# DRIVER ESTIMATION IN NON-LINEAR AUTOREGRESSIVE MODELS

*Tom Dupré la Tour*[†] *Yves Grenier*[†] *Alexandre Gramfort*[†‡]

[†] LTCI, Télécom ParisTech, Université Paris-Saclay, Paris, France
[‡]Inria, Parietal team, Université Paris-Saclay, Saclay, France

## ABSTRACT

In non-linear autoregressive models, the time dependency of coefficients is often driven by a particular time-series which is not given and thus has to be estimated from the data. To allow model evaluation on a validation set, we describe a parametric approach for such driver estimation. After estimating the driver as a weighted sum of potential drivers, we use it in a non-linear autoregressive model with a polynomial parametrization. Using gradient descent, we optimize the linear filter extracting the driver, outperforming a typical grid-search on predefined filters.

***Index Terms*—** non-linear autoregressive models, spectrum estimation, electrophysiology, cross-frequency coupling

## 1. INTRODUCTION

Autoregressive (AR) models are stochastic signal models which have been used for spectral estimation in a wide variety of fields, including geophysics, radio astronomy, speech processing, or neuroscience [1]. Since AR models are linear and stationary, they assume signal statistics to be constant over time, which is not sufficient in many applications.

To overcome this limitation, a large variety of *non-linear* AR models have been proposed, especially in audio signal processing and econometrics, to model fluctuations in mean, spectrum, or energy in the signal. The seminal work of Tong and Lim [2] introduced the threshold AR (TAR) model, where a driving time series $x$ acts as a switching mechanism between several AR models applied on the signal $y$. Several extensions have been developed to get a smoother transition between regimes, like exponential AR (EAR) [3] or smooth transition AR (STAR) [4] models.

Concerning the driver $x$, some models consider it to be hidden, assuming for instance a Markov chain structure [5]. Such probabilistic inference is computationally intensive and cannot be evaluated on a validation set. In other models, a parametric approach enables model evaluation on a validation set, which makes model comparison easy. For instance, the driver can be a function of the signal $y$ itself, as in self-exciting TAR (SETAR) [2, 6] model. A typical choice is $x(t) = y(t-d)$ with a delay $d > 0$. The driver can also be optimized as a weighted average of several potential drivers [7, 8], before being used in a deterministic [7] or a probabilistic [8] TAR model. The set of potential drivers can also be used directly to linearly parametrize the AR coefficients [9, 10, 11].

Our work builds upon driven AR (DAR) models [12], which have been used in particular to estimate cross-frequency coupling (CFC) in neural time-series [13]. In a word, CFC is an inter-frequency coupling phenomenon observed in electrophysiology signals, that is believed to play a central role in functional interactions between neural ensembles [14].

DAR models use a polynomial parametrization over a single driver, which gives a continuous transition between regimes while allowing fast model estimation. The single driver is also essential for interpretability. A limitation of DAR models is the assumption that the driver is given. In practice, the driver is obtained by filtering an exogenous time-series, which requires to search for filter parameters over a grid of values [13].

To soften this known-driver assumption, one could potentially add more drivers directly into DAR models, but that would lead to a very large number of degrees of freedom. Estimation would have high variance, making the risk of model overfit high. We would also lose the interpretability of the single driver, which is key in neuroscience applications.

Instead, we propose to build a weighted average of potential drivers as in [7, 8], and to use it as a single driver in the polynomial parametrization of DAR models [12]. The optimization is thus separated into two steps: optimizing the driver, and optimizing the DAR model. For the former, we propose

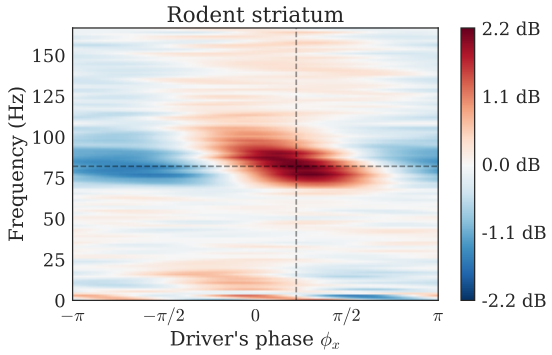


**Fig. 1**. Driven power spectral density of a DAR model fitted on electrophysiology data. The driver's phase is synchronized with a strong amplitude fluctuation around 80 Hz. This phenomenon is known as cross-frequency coupling (CFC).

a fast optimization scheme based on quasi-Newton L-BFGS algorithm [15]. For the latter, we refer the reader to [13].

This paper is organized as follows. First we present the necessary background on DAR. Then we describe the driver decomposition and the proposed gradient descent optimization scheme. Finally we present an extensive validation on both simulations and electrophysiology signals.

## 2. DRIVEN AUTOREGRESSIVE MODELS

Let $y$ be a univariate locally stationary signal, as defined in [16]. An autoregressive (AR) model states that $y$ depends linearly on its own $p$ past values, where $p$ is the *order* of the model:

$$y(t) + \sum_{i=1}^{p} a_i y(t-i) = \varepsilon(t) \tag{1}$$

for all $t \in [p+1, T]$, where $T$ is the length of the signal, and $\varepsilon$ is the *innovation* (or *residual*) modeled with a Gaussian white noise: $\varepsilon(t) \sim \mathcal{N}(0, \sigma(t)^2)$.

To extend this AR model to a non-linear model, one can assume that the AR coefficients $a_i$ are non-linear functions of a given exogenous signal $x$, here called the *driver*. As proposed in [12], we consider these non-linear functions to be polynomials:

$$a_i(t) = \sum_{k=0}^{m} a_{ik}\ x(t)^k \tag{2}$$

This parametrization allows the instantaneous AR model to smoothly change between different regimes, following the fluctuations of the driver $x$.

However, since the model is based only on the driver's value, it does not disentangle the ascending phase from the descending phase of the driver. To fix this issue and obtain phase invariance, the parametrization can be improved using a complex-valued driver $x = x_{re} + jx_{im}$ [13]. The parametrization is now:

$$a_i(t) = \sum_{0 \le k+l \le m} a_{ikl}\ x_{re}(t)^k x_{im}(t)^l = A_i^\top X(t) \tag{3}$$

where $A_i, X(t) \in \mathbb{R}^{\tilde{m}}$ and $\tilde{m} = (m+1)(m+2)/2$.

To improve stability of the estimation, we ortho-normalize the basis $\{x_{re}^k x_{im}^l\}_{0 \le k+l \le m}$, which changes (3) into:

$$a_i(t) = A_i^\top G X(t) \tag{4}$$

with $G \in \mathbb{R}^{(\tilde{m},\tilde{m})}$ such that $(GX(t))_{t\in\Theta}$ is composed of orthogonal and unit-norm vectors. We use Gram-Schmidt process to build $G$.

To allow general power fluctuation over the entire spectrum, the innovation variance is also parametrized by the driver:

$$\log(\sigma(t)) = \sum_{0 \le k+l \le m} b_{kl}\ x_{re}(t)^k x_{im}(t)^l k = B^\top X(t) \tag{5}$$

This model is called a driven AR (DAR) model [13]. A different parametrization can be found in [17], which guarantees stability of the instantaneous AR models. Model parameters $(A_0, ..., A_p, B)$ are estimated by maximizing the model likelihood, and inference is very fast. See [13] for more details.

## 3. DRIVER ESTIMATION

### 3.1. Driver decomposition

In DAR models, the driver $x$ is assumed to be known, but it might not be the case in practice. To have a weaker assumption, we assume here that the driver can be decomposed into a finite set of signals, as in [7, 8]:

$$x(t) = \sum_{n=1}^{N} \alpha_n x_n(t) \tag{6}$$

This set of potential drivers can be, for instance, a Fourier basis $x_n(t) = \exp(j2\pi nt)$, or a Gabor dictionary [18]. Another choice is to use a set of delayed signals $x_n(t) = z(t-n)$ with $-M \le n \le M$. In this case, the coefficients $\alpha_n$ define a linear filter applied on $z$. We used this set in our experiments.

Importantly, we do not use this set of drivers $x_n$ to linearly parametrized AR coefficients as in [9, 10, 11]. Instead, we use the weighted sum $x$ in a DAR model, i.e. in polynomial expressions for AR coefficients and innovation variance.

### 3.2. Model likelihood

We estimate the optimal weights $\alpha_n$ by maximizing the likelihood $L$ of the model:

$$L = \prod_{t=p+1}^{T} \frac{1}{\sqrt{2\pi\sigma(t)^2}} \exp\left(-\frac{\varepsilon(t)^2}{2\sigma(t)^2}\right) \tag{7}$$

$$-2\log(L) = T\log(2\pi) + \sum_{t=p+1}^{T} \frac{\varepsilon(t)^2}{\sigma(t)^2} + 2\sum_{t=p+1}^{T} \log(\sigma(t))$$

Using an alternating optimization approach, we optimize DAR model parameters $(A_0, ..., A_p, B)$ while keeping the driver fixed, and optimizing the driver weights $\alpha_n$ while keeping the DAR model fixed. As this problem is non-convex, weights initialization is key to find good local minima. Optimizing the driver weights can be done with various optimization algorithms. Here, we choose the quasi-Newton L-BFGS algorithm [15], which only requires to compute gradients.

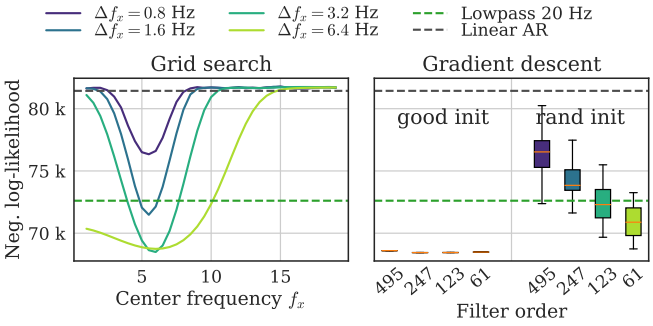


**Fig. 2**. Negative log-likelihood of DAR models fitted with different drivers (lower is better) and evaluated on a validation set. (Left) Grid search: The drivers were bandpass filtered at center frequency $f_x$ with a bandwidth $\Delta f_x$. (Right) Gradient descent: The filters extracting the drivers were optimized by gradient descent, using either several bandpass filter initializations or some random initializations. All bandpass filter initializations with center frequency ranging from 2 Hz to 8 Hz gave optimal and comparable likelihoods. Filter order (495, 247, 123, 61) respectively correspond to bandwidths (0.8, 1.6, 3.2, 6.4) Hz.

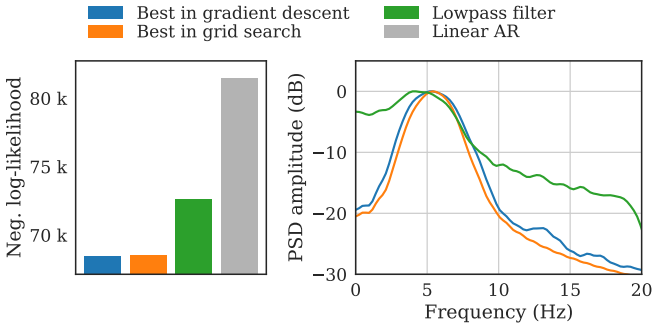


**Fig. 3**. Comparison of 4 models: 3 DAR fitted with different drivers, and 1 linear AR for reference. Both gradient descent and grid search strategies give comparable results, which are much better than when using the driver on the entire band [0, 20] Hz. (Left) Negative log-likelihood on a validation set (lower is better). (Right) Power spectral density of the best driver for each strategy.

### 3.3. Gradient of the log-likehood

The gradient with respect to the weights reads:

$$\frac{\partial \log L}{\partial \alpha_n} = -\sum_{t\in\Theta}\left(\frac{\varepsilon(t)}{\sigma(t)^2}\frac{\partial \varepsilon(t)}{\partial \alpha_n} + (1-\frac{\varepsilon(t)^2}{\sigma(t)^2})\frac{\partial \log\sigma(t)}{\partial \alpha_n}\right)$$

where $\Theta = [p+1, T]$ in the general case. In our experiments, we restricted the sum to $\Theta = [\max(p+1, M), T-M]$ to avoid filtering issue at the edges. In particular, when multiple values of $M$ are compared, we need to restrict the comparison to $\Theta = [\max(p+1, M_{max}), T-M_{max}]$.

The partial derivatives read:

$$\frac{\partial \varepsilon(t)}{\partial \alpha_n} = x_{re,n}(t)\frac{\partial \varepsilon(t)}{\partial x_{re}} + x_{im,n}(t)\frac{\partial \varepsilon(t)}{\partial x_{im}} \quad (8)$$

$$\frac{\partial \log\sigma(t)}{\partial \alpha_n} = x_{re,n}(t)\frac{\partial \log\sigma(t)}{\partial x_{re}} + x_{im,n}(t)\frac{\partial \log\sigma(t)}{\partial x_{im}} \quad (9)$$

Let's note $x_{\_}$ when an expression is similar for both $x_{re}$ and $x_{im}$. From equations (1), (3), and (5), we obtain:

$$\frac{\partial \varepsilon(t)}{\partial x_{\_}} = \sum_{i=1}^{p} A_i^\top G \frac{\partial X(t)}{\partial x_{\_}} y(t-i) \quad (10)$$

$$\frac{\partial \log\sigma(t)}{\partial x_{\_}} = B^\top \frac{\partial X(t)}{\partial x_{\_}} \quad (11)$$

Finally, we can rewrite:

$$\frac{\partial \log L}{\partial \alpha_n} = -\sum_{t\in\Theta}(x_{re,n}(t)g_{re}(t) + x_{im,n}(t)g_{im}(t)) \quad (12)$$

with

$$g_{\_}(t) = \left(\frac{\varepsilon(t)}{\sigma(t)^2}\frac{\partial \varepsilon(t)}{\partial x_{\_}} + (1-\frac{\varepsilon(t)^2}{\sigma(t)^2})\frac{\partial \log\sigma(t)}{\partial x_{\_}}\right) \quad (13)$$

Computing the gradient involves $\mathcal{O}(Tp\tilde{m})$ operations to compute $g_{\_}$, and $\mathcal{O}(TN)$ operations to compute the gradient in (12). In the special case $x_{\_,n}(t) = z_{\_}(t-n)$, we can rewrite (12) into a convolution, which can be performed in $\mathcal{O}(T\log(T))$ using the fast Fourier transform.

### 3.4. Adding a symmetry constraint

In the special case $x_n(t) = z(t-n)$, if we want to make sure the filter is zero-phase, we just need to make the filter symmetric. We rewrite the driver as $x = \alpha_0 x_0 + \sum_{n=1}^{M}\alpha_n(x_n + x_{-n})$, where $N = 2M+1$. The gradient is simply updated into $\frac{\partial x}{\partial \alpha_n} = x_n + x_{-n}$ if $n > 0$ and $\frac{\partial x}{\partial \alpha_n} = x_0$ if $n = 0$.

## 4. RESULTS

### 4.1. Simulations

We created simulated signals with artificial coupling between a driver and a sinusoid. The signals are sampled at $f_s = 240$ Hz, and have a length $T = 10^5$.

We first created a driver $x$ by filtering a Gaussian white noise with a filter $w(t) = b(t)\exp(2j\pi f_x t)$, where $b$ is a Blackman window of order $2\lfloor 1.65 f_s/\Delta f_x \rfloor + 1$, chosen to have a bandwidth of $\Delta f_x$ at $-3$ dB.

This driver $x$ was then used to modulate the amplitude of a sinusoid $y(t) = s(x_{re}(t))\sin(2\pi f_y t)$ where $s$ is a sigmoid function. The modulated sinusoid and the driver were summed up, along with some noise. The noise was pink with a frequency slope $f^{-2}$ above 3 Hz and a plateau below 3 Hz, to mimic electrophysiology signals. The amplitude of the three signals were chosen to have a signal-to-noise ratio (SNR) of 5 dB at $f_x$ and of 20 dB at $f_y$. Importantly, we do not use a DAR model to simulate such data.

We compared different choices of driver, using DAR models of order $(p, m) = (10, 2)$, and comparing their negative

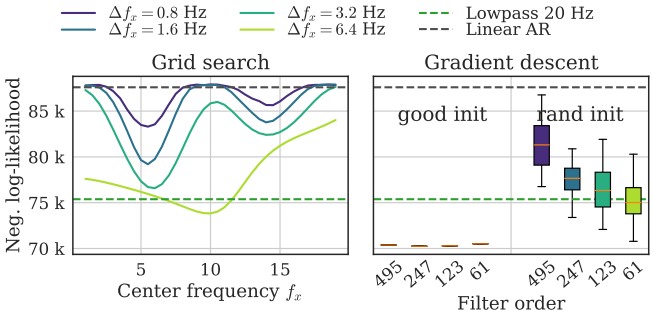


**Fig. 4**. Same as Fig. 2, but using a bimodal driver at 5 and 14 Hz. The gradient descent strategy gave better results than grid-search, when the initial filter was not too poor.

log-likelihood on a validation set using cross-validation. We split the signal into 10 parts of equal size, fitted a DAR model on 5 random parts, and estimating the negative log-likelihood on the 5 other parts, and repeating this process 10 times. To fit the models, we first separated the low frequencies from the high frequencies using a low-pass filter at 20 Hz, which gave $z$ and $y$ respectively. We extracted the driver $x$ from $z$ using different strategies described below, and fitted DAR models on signal $y$ with driver $x$.

The first strategy was grid-search, which searched over a set of bandpass filters as described above. The second strategy used the proposed gradient descent to optimize freely the filter extracting the driver. In this strategy, we used different initializations, since the problem is non-convex and thus may lead to different local minima. Initial filters where either bandpass filters as in the first strategy with center frequency ranging from 2 Hz to 8 Hz, or random filters generated with Gaussian white noise. We also compared with the entire low-pass filter $z$, and with a linear AR which uses no driver.

The first simulation used a single-band $(f_x, \Delta f_x) = (5, 3)$ ground-truth driver, and results are presented in Fig. 2 and 3. Both strategies gave the same best results. We also observed that gradient descent converged to about the same log-likelihood for a large set of reasonable initializations. However, if the initialization does not capture CFC, the optimization leads to poorer results (yet better than the linear AR, even on the validation set).

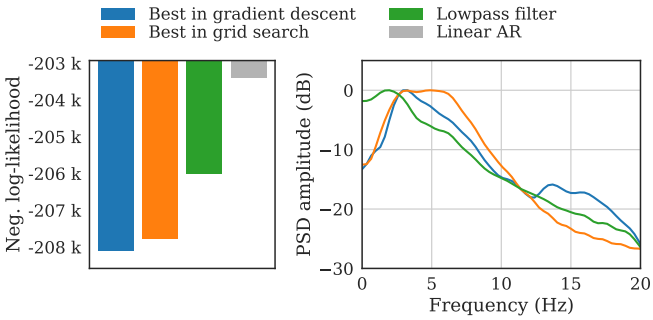


**Fig. 5**. Same as Fig. 3, using electrophysiology data. Gradient descent strategy leads to better results than grid-search.

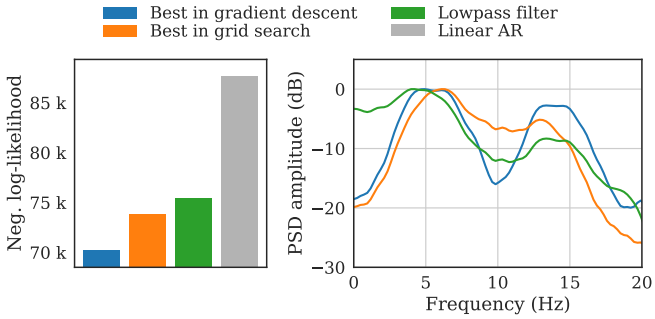


**Fig. 6**. Same as Fig. 3, but using a bimodal driver. With a more complex spectral structure, the gradient descent strategy gives much better results than the grid search one, which is limited to single mode bandpass filters.

The second simulation used a bimodal ground-truth driver, built as the sum of two drivers $x = x_1 + 0.4x_2$, filtered respectively with $(f_{x_1}, \Delta f_{x_1}) = (5, 3)$ and $(f_{x_2}, \Delta f_{x_2}) = (14, 3)$. Results are presented in Fig. 4 and 6. In this case, the grid-search strategy could not correctly capture the two bands, and chose a large filter centered at 10 Hz. It performed only marginally better than the full low-pass signal $z$. In contrast, the optimization by gradient descent correctly captured the two bands, leading to much better results.

### 4.2. Empirical data

We also validated our approach on empirical electrophysiology data containing CFC. The signal is an electro-corticogram (ECoG) channel, recorded on human auditory cortex [19]. It lasts 730 seconds and is sampled at 333.8 Hz. The results presented in Fig. 5 show that the gradient descent strategy leads to a lower negative log-likelihood than the grid-search strategy. In this case, the difference could be related to an asymmetrical shape of the driver spectral peak at 4 Hz.

## 5. CONCLUSION

In this work, we describe how to estimate the driving signal in non-linear time-dependent autoregressive models. By decomposing the driver as a weighted average of potential drivers, we are able to optimize the weights by gradient descent. As a special case, we infer the linear filter to apply to an exogenous signal in order to obtain the driver, and demonstrate the good performance of such driver on both simulated and empirical data, using cross-validation.

## 6. ACKNOWLEDGMENTS

This work was supported by the ERC Starting Grant SLAB ERC-YStG-676943.

## 7. REFERENCES


[1] Steven M. Kay and Stanley L. Marple, "Spectrum analysis–a modern perspective," *Proceedings of the IEEE*, vol. 69, no. 11, pp. 1380–1419, 1981.

[2] Howell Tong and Keng S. Lim, "Threshold autoregression, limit cycles and cyclical data," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 245–292, 1980.

[3] Valérie Haggan and Tohru Ozaki, "Modelling nonlinear random vibrations using an amplitude-dependent autoregressive time series model," *Biometrika*, vol. 68, no. 1, pp. 189–196, 1981.

[4] Kung Sik Chan and Howell Tong, "On estimating thresholds in autoregressive models," *Journal of Time Series Analysis*, vol. 7, no. 3, pp. 179–190, 1986.

[5] James D. Hamilton, "A new approach to the economic analysis of nonstationary time series and the business cycle," *Econometrica: Journal of the Econometric Society*, pp. 357–384, 1989.

[6] Dick van Dijk, Timo Teräsvirta, and Philip Hans Franses, "Smooth transition autoregressive models–a survey of recent developments," *Econometric reviews*, vol. 21, no. 1, pp. 1–47, 2002.

[7] Cathy W.S. Chen and Mike K.P. So, "On a threshold heteroscedastic model," *International Journal of Forecasting*, vol. 22, no. 1, pp. 73–89, 2006.

[8] Senlin Wu and Rong Chen, "Threshold variable determination and threshold variable driven switching autoregressive models," *Statistica Sinica*, vol. 17, no. 1, pp. 241, 2007.

[9] Yves Grenier, "Time-dependent ARMA modeling of nonstationary signals," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 31, no. 4, pp. 899–911, 1983.

[10] Michael Jachan, Gerald Matz, and Franz Hlawatsch, "Time-frequency ARMA models and parameter estimators for underspread nonstationary random processes," *IEEE Transactions on Signal Processing*, vol. 55, no. 9, pp. 4366–4381, 2007.

[11] M.D. Spiridonakos and S.D. Fassois, "Non-stationary random vibration modelling and analysis via functional series time-dependent ARMA (FS-TARMA) models–a critical survey," *Mechanical Systems and Signal Processing*, vol. 47, no. 1, pp. 175–224, 2014.

[12] Yves Grenier, "Estimating an AR model with exogenous driver," Tech. Rep. 2013D007, Telecom ParisTech, 2013.

[13] Tom Dupré la Tour, Lucille Tallot, Laetitia Grabot, Valérie Doyère, Virginie van Wassenhove, Yves Grenier, and Alexandre Gramfort, "Non-linear auto-regressive models for cross- frequency coupling in neural time series," *bioRxiv*, 2017.

[14] Ole Jensen and Laura L. Colgin, "Cross-frequency coupling between neuronal oscillations," *Trends in cognitive sciences*, vol. 11, no. 7, pp. 267–269, 2007.

[15] Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu, "A limited memory algorithm for bound constrained optimization," *SIAM Journal on Scientific Computing*, vol. 16, no. 5, pp. 1190–1208, 1995.

[16] Rainer Dahlhaus, "On the Kullback-Leibler information divergence of locally stationary processes," *Stochastic Processes and their Applications*, vol. 62, no. 1, pp. 139–168, 1996.

[17] Tom Dupré la Tour, Yves Grenier, and Alexandre Gramfort, "Parametric estimation of spectrum driven by an exogenous signal," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 4301–4305.

[18] Hans G. Feichtinger and Thomas Strohmer, *Gabor analysis and algorithms: Theory and applications*, Springer Science & Business Media, 2012.

[19] Ryan T. Canolty, Erik Edwards, Sarang S. Dalal, Maryam Soltani, Srikantan S. Nagarajan, Heidi E. Kirsch, Mitchel S. Berger, Nicholas M. Barbaro, and Robert T. Knight, "High gamma power is phase-locked to theta oscillations in human neocortex," *Science*, vol. 313, no. 5793, pp. 1626–1628, 2006.