d,1:T}$ we can write

$$\mathrm{y}_t = \sum_{d=1}^{D} \Re(\mathrm{z}_{d,t}) + \sigma_{\mathrm{y}_t}\epsilon_t. \quad (18)$$

The sequence σ_{y_t} tracks the change in noise variance. Missing data are treated by taking the corresponding $\sigma_{\mathrm{y}_t} \to \infty$, so that the processes $\mathrm{z}_{1:D,t}$ are not constrained by the data occurring at these times.

The non-stationarity of the signal in this model breaks the simple convolutional filter relationship of (34), even for time-invariant prior covariances on $\mathrm{z}_{d,1:T}$. Consequently, inference in the GTF models of (8) or (12) would require the expensive inversion of the full expected signal covariance matrix. An alternative is to redefine the prior on $\mathrm{z}_{d,1:T}$ as a Gauss-Markov auto-regressive process. This choice ensures that the inverse covariance matrix is band-diagonal and that the necessary computations can be performed efficiently, for example by a Kalman-filter based message passing approach.

Following (12), the dth complex sub-band process is expected to be an amplitude- and phase-modulated ω_d-wave (or "phasor") $\mathrm{z}_{d,t} = \mathrm{a}_{d,t}e^{i(\omega_d t + \phi_{t,d})}$. Such a phasor can be defined by a complex first-order auto-regressive or AR(1) process:

$$\mathrm{z}_{d,t} = \psi_d e^{i\omega_d}\mathrm{z}_{d,t-1} + \rho_d\epsilon_{d,t}, \quad \text{where } \Re(\epsilon_{d,t}) \sim \mathcal{N}(0,1) \text{ and } \Im(\epsilon_{d,t}) \sim \mathcal{N}(0,1). \quad (19)$$

The complex innovations term, which has independent real and imaginary components with variance ρ_d^2, induces slow variation in the amplitude and phase perturbations over time, with the shrinkage parameter $(0 < \psi_d < 1)$ ensuring the amplitude remains bounded. This model has been introduced as the Probabilistic Phase Vocoder (PPV) and exact inference is possible using the Kalman smoothing algorithm (see [24] and supplementary material). The cost of inference scales as $\mathcal{O}(D^2T)$, i.e. linear in time, but quadratic in the number of filters.

The AR(1) prior induces a bandpass expected spectrum which is suitable for many applications,

$$\gamma_{d,k} = \chi_d(\omega_k - \omega_d) + \chi_d(\omega_k + \omega_d), \quad \text{where } \chi_d(\omega) = \frac{\sigma_d^2}{1 + \psi_d^2 - 2\psi_d\cos(\omega)}. \quad (20)$$

The center frequency is set by ω_d, and the bandwidth by ψ_d. The skirts of this spectral distribution are broad, but can be sharpened by constructing a cascade of AR processes. However, since this increases the computational complexity we focus on the single AR(1) case here.

The same model can be defined in terms of STFT coefficients: $\mathrm{s}_{d,t} = e^{-i\omega_d t}\mathrm{z}_{d,t}$. Substituting this expression into (18) and (19) yields:

$$\mathrm{y}_t = \sum_{d=1}^{D} \Re(e^{i\omega_d t}\mathrm{s}_{d,t}) + \sigma_{\mathrm{y}_t}\epsilon_t,\ \mathrm{s}_{d,t} = \psi_d\mathrm{s}_{d,t-1} + \rho_d\epsilon'_{d,t} \quad \text{where } \Re\left(\epsilon'_{d,t}\right) \sim \mathcal{N}(0,1) \text{ and } \Im\left(\epsilon'_{d,t}\right) \sim \mathcal{N}(0,1). \quad (21)$$

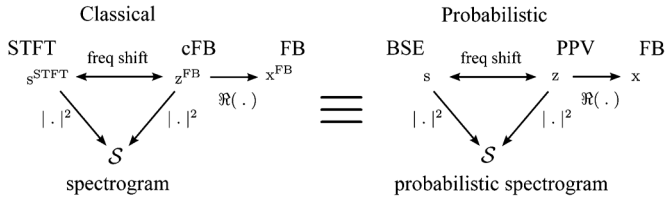


Fig. 1. Relationships between classical and probabilistic time-frequency analysis. A complex filter bank (cFB, z^{FB}) is formed from a set of filters that define the real part of the response (FB, x^{FB}) and their quadrature pairs. The complex filter bank is related to the short-time Fourier transform (STFT, s^{STFT}) via a frequency shift. The spectrogram ($\mathcal{S}$) can either be viewed as the square magnitude of the STFT or the complex filter bank coefficients, which is the Hilbert envelope of the filter bank. There are equivalent probabilistic versions of these representations. For example, the probabilistic phase vocoder (PPV, z) recovers complex filter bank coefficients and Bayesian spectrum estimation (BSE, s) recovers the STFT. The magnitude of these representations is a probabilistic spectrogram.

Thus, the real and imaginary parts of the STFT coefficients $\mathrm{s}_{d,t}$ evolve according to independent AR(1) processes. This model is used in the Bayesian Spectrum Estimation (BSE) framework of [23] (with $\psi_d = 1$) which thus proves to be equivalent to PPV. Exact inference is again possible by Kalman smoothing (see supplementary material).

The PPV and BSE coefficients are related by a frequency shift. If the noise is stationary ($\sigma^2_{\mathrm{y}_t} = \sigma^2_{\mathrm{y}}$), the posterior mean recovered by the Kalman filter in each case is equal to that returned by the associated filter bank or spectrogram. In this sense, BSE and PPV can be identified as more general probabilistic counterparts of the STFT ($\mathrm{s}_{d,t}$) and complex filter bank ($\mathrm{z}_{d,t}$). The amplitudes of these quantities (which are equal) give the corresponding probabilistic spectrogram, or, equivalently, the probabilistic version of the sub-band Hilbert envelopes.[2] The relationships between classical and probabilistic time-frequency analysis are summarized in Fig. 1. Matlab implementations of PPV and BSE, including methods for maximum likelihood learning of the parameters and accelerated inference using the FFT for the stationary noise case, are available from http://learning.eng.cam.ac.uk/Public/Turner/GTFtNMF.

VI. Combining Probabilistic Time-Frequency Analysis With Non-Negative Matrix Factorization

In this final section of theoretical development, we examine how probabilistic time-frequency analysis can be combined with an analytic module defined on the time-frequency representation to allow distributional information (including information about the representational support and uncertainty) to be propagated between the analysis levels.

For illustration, we focus on non-negative matrix factorization (NMF, [26]) applied to a spectrogram. NMF finds a factored approximation which describes the spectrogram as a time-varying sum of a small number of positively weighted spectral basis functions. It is an increasingly popular algorithm in audio analysis, being used, for example, for music transcription [27] and source separation [28]. There are many versions of NMF. Here we concentrate on a probabilistic version of Itakura-Saito NMF (IS-NMF), related to that presented in [29].

[2]As the spectrum of the AR(1) process is not band-limited the Hilbert envelope of the corresponding filter bank output is not precisely equal to the spectrogram, although in practice it is often extremely close.

A. A Probabilistic Interpretation of Itakura-Saito NMF

IS-NMF can be interpreted as a non-negative factored model for the expected squared amplitudes of random STFT coefficients $\mathrm{s}^{\mathrm{STFT}}_{d,t}$. Specifically, we construct the STFT coefficients by scaling unit-variance random complex Gaussian coefficients ($\mathrm{s}_{d,t}$) by non-negative amplitudes $\mathrm{a}_{d,t}$, possibly adding complex Gaussian noise ($\eta_{d,t}$),

$$\mathrm{s}^{\mathrm{STFT}}_{d,t} = \mathrm{s}_{d,t}\mathrm{a}_{d,t} + \sigma_{\mathrm{s}_{d,t}}\eta_{d,t}, \quad \text{with } \Re(\mathrm{s}_{d,t}), \Im(\mathrm{s}_{d,t}), \Re(\eta_{d,t}), \Im(\eta_{d,t}) \overset{\text{iid}}{\sim} \mathcal{N}(0,1). \tag{22}$$

The squares of the amplitude variables are formed by the product of two sets of non-negative basis-functions: temporal basis functions $\mathrm{h}_{l,t}$ and spectral basis-functions $\mathrm{w}_{l,d}$,

$$\mathrm{a}^2_{d,t} = \frac{1}{2}\sum_{l=1}^{L} \mathrm{h}_{l,t}\mathrm{w}_{l,d}. \tag{23}$$

Intuitively, the IS-NMF model describes the STFT coefficients as (complex) Gaussian noise which is modulated in specific cross-band and cross-time patterns.

The Gaussian components of the model define an effective likelihood for the temporal ($\mathrm{H} = \mathrm{h}_{1:L,1:T}$) and spectral basis functions ($\mathrm{W} = \mathrm{w}_{1:L,1:D}$),

$$p\left(\mathrm{s}^{\mathrm{STFT}}_{d,t}|\mathrm{H},\mathrm{W}\right) = \frac{1}{\pi\alpha_{d,t}(\mathrm{H},\mathrm{W})}\exp\left(-\frac{\left|\mathrm{s}^{\mathrm{STFT}}_{d,t}\right|^2}{\alpha_{d,t}(\mathrm{H},\mathrm{W})}\right) \tag{24}$$

where $\alpha_{d,t}(\mathrm{H},\mathrm{W}) = \sum_{l=1}^{L}\mathrm{h}_{l,t}\mathrm{w}_{l,d} + \sigma^2_{\mathrm{s}_{d,t}}$. Optimization of this likelihood to identify the temporal and spectral basis functions is thus equivalent to minimizing the Itakura-Saito divergence between the spectrogram ($\mathcal{S}_{d,t} = |\mathrm{s}^{\mathrm{STFT}}_{d,t}|^2$) and the basis-function approximation ($\alpha_{d,t}(\mathrm{H},\mathrm{W})$).

Further constraints may be imposed on the basis functions to shape the properties of the factorization. For example, temporal NMF (tNMF) adds a cost function to penalize temporal basis functions that change abruptly [28]–[33]. We adopt an analogous approach here, adding a Gaussian process prior on the logarithm of the temporal basis functions with an exponentiated quadratic covariance function [19],

$$\log \mathrm{h}_{l,t} = r_{l,t} + \mu_l \text{ where } r_{l,1:T} \sim \mathcal{N}(0, \Sigma_{l,1:T,1:T})$$
$$\text{and } \Sigma_{l,t,t'} = \sigma^2_{r,l}\exp\left(-\frac{1}{2\tau_l^2}(t-t')^2\right). \tag{25}$$

The hyper-parameter τ_l sets the typical time-scale of variation of the temporal basis functions (measured in samples). The hyper-parameter μ_l controls the mean of the marginal distribution of the temporal basis functions, $p(\mathrm{h}_{l,t}|\theta)$, which takes a log-normal form. This distribution is sparse (the excess kurtosis is positive) and the degree of sparsity is controlled by $\sigma^2_{r,l}$. We choose not to include an explicit prior over spectral basis functions.

B. Limitations of NMF on the Classical Spectrogram

The common use of NMF with pre-computed spectrograms suffers from many of the limitations associated with step-wise algorithms. First, NMF does not capture the dependencies in the spectrogram introduced by the time-frequency analysis. These dependencies are a consequence of the linear injective mapping from signal to STFT which constrains the time-frequency coefficients to lie on a hyperplane [34]. Incorporating these constraints into the probabilistic model should improve prediction (e.g. in denoising or restoration tasks) and learning. This incorporation could be explicit [35] or implicit, as here where we switch from modeling time-frequency coefficients to the waveform.

Second, and similarly, NMF does not model the phase of the STFT coefficients, even though these carry important information with densely sampled narrowband coefficients. Extensions such as the high resolution non-negative matrix factorization model [36], [37] have sought to learn the phase dependencies in the STFT output. Again, such dependence can also be implicitly handled by modeling the waveform directly. In particular, complex phase interplay due to window-function modulation or simultaneous excitation of overlapping filters by narrow-band signals can then be handled automatically.

Third, it is difficult for NMF to accurately capture distortions in the STFT coefficients arising from corruption of the signal. Although NMF can capture independent noise in the STFT coefficients, as in (22), it is more challenging to handle complex correlations. However, even simple forms of waveform noise, such as independent Gaussian distortions, can result in correlated noise in the STFT coefficients. Similarly, it is not completely clear how to compute the STFT coefficients when some waveform data are missing. One option is to discard all coefficients where the window overlaps with the missing region [38], but this can lead to very large segments of the spectrogram being deleted. Again, direct waveform models should be able to limit the impact of missing data to the samples that are directly affected. This approach also has the advantage that once inference has been performed, restoration of the signal waveform is simple and does not require iterative methods [34], [39].

Some previous work has sought to combine NMF with a GTF-like approach to provide a model in the waveform domain [40]. However, in this proposal the signal was first divided into frames, which were modeled as statistically independent and stationary conditioned on the NMF coefficients. Although the independence assumption led to efficient inference schemes based on the STFT, the model was still unable to capture temporal dependencies in the time-frequency coefficients between frames. In the full combination of GTF and NMF we describe below, the STFT window function emerges from assumptions about the bandwidth of the time-frequency coefficients, and the model is able to capture the relevant coefficient interactions.

C. Combining NMF and GTF Models

The probabilistic formulations of time-frequency analysis and NMF are straightforward to combine to provide a time-frequency structured model of the signal waveform. We refer to this class of models (and associated algorithms) as GTF-NMF, or with priors on the temporal basis functions, GTF-tNMF.

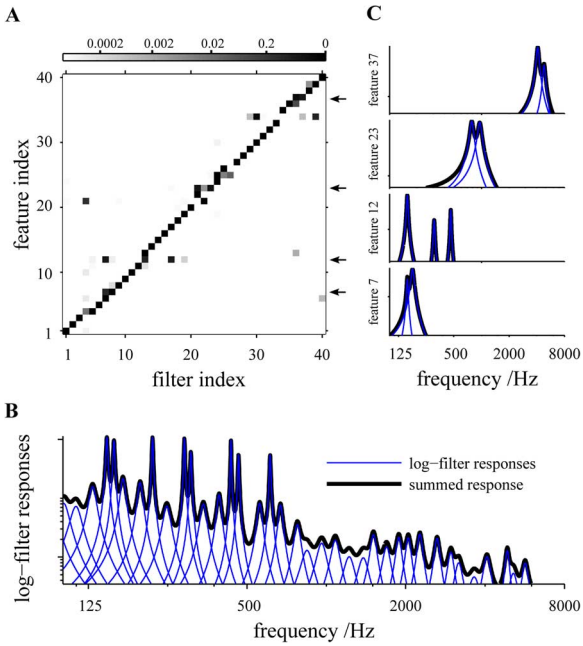


Fig. 2. GTF-tNMF model trained on speech. The model was trained on 3 s of speech from a female speaker. Panel A shows the $L = 40$ learned spectral features in the NMF part of the model, W. These features are very sparse. Panel B shows the $D = 40$ learned filter responses in the GTF part of the model. Panel C shows the filters which are activated by four example spectral features (corresponding to the rows in panel A indicated by arrows). Most spectral features activate filters that are adjacent in frequency, but some activate more widely separated filters (e.g. feature 12). (a) Special features W. (b) Filterbank $\gamma(\theta)$. (c) Example features.

The key step is to replace the independence of $\mathrm{s}_{d,t}$ assumed by probabilistic IS-NMF in (22) with the GTF assumptions of (12) or (21). Here, we choose the BSE approach of (21), although more general sub-band covariance functions could also be used at greater computational cost. The combined model is thus given by:

$$\begin{aligned} \mathrm{y}_t &= \sum_{d=1}^{D} \mathrm{a}_{d,t} \Re(\mathrm{s}_{d,t} e^{i\omega_d t}) + \sigma_{\mathrm{y}_t} \epsilon_t, \text{ where } \epsilon_t \sim \mathcal{N}(0,1) \\ \mathrm{s}_{d,t} &= \psi_d \mathrm{s}_{d,t-1} + \rho_d \epsilon_{d,t} \; \Re(\epsilon_{d,t}), \Im(\epsilon_{d,t}) \sim \mathcal{N}(0,1), \end{aligned} \tag{26}$$

with the amplitudes, $\mathrm{a}_{d,t}$, being factored according to (23) as in IS-NMF, and possibly subject to priors of the form (25) as in tNMF. As before, the model could equivalently be formulated in terms of sub-band processes $(z_{d,t})$, but we will not do so here. We also defer to the supplementary material discussion of equivalent hierarchical rather than product versions.

Whereas the GTF model describes signal waveforms as comprising a sum of independent band-limited Gaussian noise processes, the GTF-tNMF models allow these processes to be modulated with specific cross-band patterns (described by $\mathrm{w}_{l,d}$) that vary over time (according to $\mathrm{h}_{l,t}$) (see Fig. 2 for a visualization of the parameters learned from a speech signal). In either case, the overall model is non-stationary.

D. Learning and Inference in GTF-tNMF Models

As is common in NMF, we seek to learn temporal and spectral basis functions by maximizing the likelihood:

$$\mathrm{W}^{\mathrm{ML}}, \mathrm{H}^{\mathrm{ML}} = \arg\max_{\mathrm{W},\mathrm{H}} \log p(\mathrm{Y}|\mathrm{W}, \mathrm{H}, \theta). \tag{27}$$

As with IS-NMF, the likelihood is formed by integrating out the time-frequency coefficients. Collecting these coefficients together, $\mathrm{S} = \mathrm{s}_{1:D,1:T}$, the log-likelihood is given by,

$$\log p(\mathrm{Y}|\mathrm{W},\mathrm{H},\theta) = \log \int \mathrm{dS}\; p(\mathrm{Y},\mathrm{S}|\mathrm{W},\mathrm{H},\theta). \tag{28}$$

Since the coefficients are now dependent, the integral is harder to perform. However, for fixed basis functions, the model becomes equivalent to a linear-Gaussian state space model with time-varying weights (given by the time-varying amplitudes and sinusoidal basis functions). Thus, the likelihood, along with its derivatives with respect to W, H and the time-frequency parameters can all be computed using the Kalman smoother (see supplementary material). The computational complexity is determined by the cost of the Kalman smoother to be $\mathcal{O}(TD^2)$. In practice, optimization of the likelihood by the conjugate gradient algorithm converged far more quickly than alternatives like Expectation Maximization (EM), as has been observed for other models [41]. Temporal constraints are incorporated as with standard NMF, by including the log-prior $\log p(\mathrm{H}|\theta)$ in the cost function. Learning of the temporal basis functions then returns the *maximum a posteriori* (MAP) rather than maximum likelihood value.

The posterior distribution over the time-frequency coefficients given the signal, $p(\mathrm{S}|\mathrm{Y},\mathrm{W},\mathrm{H},\theta)$, is obtained by the same Kalman smoother employed during learning. Intuitively, this inference amounts to a time-frequency analysis using time-varying filters that slowly adapt to the local spectral content of the signal estimated by NMF.

Matlab implementations of NMF, tNMF and GTF-tNMF can be obtained from the website http://learning.eng.cam.ac.uk/Public/Turner/GTFtNMF.

E. Relationship to Existing Models

The GTF-tNMF model generalizes both GTF and tNMF models. When the amplitude variables are constant, $\mathrm{a}_{d,t} = \mathrm{a}$, it reduces to BSE. Although there is no similar limit in which GTF-tNMF reduces to tNMF exactly, a tNMF-like approach may be recovered when a modified EM algorithm is used for training as follows: 1) the amplitudes are initialized to a constant value so that the first E-step corresponds to computing the STFT, 2) uncertainty in the E-step is discarded as NMF does not treat uncertainty in the STFT coefficients (so-called zero-temperature EM [42]), 3) the E-step is not updated, since tNMF does not re-estimate the STFT coefficients. That is, tNMF is similar to a single iterative M-step in zero-temperature EM learning of the GTF-tNMF model (see supplementary material). More generally, the GTF-tNMF model extends probabilistic amplitude demodulation [16], [22], [43], [44] to the multi-carrier setting. It is a member of the generalized Gaussian scale mixture model family which has been used for image [45] and video [46] modeling. The approach does not treat the time-frequency representation as data [47], but rather as latent variables that are inferred from the signal.

VII. Evaluation

Although the objective of this paper is largely to lay out the theory of probabilistic time-frequency analysis, the practical performance of the new methods developed here was also evaluated on a denoising task and a missing data imputation task. Performance of instances of GTF, NMF, tNMF and GTF-tNMF model classes were compared to one another.

The models were tested on spoken speech from three male and three female speakers in the TIMIT database (arc0, cpm0, adc0 and aem0, vmh0, alk0) [48]. Speaker-specific models were produced by training the models on 6 s of speech from each speaker (sampling rate 16 kHz) and then testing on novel sentences spoken by the same speaker (of duration 3–3.5 s) that were either corrupted by white Gaussian noise or which contained missing sections.

The GTF model used was BSE (21). The NMF models were IS-NMF (22) and its temporal extension IS-tNMF (25). The GTF-tNMF model used was the combination of BSE with IS-tNMF as introduced in Section VI-C (25), (26). All models used $D = 40$ time-frequency components, but results were found to be robust in range $D = 30$–50. For all of these models, we tested versions using an unadapted time-frequency representation with center-frequencies linearly spaced from 50 to 6000 Hz on a log-scale with full-width-half-maximum bandwidths that were 5 percent of the center-frequency. For GTF and GTF-tNMF models we also tested adapted versions in which the filter center-frequencies, bandwidths and variances were learned using maximum-likelihood and approximate maximum likelihood respectively.

The models that included spectro-temporal modulation features (NMF, tNMF and GTF-tNMF) each had $L = 40$ components. In the temporal versions, the time constants were set to 47 ms ($\tau_l = 750$ samples), but performance on held-out data was stable over a wide range of values ($\tau_l = 500$–1000 samples) roughly corresponding to the syllable rate. The basis functions learned using NMF were used to initialize tNMF. In turn, these were used to initialize the GTF-tNMF model and to set the hyper-parameters μ_l and $\sigma^2_{r,l}$. The basis functions were then fine-tuned along with the filter parameters for the filter-adapted version. The parameters learned for the filter-adapted GTF-tNMF model for speaker vmh0 are shown in Fig. 2.

A. Denoising

In the denoising test, Gaussian noise of known variance was added to the test signal and the models were used to reconstruct the signal waveform and spectrogram. For the GTF model this was simple since the denoised coefficients could be estimated using (13) and then added together to reconstruct the signal. The procedure was also straightforward for GTF-tNMF. The MAP temporal basis functions were inferred (using 50 conjugate gradient iterations) and then the waveform was reconstructed as in the GTF models. NMF and tNMF provide denoised estimates of the spectrogram coefficients ($\hat{\mathcal{S}}_{d,t} = \sum_{l=1}^{L} \mathrm{h}_{l,t}\mathrm{w}_{l,d}$), but they do not provide an automatic method to reconstruct the signal waveform. We used an iterative method which refines a signal waveform until it matches the estimated denoised spectrogram. Experiments indicated that the method of [39] scored better on the evaluation metrics than that of [34]. In order for NMF/tNMF to be used for denoising, the noise variance in the STFT coefficients must be estimated ($\sigma^2_{\mathrm{s}_{dt}}$, see (22)). This was done by com-

puting the STFT of Gaussian noise and evaluating the long-term variance within each band.

Three metrics were used to evaluate the models' reconstruction quality: the signal-to-noise ratio (SNR) between the reconstruction and ground-truth waveforms, $\mathrm{SNR}_{\mathrm{DB}}(\mathrm{y},\hat{\mathrm{y}}) = 10\log_{10}\frac{\sum_t \mathrm{y}_t^2}{\sum_t (\mathrm{y}_t-\hat{\mathrm{y}}_t)^2}$; the perceptual quality metric PESQ [49], and the SNR between the logarithm of the reconstructed spectrogram sub-bands and the ground-truth spectrogram, averaged over sub-bands $\frac{1}{D}\sum_{d=1}^{D}\mathrm{SNR}_{\mathrm{DB}}(\log(\mathcal{S}_{d,t}),\log(\hat{\mathcal{S}}_{d,t}))$. The spectrogram was computed using a Hamming window of duration 128 samples with 50% overlap. The last criterion was chosen due to its phase insensitivity and to favor the NMF models as NMF specifically models the time-frequency representation and this measure is not sensitive to noise introduced by the reconstruction step.

The denoising results are shown in Fig. 3. The trends for different speakers were similar and so results were averaged. Errorbars indicate the standard error of the mean, but these errors are dominated by the fact that all methods perform more poorly on some speakers than others. The filter-adapted GTF-tNMF model (red line) generally outperformed the other probabilistic models, leading to average improvements (across all noise conditions and speakers) over the next best method of 2.8 dB for waveform reconstruction, 0.48 for PESQ and 3.2 dB for log-spectrogram reconstruction. The unadapted GTF-tNMF model was slightly better on average for the high-noise conditions, but this was not consistent across subjects. The filter-adapted GTF model (blue line) outperformed tNMF (green line) at waveform reconstruction, as NMF does not model phase and the iterative reconstruction method introduced noise. The snippets shown in the bottom panels in the figure are taken from three different speakers and are indicative of typical behavior. For sections with a spectrum close to that of the long-term average of the whole signal, the GTF model performed similarly to the GTF-tNMF model (bottom snippet), but more generally the ability of GTF-tNMF to adapt to the local spectro-temporal statistics of the signal allowed it to more accurately capture the local spectrum (as for the high frequency content in the top two snippets). Both methods performed more poorly on fricatives than on voiced sections of speech because the phases are more variable and therefore harder to predict. tNMF sometimes performed better than the GTF model on the spectrogram reconstruction measure, presumably because this closely reflected the optimized objective. In this context, it is perhaps surprising that GTF-tNMF still consistently outperformed tNMF on this metric. NMF performed poorly across all measures, as expected, as it does not model temporal dependencies and therefore cannot smooth out the noise. At very high noise levels the methods performed more and more similarly and the PESQ measure began to break down.

Learning the center-frequencies, bandwidths and variances of the time-frequency representation improved the GTF model results significantly (average improvements of 1.0 dB for waveform reconstruction, 0.28 for PESQ and 0.80 dB for log-spectrogram reconstruction, compare solid and dashed blue lines in Fig. 3). The improvement was much more modest for the GTF-tNMF model (0.33 dB for waveform reconstruction, 0.05 for PESQ and 0.36 dB for log-spectrogram reconstruction) possibly because GTF-tNMF could partly compensate for the misspecified time-frequency representation by adapting the spectral basis functions accordingly. However, the magnitude of the effect of this compensation appears to be task-dependent, as filter adaptation significantly improved GTF-tNMF performance on the missing data task.

All of the experiments were conducted on a desktop machine with an Intel i7-3930K 3.20 GHz (12 thread) processor and 64 GB memory. Test times for processing one second of audio (processing times scaled linearly with the signal duration) were: NMF 20 s, tNMF 50 s, GTF 0.05 s, GTF-tNMF 300 s. No decimation was used and all of the time frequency coefficients were maintained at the same sampling rate as the signal. All code was implemented in Matlab with associated overhead for the loops required for the Kalman filter.

Although the focus of this paper is on probabilistic models, we also compared to three well-known audio denoising methods for reference: block thresholding [50], Wiener filtering using decision directed SNR tracking [51], and spectral subtraction [52]. Of these methods, block thresholding performed best on the metrics considered here, but the filter adapted GTF-tNMF model significantly outperformed it (average improvements of 0.94 dB for waveform reconstruction, 0.46 for PESQ and 1.35 dB for log-spectrogram reconstruction).

B. Missing Data Imputation

In the missing data experiment, 0.62–19 ms gaps were introduced into the high energy regions of the test signals (so as to avoid placing gaps in silence). tNMF, GTF and GTF-tNMF models were used to reconstruct the missing regions. NMF was excluded from these experiments as its failure to model temporal dependence precludes interpolation. For the remaining models, missing data interpolation was handled as for noisy data with the noise variance set to infinity in the missing regions ($\sigma^2_{\mathrm{y}_t}=\infty$ and $\sigma^2_{\mathrm{y}_{t,d}}=\infty$). For tNMF these regions in the spectrogram were extended to all affected spectrogram frames.

The missing data results are shown in Fig. 4. The same evaluation measures were used as in the denoising experiments, with the SNR measures being computed only for the reconstructions of the missing regions rather than over the entire waveform. However, since the PESQ measure requires input signals that are longer than the missing sections used here, this measure was computed on the whole waveform. The filter-adapted GTF-tNMF model (red line) generally outperformed the filter-adapted GTF model (blue line) with average improvements of 4.2 dB for waveform reconstruction, 0.57 for PESQ and 4.7 dB for log-spectrogram reconstruction. The snippets shown in the bottom panels in the figure are indicative of typical behavior. Again the GTF-tNMF model made better interpolations than the GTF model when the segment was a poor match to the long-term spectrum of the signal. In the reconstruction experiments, adapting the time-frequency representation led to significant improvement for both the GTF (1.8 dB for waveform reconstruction, 0.22 for PESQ and 1.4 dB for log-spectrogram reconstruction) and GTF-tNMF models (2.7 dB for waveform reconstruction, 0.34 for PESQ and 5.0 dB for log-spectrogram

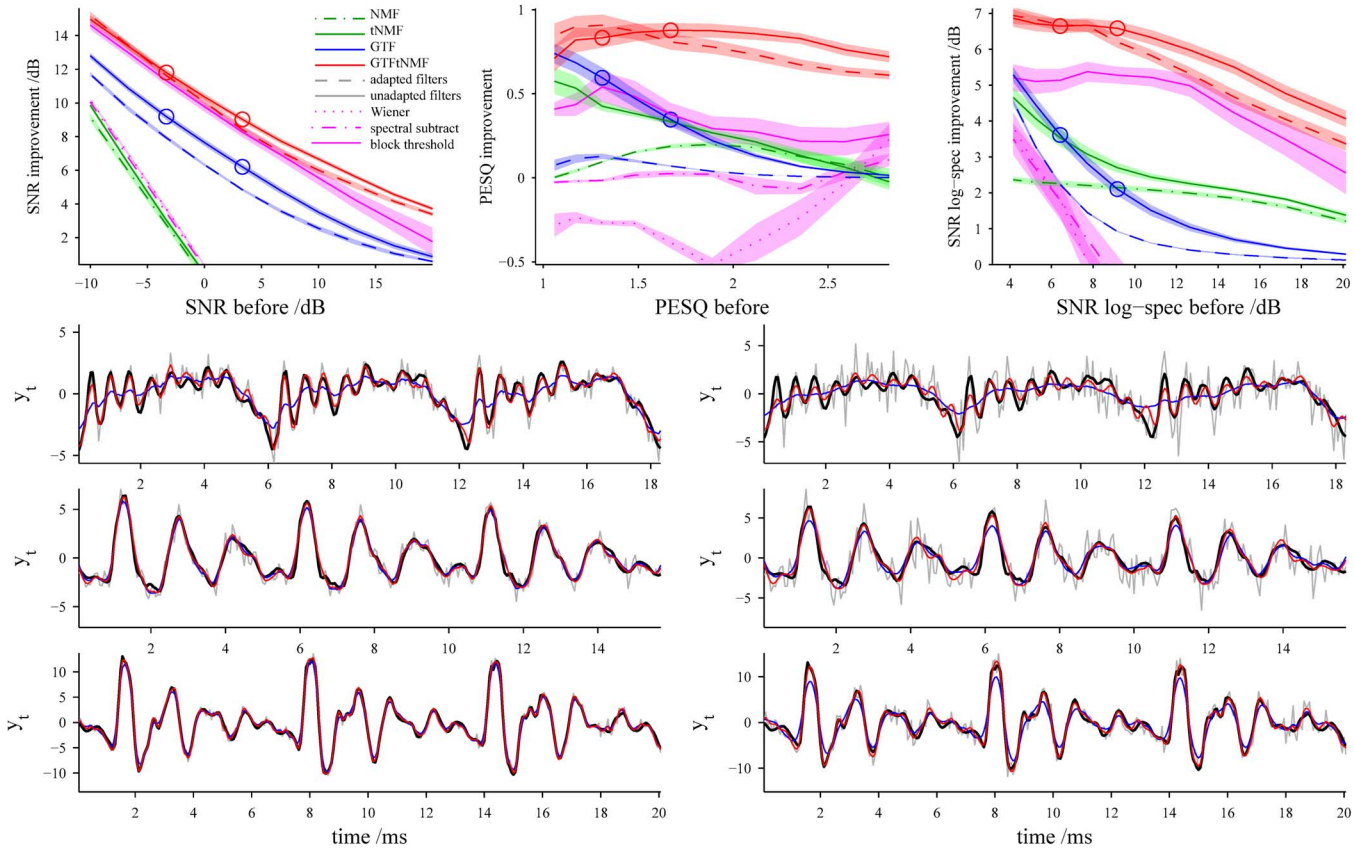


Fig. 3. Denoising results. White Gaussian noise was added to clean speech signals from six speakers. Six probabilistic models and three classical denoising algorithms used to reconstruct the signal and spectrogram. Reconstruction quality compared to ground truth was evaluated as a function of the initial quality using three different measures: SNR of the waveform (top left), PESQ (top center) and spectrogram SNR (top right). The abscissa shows the quality of the noisy signal and the ordinate shows the improvement (quality of reconstruction minus quality of noisy signal). The lines are averages across the six speakers and the shaded regions around them are the standard errors. The lower panels show small sections of the reconstructions for three different speakers for the two noise conditions indicated by open circles in the top panels. The higher SNR condition is shown in the left column. Each panel shows the original waveform (black line), the noisy version (gray line) and the reconstructions from the filter-adapted GTF (blue line) and GTF-tNMF models (red line). Spectrograms are shown in the supplementary material. For full details see Section VII in the main text.

reconstruction). Test times were the same as those reported for the denoising experiments, except for the GTF method. Here it was necessary to use the Kalman filter, rather than FFT-based methods, which took 5 s per second of audio.

VIII. Discussion

The preliminary experiments reported in the previous section indicate that the probabilistic interpretation of time-frequency analysis can translate into improved performance on audio-denoising and missing data imputation tasks. The result is promising, but the critical limitation of the methods developed in this paper is their computational complexity, and in particular the quadratic scaling with the number of sub-bands. However, here there is room for optimism since signal processing has many well developed methods for efficient implementation of time-frequency analysis and there is scope to incorporate these into approximate inference methods. For example, in the models considered here, the sub-bands are not decimated. In principle, there is no reason why models based on multi-rate signal processing could not be treated in the same way, an approach which connects to popular approximation methods for Gaussian processes based upon pseudo-points [40], [53], [54].

IX. Conclusion

This paper introduced a new way of framing time-frequency analysis in terms of a probabilistic inference problem. The new view led to methods for automatically adapting the time-frequency analysis to the statistics of the signal (using maximum-likelihood) and handling non-stationary noise and missing data (using the Kalman filter). The perspective also connected together a number of existing models and algorithms thereby simplifying the literature. Perhaps the most important benefit of the new approach is that it enables time-frequency analysis to be combined with down-stream processing modules that have a probabilistic interpretation. We provide an example in which non-negative matrix factorization was combined with a probabilistic time-frequency model. The hybrid approach was evaluated on two audio-reconstruction tasks involving denoising and missing data imputation of speech. The hybrid model outperformed the component models in terms of waveform SNR, spectrogram SNR and a perceptual quality metric. Future work will focus on reducing the significant computational complexity of the new probabilistic time-frequency approaches by fusing efficient methods from signal processing with approximate inference techniques.

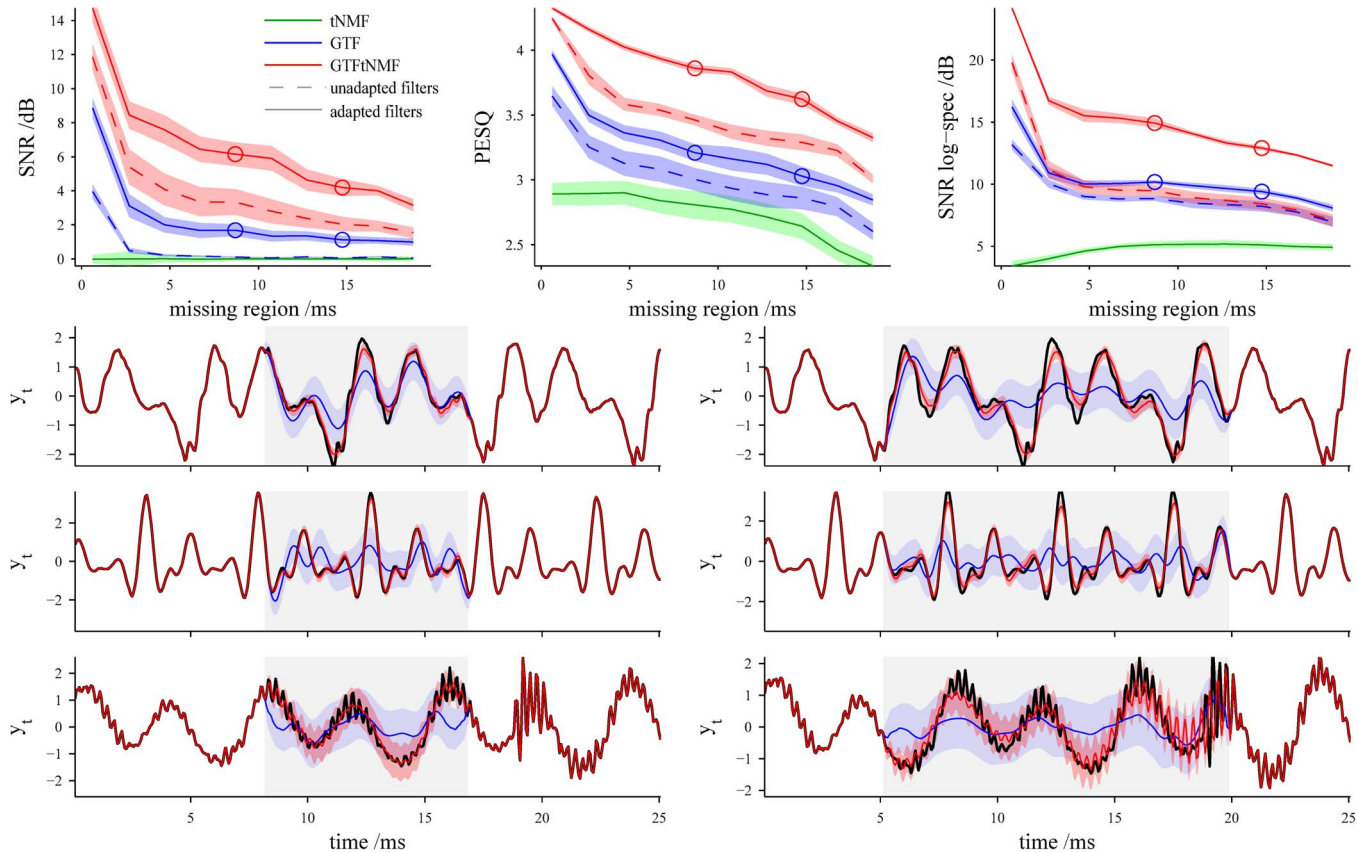


Fig. 4. Missing data results. Short sections were removed from six speech signals from different speakers and five different probabilistic models were used to reconstruct the missing sections. Reconstruction quality compared to ground truth was evaluated as a function of the duration of the missing sections (0.62–19 ms) using three different quality measures: SNR of the waveform (top left), PESQ (top center) and spectrogram SNR (top right). The abscissa shows the gap duration and the ordinate shows the quality of the reconstruction. The lines are averages across the six speakers and the shaded regions around them are the standard errors. The lower panels show small sections of the reconstructions for the two missing data conditions indicated by open circles in the top panels. The shorter gap duration condition is shown in the left column. Each panel shows the original waveform (black line), the missing section (gray shaded region) and the reconstructions from the filter-adapted GTF (blue line) and GTF-tNMF models (red line). For full details see Section VII in the main text.

Appendix A
Estimation in Probabilistic Filter Banks

We show that the posterior mean of a GTF model recovers the Wiener Filter. Consider a vector of stacked time-frequency coefficients, $\mathbf{x} = [\mathrm{x}_{1,1:T}, \mathrm{x}_{2,1:T}, \ldots, \mathrm{x}_{D,1:T}]^{\mathsf{T}}$. The model (8) can be written in terms of this vector,

$$\begin{aligned} p(\mathbf{x}|\theta) &= \mathcal{N}(\mathbf{x}; \mathbf{0}, \Gamma_{\mathrm{pri}}), \\ p(\mathrm{y}_{1:T}|\mathbf{x}, \theta) &= \mathcal{N}(\mathrm{y}_{1:T}; \mathrm{C}\mathbf{x}, \sigma_{\mathrm{y}}^2 \mathrm{I}) \end{aligned} \tag{29}$$

where Γ_{pri} collects the prior covariance matrices,

$$\Gamma_{\mathrm{pri}} = \begin{bmatrix} \Sigma_1 & 0 & \ldots & 0 \\ 0 & \Sigma_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \Sigma_D \end{bmatrix}. \tag{30}$$

Block diagonal matrices like this will be written as $\Gamma_{\mathrm{pri}} = \mathrm{blockdiag}([\Sigma_1 \ \Sigma_2 \ \ldots \ \Sigma_D])$. The weights C select the contributing entries of $\mathbf{x}$ at each time-point, $\mathrm{C} = [\mathrm{I} \ \mathrm{I} \ \ldots \ \mathrm{I}]$. Since the prior and the likelihood are Gaussian, so too is the posterior distribution, $p(\mathbf{x}|\mathrm{y}_{1:T}, \theta) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{\mathrm{pos}}, \Gamma_{\mathrm{pos}})$. The posterior mean ($\boldsymbol{\mu}_{\mathrm{pos}} = \langle \mathbf{x}|\mathrm{y}_{1:T} \rangle$) and the posterior covariance (Γ_{pos}) are found using Bayes' rule,

$$\boldsymbol{\mu}_{\mathrm{pos}} = \frac{1}{\sigma_{\mathrm{y}}^2} \Gamma_{\mathrm{pos}} \mathrm{C}^{\mathsf{T}} \mathrm{y}_{1:T}, \ \Gamma_{\mathrm{pos}}^{-1} = \Gamma_{\mathrm{pri}}^{-1} + \frac{1}{\sigma_{\mathrm{y}}^2} \mathrm{C}^{\mathsf{T}} \mathrm{C}. \tag{31}$$

The posterior covariance consists of two terms, one from the likelihood and the other from the prior. Neither of these two terms depend on the signal and so the uncertainty information is independent of the signal. The posterior covariance can be rewritten $\Gamma_{\mathrm{pos}} = \Gamma_{\mathrm{pri}} - \Gamma_{\mathrm{pri}} \mathrm{C}^{\mathsf{T}} \Sigma_{\mathrm{y}}^{-1} \mathrm{C} \Gamma_{\mathrm{pri}}$ where $\Sigma_{\mathrm{y}} = \sigma_{\mathrm{y}}^2 \mathrm{I} + \mathrm{C}\Gamma_{\mathrm{pri}}\mathrm{C}^{\mathsf{T}} = \sigma_{\mathrm{y}}^2 \mathrm{I} + \sum_{d=1}^{D} \Sigma_d$. Substituting this expression for the posterior covariance into the expression for the posterior mean yields,

$$\boldsymbol{\mu}_{\mathrm{pos}} = \Gamma_{\mathrm{pri}} \mathrm{C}^{\mathsf{T}} \Sigma_{\mathrm{y}}^{-1} \mathrm{y}_{1:T}. \tag{32}$$

Concentrating on the estimates for the dth component, $\langle \mathrm{x}_{d,1:T} | \mathrm{y}_{1:T} \rangle = \Sigma_d \Sigma_{\mathrm{y}}^{-1} \mathrm{y}_{1:T}$, we note that since the prior covariance matrices are stationary, the expression reduces to an acausal filtering operation,

$$\langle \mathrm{x}_{d,t} | \mathrm{y}_{1:T} \rangle = \sum_{t'} \mathrm{V}_{d,t-t'} \mathrm{y}_{t'}, \tag{33}$$

where the filter is given by the convolution between the prior covariance and the inverse signal covariance,

$$\mathrm{V}_{d,t} = \sum_{t'} \Sigma_{d,t'} \Sigma^{-1}_{\mathrm{y},t-t'}. \tag{34}$$

This is the Wiener Filter [12]. The frequency domain view is perhaps more intuitive. To derive it, we use the fact that stationary covariance matrices can be written in terms of the power-spectrum (7) and hence, $\sum_t \mathrm{FT}_{k,t}\mathrm{V}_{d,t} = \gamma_{d,k}/\gamma_{y,k}$. The discrete Fourier transform of the filter is the ratio of the component spectrum to the signal spectrum.

Appendix B
Estimation in Probabilistic STFTs

This section proves the connection between GTF models and the STFT. The model (12) can be written in the form of (29) using a state vector, which plays the role of $\mathbf{x}$, formed by stacking the real and imaginary parts of the coefficients, $\mathbf{s} = [\Re(\mathrm{s}_{1,1:T}), \Im(\mathrm{s}_{1,1:T}), \ldots, \Re(\mathrm{s}_{D,1:T}), \Im(\mathrm{s}_{D,1:T})]^{\mathsf{T}}$. The prior covariance becomes, $\Gamma_{\mathrm{pri}} = \mathrm{blockdiag}([\Gamma_1 \quad \Gamma_1 \quad \ldots \quad \Gamma_D \quad \Gamma_D])$, and the weights C select the contributing entries of $\mathbf{s}$ at each time-point and multiply them by the appropriate sinusoid,

$$\mathrm{C} = [\mathrm{C}_1 \quad -\mathrm{S}_1 \quad \ldots \quad \mathrm{C}_D \quad -\mathrm{S}_D]. \tag{35}$$

The component matrices have sinusoids along the diagonals, $\mathrm{C}_{d,t,t'} = \cos(\omega_d t)\delta_{t,t'}$ and $\mathrm{S}_{d,t,t'} = \sin(\omega_d t)\delta_{t,t'}$.

The posterior mean for the probabilistic STFT coefficients is recovered by substituting these parameters into (32). Collecting real and imaginary parts, this gives,

$$\langle \mathrm{s}_{d,t} | \mathrm{y}_{1:T} \rangle = \sum_{t'} \Gamma_{d,t-t'} e^{-i\omega_d t'} \sum_{t''} \Sigma^{-1}_{\mathrm{y},t'-t''} \mathrm{y}_{t''}. \tag{36}$$

That is, the posterior mean is obtained by filtering through the inverse signal covariance, frequency shifting down towards the base-band, and low-pass filtering through the STFT prior covariance. The order of these operations can be exchanged, for example we can frequency shift the inverse covariance toward the base-band, $\Sigma^{-1}_{\mathrm{y},t} e^{-i\omega_d t}$, to recover the STFT

$$\langle \mathrm{s}_{d,t} | \mathrm{y}_{1:T} \rangle = \sum_{t'} e^{-i\omega_d t'} \mathrm{W}_{d,t-t'} \mathrm{y}_{t'}. \tag{37}$$

Here the window function is the convolution of the frequency-shifted inverse covariance matrix and the component prior covariance, and so will typically be low-pass,

$$\mathrm{W}_{d,t} = \sum_{t'} \Gamma_{d,t-t'} \Sigma^{-1}_{\mathrm{y},t'} e^{-i\omega_d t'}, \tag{38}$$

obtained by frequency shifting the inferential filters to DC.

Appendix C
Estimation in Probabilistic Complex Filter Banks

We begin by connecting the probabilistic complex filter bank to the probabilistic filter bank. The relationship between the likelihood is simple and it has already been established, so here the focus is on the prior over the complex filter bank coefficients. We know the prior over the STFT coefficients is a zero-mean Gaussian with covariance, $\langle \mathrm{s}_{d,t}\mathrm{s}_{d,t'} \rangle = \Gamma_{d,t-t'}$. Combining this with the frequency-shift relationship between the complex filter bank and STFT, $\mathrm{z}_{d,t} = e^{i\omega_d t}\mathrm{s}_{d,t}$, we can compute the covariance of the real and imaginary parts of the coefficients,

$$\begin{aligned} \langle \Re(\mathrm{z}_{d,t})\Re(\mathrm{z}_{d,t'}) \rangle &= \langle \Im(\mathrm{z}_{d,t})\Im(\mathrm{z}_{d,t'}) \rangle = \cos(\omega_d(t-t'))\,\Gamma_{d,t-t'} \\ \langle \Re(\mathrm{z}_{d,t})\Im(\mathrm{z}_{d,t'}) \rangle &= \cos\left(\omega_d(t+t')\right)\Gamma_{d,t-t'}. \end{aligned} \tag{39}$$

Focusing on the covariance of the real components, we note the prior is equivalent to that assumed in the filter bank when the covariances are related by a frequency shift, $\Sigma_{d,t-t'} = \cos(\omega_d(t-t'))\Gamma_{d,t-t'}$.

We now consider estimation. Most of the hard work has already been done in the last section because the posterior mean of the complex filter bank is just a frequency shifted version of the STFT,

$$\langle \mathrm{z}_{d,t} | \mathrm{y}_{1:T} \rangle = e^{i\omega_d t} \langle \mathrm{s}_{d,t} | \mathrm{y}_{1:T} \rangle = \sum_{t'} \mathrm{V}_{d,t-t'} \mathrm{y}_{t'}. \tag{40}$$

where the filter is

$$\mathrm{V}_{d,t} = \sum_{t'} \Gamma_{d,t'} e^{i\omega_d t'} \Sigma^{-1}_{\mathrm{y},t-t'} = e^{i\omega_d t} \mathrm{W}_{d,t}. \tag{41}$$

The real part of the posterior mean is equal to the posterior mean of the real filter bank derived earlier (compare to (34)). Moreover, the filter is equal to the frequency shifted window used in the probabilistic STFT.

References

[1] L. Cohen, *Time Frequency Analysis: Theory and Applications*, 1st ed. Englewood Cliffs, NJ, USA: Prentice-Hall, 1994.

[2] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, ser. Prentice Hall Signal Processing, 3rd ed. Englewood Cliffs, NJ, USA: Prentice-Hall, Aug. 2009.

[3] D. Jones and T. Parks, "A high resolution data-adaptive time-frequency representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 12, pp. 2127–2135, 1990.

[4] R. G. Baraniuk and D. L. Jones, "A signal-dependent time-frequency representation: Optimal kernel design," *IEEE Trans. Signal Process.*, vol. 41, pp. 1589–1602, 1993.

[5] D. Jones and R. Baraniuk, "An adaptive optimal-kernel time-frequency representation," *IEEE Trans. Signal Process.*, vol. 43, no. 10, pp. 2361–2371, 1995.

[6] B. Gillespie and L. Atlas, "Optimizing time-frequency kernels for classification," *IEEE Trans. Signal Process.*, vol. 49, no. 3, pp. 485–496, 2001.

[7] E. Sejdi, I. Djurovi, and J. Jiang, "Time-frequency feature representation using energy concentration: An overview of recent advances," *Digit. Signal Process.*, vol. 19, no. 1, pp. 153–183, 2009.

[8] J. Zhong and Y. Huang, "Time-frequency representation based on an adaptive short-time Fourier transform," *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 5118–5128, 2010.

[9] J. M. Coughlan and A. L. Yuille, "The G factor: Relating distributions on features to distributions on images," in *Advances in Neural Information Processing Systems 14*, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds. Cambridge, MA, USA: MIT Press, 2001, pp. 1231–1238.

[10] R. E. Turner and M. Sahani, "A maximum-likelihood interpretation for slow feature analysis," *Neural Comput.*, vol. 19, no. 4, pp. 1022–1038, 2007.

[11] J. R. Hershey, P. A. Olsen, and S. J. Rennie, "Signal interaction and the devil function," in *Proc. INTERSPEECH (ISCA)*, 2010, pp. 334–337.

[12] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. Cambridge, MA, USA: MIT Press, 1964.

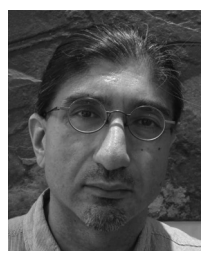
[13] R. Kalman, "A new approach to linear filtering and prediction problems," *Trans. ASME J. Basic Eng.*, ser. D, no. 82, pp. 35–45, 1960.

[14] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[15] G. Sell and M. Slaney, "Solving demodulation as an optimization problem," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 2051–2066, 2010.
[16] R. E. Turner and M. Sahani, "Demodulation as probabilistic inference," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 8, pp. 2398–2411, 2011.
[17] J. L. Flanagan, "Parametric coding of speech spectra," *J. Acoust. Soc. Amer.*, vol. 68, pp. 412–419, 1980.
[18] G. L. Bretthorst, *Bayesian Spectrum Analysis and Parameter Estimation*, J. Berger, S. Fienberg, J. Gani, K. Krickeberg, and B. Singer, Eds. New York, NY, USA: Springer, 1988.
[19] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: MIT Press, 2006.
[20] E. T. Jaynes, "Information theory and statistical mechanics," *Phys. Rev. Series II*, vol. 106, no. 4, pp. 620–630, 1957.
[21] R. M. Gray, *Toeplitz and Circulant Matrices: A Review*. Norwell, MA, USA: Now Publishers, Jan. 2006.
[22] R. E. Turner, "Statistical models for natural sounds," Ph.D. dissertation, Gatsby Computational Neuroscience Unit, Univ. College London, London, U.K., 2010.
[23] Y. Qi, T. P. Minka, and R. W. Picard, "Bayesian spectrum estimation of unevenly sampled nonstationary data," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2002, pp. 1473–1476.
[24] A. T. Cemgil and S. J. Godsill, "Probabilistic Phase Vocoder and its application to interpolation of missing values in audio signals," in *Proc. Eur. Signal Process. Conf.*, Antalya, Turkey, 2005.
[25] J. K. Nielsen, M. G. Christensen, A. T. Cemgil, S. J. Godsill, and S. J. Jensen, "Bayesian interpolation and parameter estimation in a dynamic sinusoidal model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 1986–1998, 2011.
[26] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.
[27] P. Smaragdis and J. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2003, pp. 177–180.
[28] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1066–1074, 2007.
[29] C. Févotte, N. Bertin, and J. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Comput.*, vol. 21, no. 3, pp. 793–830, 2008.
[30] A. T. Cemgil and O. Dikmen, "Conjugate Gamma Markov random fields for modelling nonstationary sources," in *Independent Component Analysis and Signal Separation*, ser. Lecture Notes in Computer Science, M. E. Davies, C. J. James, S. A. Abdallah, and M. D. Plumbley, Eds. Berlin, Germany: Springer, 2007, pp. 697–705.
[31] A. T. Cemgil, P. Peeling, O. Dikmen, and S. Godsill, "Prior structures for Time-Frequency energy distributions," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2007, pp. 151–154.
[32] N. Bertin, R. Badeau, and E. Vincent, "Fast Bayesian NMF algorithms enforcing harmonicity and temporal continuity in polyphonic music transcription," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, 2009, pp. 29–32.
[33] N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 538–549, 2010.
[34] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 2, pp. 236–243, 1984.
[35] J. L. Roux, H. Kameoka, E. Vincent, N. Ono, K. Kashino, and S. Sagayama, "Complex NMF under spectrogram consistency constraints," in *Proc. Acoust. Soc. Jpn. Autumn Meet.*, 2009, no. 2-4-5.
[36] R. Badeau, "Gaussian modeling of mixtures of non-stationary signals in the time-frequency domain (HR-NMF)," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, 2011, pp. 253–256.
[37] R. Badeau and A. Dremeau, "Variational Bayesian EM algorithm for modelling mixtures of non-stationary in the time-frequency domain (HRNMF)," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2013.
[38] P. Clark and L. E. Atlas, "Modulation decompositions for the interpolation of long gaps in acoustic signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2008, pp. 3741–3744.
[39] K. Achan, S. T. Roweis, and B. J. Frey, "Probabilistic inference of speech signals from phaseless spectrograms," in *In Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2003, vol. 16, pp. 1393–1400.
[40] A. Liutkus, R. Badeau, and G. Richard, "Gaussian processes for underdetermined source separation," *IEEE Trans. Signal Process.*, vol. 59, no. 7, pp. 3155–3167, 2011.
[41] K. B. Petersen, O. Winther, and L. K. Hansen, "On the slow convergence of EM and VBEM in low-noise linear models," *Neural Comput.*, vol. 17, no. 9, pp. 1921–1926, 2005.
[42] R. Neal and G. E. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in Graphical Models*. Norwell, MA, USA: Kluwer, 1998, pp. 355–368.
[43] R. E. Turner and M. Sahani, "Probabilistic amplitude demodulation," in *Proc. Independ. Compon. Anal. Signal Separat. (ICA)*, 2007, pp. 544–551.
[44] R. E. Turner and M. Sahani, "Probabilistic amplitude and frequency demodulation," in *Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates, 2012, vol. 24, pp. 981–989.
[45] Y. Karklin and M. S. Lewicki, "A hierarchical Bayesian model for learning nonlinear statistical regularities in nonstationary natural signals," *Neural Comput.*, vol. 17, no. 2, pp. 397–423, 2005.
[46] P. Berkes, R. E. Turner, and M. Sahani, "A structured model of video reproduces primary visual cortical organisation," *PLoS Comput. Biol.*, vol. 5, no. 9, 2009.
[47] R. Everitt, C. Andrieu, and M. Davy, "Online Bayesian inference in some time-frequency representations of non-stationary processes," *IEEE Trans. Signal Process.*, vol. 61, no. 22, pp. 5755–5766, 2013.
[48] W. Fisher, G. Doddington, and G. K. Marshall, "The DARPA speech recognition research database: Specification and status," in *Proc. DARPA Speech Recognit. Workshop*, 1986, pp. 93–100.
[49] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2001, vol. 2, pp. 749–752.
[50] G. Yu, S. Mallat, and E. Bacry, "Audio denoising by time-frequency block thresholding," *IEEE Trans. Signal Process.*, vol. 56, no. 5, pp. 1830–1839, 2008.
[51] P. Scalart and J. Filho, "Speech enhancement based on *a priori* signal to noise estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 1996, vol. 2, pp. 629–632.
[52] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, 1979.
[53] J. Quiñonero Candela and C. E. Rasmussen, "A unifying view of sparse approximate Gaussian process regression," *J. Mach. Learn. Res.*, vol. 6, pp. 1939–1959, 2005.
[54] E. Snelson and Z. Ghahramani, "Sparse Gaussian processes using pseudo-inputs," in *Advances in Neural Information Processing Systems 18*. Cambridge, MA, USA: MIT Press, 2006, pp. 1257–1264.

Richard E. Turner (M'14) received the M. Sci. degree in Physics from the University of Cambridge, U.K., and the Ph.D. degree in Computational Neuroscience and Machine Learning from the Gatsby Computational Neuroscience Unit, University College London (UCL), London, U.K.

He holds a Lectureship in the Computational Perception Group which is part of the Computational and Biological Learning Lab, Department of Engineering, University of Cambridge, U.K. His research interests include machine learning for signal processing and probabilistic models of perception.



Maneesh Sahani (M'10) 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 is currently Professor of Theoretical Neuroscience and Machine Learning at the Gatsby Computational Neuroscience Unit, University College London, London, U.K. His work explores the roles and uses of probabilistic inference in perception, neural processing and machine learning.