A Non-iterative Method for STFT Phase (Re)Construction

Zdeněk Průša, Peter Balazs, *Senior Member, IEEE,* and Peter L. Søndergaard

***Abstract*—A non-iterative method for the construction of the Short-Time Fourier Transform (STFT) phase from the magnitude is presented. The method is based on the direct relationship between the partial derivatives of the phase and the logarithm of the magnitude of STFT with respect to the Gaussian window. Although the theory holds in the continuous setting only, the experiments show that the algorithm performs well even in the discretized setting (Discrete Gabor transform) with low redundancy using the sampled Gaussian window, the truncated Gaussian window and even other compactly supported windows like the Hann window.**

Due to the non-iterative nature, the algorithm is very fast and it is suitable for long audio signals. Moreover, the solutions of iterative phase reconstruction algorithms can be improved considerably by initializing them with the phase estimate provided by the present algorithm.

We present an extensive and reproducible comparison with the state-of-the-art algorithms.

***Index Terms*—STFT, Gabor transform, Phase reconstruction, Griffin-Lim algorithm, Gradient theorem, Numerical integration**

I. Introduction

The phase retrieval problem has been actively researched for decades. It was first formulated for the Fourier transform and later more generally as an inverse of a generic linear system where only the abs. values of the right-hand side are known.

In this paper, we consider a particular case of the phase retrieval problem; the reconstruction from the magnitude of the STFT. More precisely from the magnitude of the Gabor transform coefficients obtained by sampling the STFT magnitude at discrete time and frequency points [1]. The need for an effective way of the phase (re)construction arises in audio processing applications such as source separation, time-stretching/pitch shifting, channel mixing, denoising and missing data inpainting.

The problem has already been addressed by many authors. Among the iterative algorithms, the most widespread and influential is the Griffin-Lim algorithm (GLA) [2] which inspired several extensions [3], [4], [5]. See [6] for a detailed overview of the Griffin-Lim-based algorithms. A different approach was taken in [7], where the authors proposed to express the problem as an unconstrained optimization problem and to solve it using the limited memory Broyden-Flatcher-Goldfarb-Shanno algorithm. Although it is again an iterative algorithm, the computational cost of a single iteration is comparable to that of GLA.

Other approaches are based on reformulating the task as a convex problem [8], [9], [10], [11]. Doing so however squares the size of the signal which makes it unsuitable for long audio signals which typically consist of tens of thousands of samples per second.

The approach from the recently published work [12] builds upon the assumption that the signal is sparse in the original domain, which is not realistic in the context of the audio processing applications mentioned above.

An interesting approach was presented in [13]. It is based on solving non-linear system of equations in each STFT frame. The authors proposed to use iterative solver and initialize it with samples obtained from previous frames. The algorithm was however designed to work exclusively with a rectangular window.

The common problem of the iterative state-of-the-art algorithms is that they require many relatively expensive iterations in order to produce acceptable results. In this paper we propose a non-iterative algorithm (Phase Gradient Heap Integration – PGHI). The theory behind that has been known at least since [14]. It is based on the relationship between gradients of the Gaussian window-based STFT phase and log-magnitude and on the gradient theorem. More precisely, the former allows expressing the phase gradient using the STFT magnitude and the latter gives a prescription how to integrate the phase gradient field to recover the phase up to a global phase shift (or up to sign ambiguity in case of real signals). To our knowledge, no such algorithm has ever been published yet. Curiously enough, in [15] it was even explicitly discouraged to use such algebraic results for practical purposes.

Recently, another non-iterative algorithm was proposed in [16]. It is based on the notion of *phase consistency* used in the phase vocoder [17]. Although the algorithm is simple, fast and it is directly suitable for the real-time setting, it relies on the fact that the signal consists of slowly varying sinusoidal components and fails for transients and broadband components in general. A similar algorithm was introduced in [18], which, in addition, tries to treat impulse-like components separately. Both of these algorithms are in fact very close to the present one since they basically do a crude integration of the estimate of *instantaneous frequency* and of the *local group delay* in case of [18] which are components of the STFT phase gradient.

In the spirit of reproducible research, the implementation of

Manuscript received April 19, 2005; revised August 26, 2015.

Z. Průša* and P. Balazs are with the Acoustics Research Institute, Austrian Academy of Sciences, Wohllebengasse 12–14, 1040 Vienna, Austria, email: zdenek.prusa@oeaw.ac.at (corresponding address), peter.balazs@oeaw.ac.at

P. L. Søndergaard is with Oticon A/S, Kongebakken 9, 2765 Smørum, Denmark, email: peter@sonderport.dk

This work was supported by the Austrian Science Fund (FWF) START-project FLAME ("Frames and Linear Operators for Acoustical Modeling and Parameter Estimation"; Y 551-N13).

the algorithms, audio examples, color version of the figures as well as scripts reproducing experiments from this manuscript are freely available at http://ltfat.github.io/notes/040. The code depends on our Matlab/GNU Octave[19] packages LTFAT [20], [21] (version 2.1.2 or above) and PHASERET (version 0.1.0 or above). Both toolboxes can be obtained freely from http://ltfat.github.io and http://ltfat.github.io/phaseret respectively.

The paper is organised as follows. Section II summarizes the necessary theory of the STFT and the Gabor analysis, Section III presents the theory behind the present algorithm, Section IV contains a detailed description of the numerical algorithm. Finally, in Section V we present an extensive evaluation of the proposed algorithm and comparison with the iterative and non-iterative state-of-the-art algorithms using the Gaussian window, the truncated Gaussian window, the Hann and the Hamming windows.

II. Gabor analysis

The *short-time Fourier transform* of a function $f \in L^2(\mathbb{R})$ with respect to a window $g \in L^2(\mathbb{R})$ can be defined as[1]

$$(\mathcal{V}_g f)(\omega,t) = \int_{\mathbb{R}} f(\tau)\overline{g(\tau-t)}\mathrm{e}^{-\mathrm{i}2\pi\omega(\tau-t)}\,\mathrm{d}\tau, \quad \omega,t\in\mathbb{R} \tag{1}$$

$$=: M_g^f(\omega,t)\cdot e^{\mathrm{i}\Phi_g^f(\omega,t)}. \tag{2}$$

Using the modulation $\mathcal{E}_\omega f(\tau) := e^{2\pi i\omega\tau}\cdot f(\tau)$ and translation $\mathcal{T}_t f(\tau) := f(\tau - t)$ we get the alternative representation $(\mathcal{V}_g f)(\omega,t) = \langle f, \mathcal{T}_t\mathcal{E}_\omega g\rangle$.

The (complex) logarithm of the STFT can be written as

$$\log(\mathcal{V}_g f)(\omega,t) = \log M_g^f(\omega,t) + \mathrm{i}\Phi_g^f(\omega,t). \tag{3}$$

The *Gaussian function* is a particularly suitable window function as it allows an algebraic treatment of the equations. It is defined by the following formula

$$\varphi_\lambda(t) = \left(\frac{\lambda}{2}\right)^{-\frac{1}{4}}\mathrm{e}^{-\pi\frac{t^2}{\lambda}} = \left(D_{\sqrt{\lambda}}\varphi_1\right)(t), \tag{4}$$

where $\lambda\in\mathbb{R}^+$ denotes the "width" or the time-frequency ratio of the Gaussian window and D_α is a dilation operator such that $(D_\alpha f)(t) = \frac{1}{\sqrt{|\alpha|}}f(t/\alpha)$. We will use the following shorthand notation for $\lambda = 1$: $\varphi(t) = \varphi_1(t)$.

A. Discrete Gabor Transform

We define a Discrete Gabor Transform – DGT of a signal $f\in\mathbb{C}^L$ with periodic boundaries with respect to a window $g\in\mathbb{C}^L$ as [22], [23]

$$c(m,n) = \sum_{l=0}^{L-1} f(l)\overline{g(l-na)}\mathrm{e}^{-\mathrm{i}2\pi m(l-na)/M} \tag{5}$$

$$=: s(m,n)\cdot\mathrm{e}^{\mathrm{i}\phi(m,n)} \tag{6}$$

for $m = 0,\ldots,M-1$ and $n = 0,\ldots,N-1$ where $M = L/b$ is the number of frequency channels, $N = L/a$ number of time shifts, a is a hop factor in samples in time and b is a hop factor in samples in frequency. The overline denotes complex conjugation and $l - na$ is assumed to be evaluated modulo L. Redundancy of the DGT is given as $MN/L = M/a$. In the matrix notation, we can write $c = F_g^* f$, where F_g^* is a $MN\times L$ matrix.

The DGT can be seen as a discretization (of ω and t) and sampling (of f and g) of STFT of one period of L-periodic continuous signal f such that

$$c(m,n) = \left(\mathcal{V}_g f\right)(bm,an) + \mathcal{E}(m,n), \tag{7}$$

for $m = 0,\ldots,M-1$, $n = 0,\ldots,N-1$ where $\mathcal{E}(m,n)$ represents numerical errors and the aliasing (both in time and frequency) introduced by the sampling.

For real signals $f\in\mathbb{R}^L$, the range of m can be shrunken to the first $\lfloor M/2\rfloor + 1$ values as the remaining coefficients can be obtained by the complex conjugation. Moreover, coefficients $c(0,n)$ are real and so are $c(M/2,n)$ if M is even.

Signal f can be recovered (up to a numerical precision error) using the following formula

$$f(l) = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} c(m,n)\widetilde{g}(l-na)\mathrm{e}^{\mathrm{i}2\pi m(l-na)/M} \tag{8}$$

for $l = 0,\ldots,L-1$, where $\widetilde{g}$ is the canonical dual window. Again, in the matrix notation, we can write $f = F_{\widetilde{g}}c$, where $F_{\widetilde{g}}$ is a $L\times MN$ matrix. The canonical dual window can be obtained as

$$\widetilde{g} = \left(F_g F_g^*\right)^{-1} g. \tag{9}$$

See e.g. [23] for conditions under which the product $F_g F_g^*$ is (easily) invertible and [24] for an overview of efficient algorithms for computing (5), (8) and (9).

The discretized and periodized Gaussian window is given by

$$\varphi_\lambda^{\mathrm{D}}(l) = \left(\frac{\lambda L}{2}\right)^{-\frac{1}{4}}\sum_{k\in\mathbb{Z}}\mathrm{e}^{-\pi\frac{(l+kL)^2}{\lambda L}}, \tag{10}$$

for $l = 0,\ldots,L-1$. We assume that L and λ are always chosen such that the time aliasing is numerically negligible and $k\in\{-1,0\}$.

The (possibly non-integer) width of the Gaussian window in samples at the relative height $h\in[0,1]$ is given by

$$w_h = \sqrt{-\frac{4\log(h)}{\pi}\lambda L}. \tag{11}$$

Note that all the windows used in this manuscript are odd symmetric, such that they have just one center sample, and they are non-causal such that they introduce no delay. Ultimately, the discrete Fourier transform of such windows is real everywhere.

III. Theory Behind The Algorithm

It has been already mentioned that the algorithm is based on the direct relationship between the partial derivatives of the phase and the log-magnitude of the STFT with respect to the (dilated) Gaussian window. In this section, we derive such relations. We include a complete derivation because the

[1]In the literature, two other STFT phase conventions can be found. The present one is the most common one in the engineering community.

relations for STFT defined as (1) has never appeared in the literature.

It is known that the Bargmann transform of $f \in L^2(\mathbb{R})$

$$(\mathcal{B}f)(z) = 2^{\frac{1}{4}} \int_{\mathbb{R}} f(\tau) \mathrm{e}^{2\pi\tau z - \pi\tau^2 - \frac{\pi}{2}z^2} \,\mathrm{d}\tau, \quad z \in \mathbb{C} \tag{12}$$

forms an entire function for all $z \in \mathbb{C}$ and that it relates to the STFT defined in (1) such that [1]

$$(\mathcal{B}f)(z) = \mathrm{e}^{\pi \mathrm{i} t\omega + \pi\frac{|z|^2}{2}} (\mathcal{V}_\varphi f)(-\omega, t), \tag{13}$$

assuming $z = t + \mathrm{i}\omega$. Due to this the right-hand side satisfies the Cauchy-Riemann equations and so does its logarithm (away from zeros). The real and imaginary parts of $\log(\mathcal{B}f)(z)$ can be written as

$$\log(\mathcal{B}f)(t + i\omega) = u(\omega, t) + \mathrm{i}v(\omega, t) \tag{14}$$
$$u(\omega, t) = \pi(t^2 + \omega^2)/2 + \log(M_\varphi^f)(-\omega, t) \tag{15}$$
$$v(\omega, t) = \pi t\omega + \Phi_\varphi^f(-\omega, t) \tag{16}$$

and since it holds that

$$\tfrac{\partial u}{\partial t}(\omega, t) = \tfrac{\partial v}{\partial \omega}(\omega, t) \tag{17}$$
$$\tfrac{\partial u}{\partial \omega}(\omega, t) = -\tfrac{\partial v}{\partial t}(\omega, t) \tag{18}$$

we can write (assuming $\omega' = -\omega$) that

$$\tfrac{\partial \Phi_\varphi}{\partial \omega'}(\omega', t) = -\tfrac{\partial}{\partial t}\log(M_\varphi)(\omega', t) \tag{19}$$
$$\tfrac{\partial \Phi_\varphi}{\partial t}(\omega', t) = \tfrac{\partial}{\partial \omega'}\log(M_\varphi)(\omega', t) + 2\pi\omega'. \tag{20}$$

The following relations between the partial derivatives of the phase and the log-magnitude for the STFT with respect to the dilated Gaussian window φ_λ

$$\tfrac{\partial \Phi_{\varphi_\lambda}^f}{\partial \omega}(\omega, t) = -\lambda \tfrac{\partial}{\partial t}\log(M_{\varphi_\lambda}^f(\omega, t)) \tag{21}$$
$$\tfrac{\partial \Phi_{\varphi_\lambda}^f}{\partial t}(\omega, t) = \frac{1}{\lambda}\tfrac{\partial}{\partial \omega}\log(M_{\varphi_\lambda}^f(\omega, t)) + 2\pi\omega \tag{22}$$

were first published in [14] and they also appear in [25], [26]. The authors however use different STFT phase conventions and therefore the equations are slightly different.

A bit more general relationships can be obtained for windows defined as $g = \mathcal{O}\varphi_1$ ($\mathcal{O}$ being a fixed bounded operator) and Proposition 1.

Proposition 1. *Let $\mathcal{O}, \mathcal{P}$ be fixed bounded operators such that for all (ω, t) there exist (differentiable) functions $\eta(t)$ and $\xi(\omega)$, such that $\mathcal{T}_t\mathcal{E}_\omega\mathcal{O} = \mathcal{P}\mathcal{T}_{\eta(t)}\mathcal{E}_{\xi(\omega)}$ and let $g = \mathcal{O}\varphi_1$. Then*

$$\tfrac{\partial}{\partial \omega}\Phi_g^f(\omega, t) = -\tfrac{\partial}{\partial t}\log M_g^f(\omega, t) \cdot \frac{\xi'(\omega)}{\eta'(t)} \tag{23}$$

$$\tfrac{\partial}{\partial t}\Phi_g^f(\omega, t) = \tfrac{\partial}{\partial \omega}\log M_g^f(\omega, t) \cdot \frac{\eta'(t)}{\xi'(\omega)} + 2\pi\eta'(t)\xi(\omega). \tag{24}$$

Proof. Consider

$$\begin{aligned}\left(\mathcal{V}_g f\right)(\omega, t) &= \langle f, \mathcal{T}_t\mathcal{E}_\omega g\rangle = \langle f, \mathcal{T}_t\mathcal{E}_\omega\mathcal{O}\varphi_1\rangle \\ &= \left\langle \mathcal{P}^* f, \mathcal{T}_{\eta(t)}\mathcal{E}_{\xi(\omega)}\varphi_1 \right\rangle \\ &= \left(\mathcal{V}_{\varphi_1}\left(\mathcal{P}^* f\right)\right)\left(\xi(\omega), \eta(t)\right)\end{aligned}$$

Therefore

$$\begin{aligned}\tfrac{\partial}{\partial t}\Phi_g^f(\omega, t) &= \tfrac{\partial}{\partial t}\Phi_{\varphi_1}^{(\mathcal{P}^* f)}(\xi(\omega), \eta(t)) \\ &= \tfrac{\partial}{\partial \eta}\Phi_{\varphi_1}^{(\mathcal{P}^* f)}(\xi(\omega), \eta(t)) \cdot \eta'(t)\end{aligned}$$

and

$$\frac{\partial}{\partial \omega}\Phi_g^f(\omega, t) = \frac{\partial}{\partial \xi}\Phi_{\varphi_1}^{(\mathcal{P}^* f)}(\xi(\omega), \eta(t)) \cdot \xi'(\omega)$$

Futrhermore

$$\frac{\partial}{\partial t}\log M_g^f(\omega, t) = \frac{\partial}{\partial \eta}\log M_{\varphi_1}^{(P^* f)}(\xi(\omega), \eta(t)) \cdot \eta'(t)$$

and

$$\frac{\partial}{\partial \omega}\log M_g^f(\omega, t) = \frac{\partial}{\partial \xi}\log M_{\varphi_1}^{(P^* f)}(\xi(\omega), \eta(t)) \cdot \xi'(\omega).$$

So combining this with (20) we get

$$\begin{aligned}\tfrac{\partial}{\partial t}\Phi_g^f(\omega, t) &= \tfrac{\partial}{\partial \eta}\Phi_{\varphi_1}^{(P^* f)}(\xi(\omega), \eta(t)) \cdot \eta'(t) \\ &= \left(\tfrac{\partial}{\partial \xi}\log(M_{\varphi_1}^{(P^* f)}(\xi(\omega), \eta(t)) + 2\pi\xi(\omega)\right) \cdot \eta'(t) \\ &= \tfrac{\partial}{\partial \omega}\log M_g^f(\omega, t) \cdot \frac{\eta'(t)}{\xi'(\omega)} + 2\pi\xi(\omega)\eta'(t)\end{aligned}$$

The other equality can be shown using the same arguments using (19). □

By setting $\mathcal{O} = D_{\sqrt{\lambda}}$, $\xi(\omega) = \sqrt{\lambda}\omega$ and $\eta(t) = t/\sqrt{\lambda}$ we arrive at equations (21) and (22).

It should be noted that the relations for general windows were already studied in [27], they however involve partial derivatives of the logarithm of the modified Bargmann transform and thus it seems they cannot be exploited directly. Moreover, the experiments show that the performance degradation is not too significant when using windows resembling the Gaussian window.

The STFT phase gradient of a signal f with respect to dilated Gaussian φ_λ will be further denoted as

$$\nabla\Phi_{\varphi_\lambda}^f(\omega, t) = \left(\tfrac{\partial \Phi_{\varphi_\lambda}^f}{\partial \omega}(\omega, t), \tfrac{\partial \Phi_{\varphi_\lambda}^f}{\partial t}(\omega, t)\right). \tag{25}$$

A. Gradient Integration

Knowing the phase gradient, one can exploit the *gradient theorem* to go back to the original phase $\Phi_{\varphi_\lambda}^f(\omega, t)$ provided the phase at a single position $\Phi_{\varphi_\lambda}^f(\omega_0, t_0)$ is known such that

$$\Phi_{\varphi_\lambda}^f(\omega, t) - \Phi_{\varphi_\lambda}^f(\omega_0, t_0) = \int_0^1 \nabla\Phi_{\varphi_\lambda}^f\left(L(\tau)\right) \cdot \frac{\mathrm{d}L}{\mathrm{d}\tau}(\tau)\,\mathrm{d}\tau, \tag{26}$$

where $L(\tau) = [L_\omega(\tau), L_t(\tau)]$ is any line $(\omega_0, t_0) \to (\omega, t)$. Without the knowledge of $\Phi_{\varphi_\lambda}^f(\omega_0, t_0)$ one obtains

$$\widetilde{\Phi}_{\varphi_\lambda}^f(\omega, t) = \Phi_{\varphi_\lambda}^f(\omega, t) + \Phi_0, \tag{27}$$

where Φ_0 is a constant global phase shift.

The global phase shift of the STFT carries over to the global phase shift of the reconstructed signal trough the linearity of the reconstruction. One must however treat real input signals with care as the phase shift breaks the complex conjugate relation of the positive and negative frequency coefficients. Therefore if one simply takes only the real part of the reconstructed signal the phase shift causes its amplitude attenuation or even causes the signal to vanish in the extreme case. To explain this phenomenon, consider the following example where we compare the effect of the phase shift on analytic and on real signals. We denote the constant phase shift as ψ_0 and define an analytic signal as $x_{\text{an}}(t) = A(t)e^{i\psi(t)}$. The real part including the global phase shift ($e^{i\psi_0}$) is given as $\mathcal{R}e(x_{\text{an}}(t)e^{i\psi_0}) = A(t)\cos(\psi(t) + \psi_0)$ which is what one would expect. Similarly, we define a real signal as $x(t) = \frac{A(t)}{2}\left(e^{i\psi(t)} + e^{-i\psi(t)}\right)$ and the real part of such signal with the global phase shift ψ_0 amounts to $\mathcal{R}e(x(t)e^{i\psi_0}) = A(t)\cos(\psi_0)\cos(\psi(t))$ which causes the signal to vanish when $\psi_0 = \pi/2 + k\pi$, $k \in \mathbb{Z}$.

In theory, the global phase shift of the STFT of a real signal can be compensated for, leaving only a global signal sign ambiguity. For real signals, it is clear that the following holds for $\omega \neq 0$

$$\widetilde{\Phi}^{f}_{\varphi_\lambda}(\omega, t) + \widetilde{\Phi}^{f}_{\varphi_\lambda}(-\omega, t) = 2\Phi_0. \tag{28}$$

Due to the phase wrapping problem, after the compensation, the phase shift is still ambiguous such that $\Phi_0 = k\pi$ for $k \in \mathbb{Z}$, which causes the aforementioned signal sign ambiguity.

In practice, the phase shift is not constant and therefore (28) cannot be used easily. Therefore, when dealing with real signals, we reconstruct the phase only for the positive frequency coefficients and enforce the conjugate symmetry to the negative frequency coefficients.

IV. The Algorithm

The STFT phase gradient approximation $\nabla\phi(m, n) = \widehat{\nabla\Phi_{\varphi_\lambda}}(bm, an)$ is obtained by numerical differentiation of $s_{\log}(m, n) = \log\big(s(m, n)\big)$ (s from (6)) as

$$\nabla\phi(m, n) = \big(\phi_\omega(m, n), \phi_t(m, n)\big) := \tag{29}$$

$$\left(-\frac{\lambda}{a}(s_{\log}D_t^T)(m, n), \frac{1}{\lambda b}(D_\omega s_{\log})(m, n) + 2\pi m/M\right) \tag{30}$$

where D_t^T, D_ω denote matrices performing the numerical differentiation along rows (in time) and columns (in frequency) of $s_{\log}$ respectively. The matrices are assumed to be scaled such that the sampling step of the differentiation scheme they represent is equal to 1. The central finite difference scheme is the most suitable because it ensures the gradient components to be sampled at the same grid.

The steps of the numerical integration will be done in either horizontal or vertical directions such that only one component of the gradient is active at a time i.e. $\frac{\mathrm{d}L}{\mathrm{d}\tau}(\tau)$ from (26) is equal to $(0, 1)$ or $(1, 0)$. Due to this, the gradient can be pre-multiplied with length of the steps such that

$$\nabla\phi^{\text{SC}}(m, n) = \big(b\phi_\omega(m, n), a\phi_t(m, n)\big) := \tag{31}$$

$$\left(-\frac{\lambda L}{aM}(s_{\log}D_t^T)(m, n), \frac{aM}{\lambda L}(D_\omega s_{\log})(m, n) + 2\pi am/M\right). \tag{32}$$

Note that the dependency on L can be avoided when (11) is used to express λL. Further, the same formula can be used when searching for λL for non-Gaussian windows.

The numerical gradient line integration is performed adaptively using the simple trapezoidal rule as summarized in Alg. 1. The algorithm makes use of a heap data structure (from the heapsort algorithm [28]). In case of the present algorithm it is used for holding tuples (m, n) and it has the property of having (m, n) of the maximum $|c(m, n)|$ always at the top. It is further equipped with efficient operations for insertion and deletion. The effect of the parameter tol is twofold. First, a random phase is assigned to small coefficients for which the phase gradient is unreliable and second, the integration is done only locally on "islands" with the max. coefficient within the island serving as the zero phase reference. The randomization of the phase of the coefficients below the tolerance is chosen over the zero phase because it helps to avoid the impulsive disturbances introduced by the small phase-aligned coefficients.

After the phase $\widehat{\phi}(m, n)$ has been estimated, it is combined with the target magnitude of the coefficients such that

$$\widehat{c}(m, n) = s(m, n)\mathrm{e}^{\mathrm{i}\widehat{\phi}(m,n)} \tag{33}$$

and the signal is constructed using (8).

A. Practical Considerations

In this section, we analyze the sources of inaccuracies coming from the discretization.

The obvious sources of error are the numerical differentiation and integration schemes. However, the aliasing introduced by subsampling in time and frequency domains is more serious. In the discrete setting, since the signal is assumed to be band-limited and periodic, the truly aliasing-free case occurs when $a = 1, b = 1$ ($M = L, N = L$) regardless of the time or the frequency effective supports of the window. DGT with such setting is however highly redundant and only signals up to several thousands samples in length can be handled effectively.

In the subsampled case, the amount of aliasing and therefore the performance of the algorithm depends on the effective support of the window. Increasing a introduces aliasing in frequency and vice versa, increasing b introduces aliasing in time. Even though the Gaussian window is in theory infinitely supported in both time and frequency, it decays exponentially and therefore the aliasing might not significantly degrade the performance of the algorithm when choosing the hop factors and the effective support carefully.

Figure 1 shows that the algorithm performs very well in the aliasing-free case ($a = 1, b = 1$) and the performance becomes worse when bigger hop factors in time are introduced while keeping the effective width of the window constant. The length of the signal is 5888 samples and the time-frequency ratio of the Gaussian window is $\lambda = 1$. The sampling step in frequency was $b = 1$ ($M = 5888$). Only the values for the 60 dB range of the highest coefficients are shown.

B. Exploiting Partially Known Phase

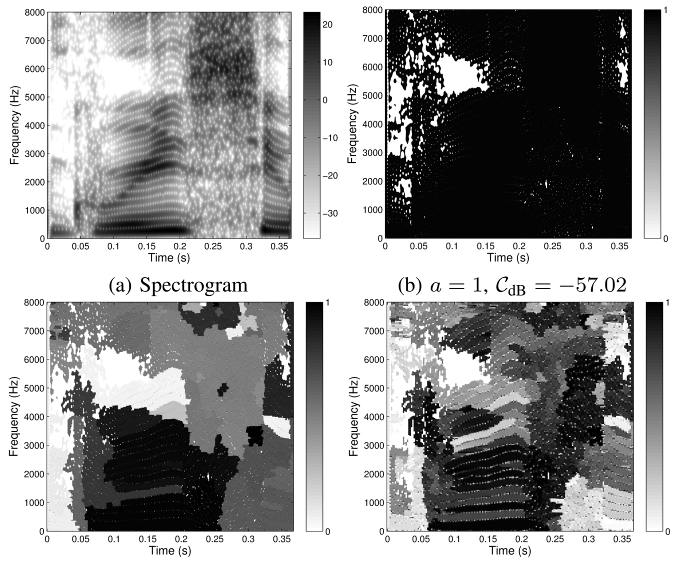
In some scenarios, the true phase of some of the coefficients or regions of coefficients is available. In order to exploit such

Algorithm 1: Phase gradient heap integration

```
Input: DGT phase gradient
       $\nabla\phi^{\mathrm{SC}}(m,n) = \left(\phi_\omega^{\mathrm{SC}}(m,n), \phi_t^{\mathrm{SC}}(m,n)\right)$ obtained
       from (32), magnitude of DGT coefficients
       $\left|c(m,n)\right|$, relative tolerance $tol$.
Output: Estimate of the DGT phase $\widehat{\phi}(m,n)$.
1  Create set
   $\mathcal{I} = \left\{(m,n) : \left|c(m,n)\right| > tol \cdot \max\left(\left|c(m,n)\right|\right)\right\}$;
2  Assign random values to $\widehat{\phi}(m,n)_{(m,n)\notin\mathcal{I}}$;
3  Construct a self-sorting heap for $(m,n)$ tuples;
4  while $\mathcal{I}$ is not $\emptyset$ do
5      if heap is empty then
6          Insert $(m,n)_{\max} = \arg\max\left(\left|c(m,n)_{(m,n)\in\mathcal{I}}\right|\right)$
           into the heap;
7          $\widehat{\phi}(m,n)_{\max} \leftarrow 0$;
8          Remove $(m,n)_{\max}$ from $\mathcal{I}$;
9      end
10     while heap is not empty do
11         $(m,n) \leftarrow$ remove the top of the heap;
12         if $(m+1,n) \in \mathcal{I}$ then
13             $\widehat{\phi}(m+1,n) \leftarrow$
               $\widehat{\phi}(m,n) + \frac{1}{2}\left(\phi_\omega^{\mathrm{SC}}(m,n) + \phi_\omega^{\mathrm{SC}}(m+1,n)\right)$;
14             Insert $(m+1,n)$ into the heap;
15             Remove $(m+1,n)$ from $\mathcal{I}$;
16         end
17         if $(m-1,n) \in \mathcal{I}$ then
18             $\widehat{\phi}(m-1,n) \leftarrow$
               $\widehat{\phi}(m,n) - \frac{1}{2}\left(\phi_\omega^{\mathrm{SC}}(m,n) + \phi_\omega^{\mathrm{SC}}(m-1,n)\right)$;
19             Insert $(m-1,n)$ into the heap;
20             Remove $(m-1,n)$ from $\mathcal{I}$;
21         end
22         if $(m,n+1) \in \mathcal{I}$ then
23             $\widehat{\phi}(m,n+1) \leftarrow$
               $\widehat{\phi}(m,n) + \frac{1}{2}\left(\phi_t^{\mathrm{SC}}(m,n) + \phi_t^{\mathrm{SC}}(m,n+1)\right)$;
24             Insert $(m,n+1)$ into the heap;
25             Remove $(m,n+1)$ from $\mathcal{I}$;
26         end
27         if $(m,n-1) \in \mathcal{I}$ then
28             $\widehat{\phi}(m,n-1) \leftarrow$
               $\widehat{\phi}(m,n) - \frac{1}{2}\left(\phi_t^{\mathrm{SC}}(m,n) + \phi_t^{\mathrm{SC}}(m,n-1)\right)$;
29             Insert $(m,n-1)$ into the heap;
30             Remove $(m,n-1)$ from $\mathcal{I}$;
31         end
32     end
33 end
```

information, the present algorithm has to be adjusted slightly. First, we introduce a mask to select the reliable coefficients and second, we select the border coefficients i.e. coefficients with at least one neighbor with unknown phase. Then we simply initialize the algorithm with the border coefficients stored in the heap.

Formally, the Algorithm 1 will be changed such that steps summarized in Algorithm 2 are pasted after line 3.



(a) Spectrogram (b) $a = 1$, $\mathcal{C}_{\mathrm{dB}} = -57.02$

(c) $a = 16$, $\mathcal{C}_{\mathrm{dB}} = -28.17$ (d) $a = 32$, $\mathcal{C}_{\mathrm{dB}} = -24.06$

Fig. 1: Spectrogram of a spoken word *greasy* (a). The absolute phase differences of the STFT of the original and reconstructed signal in rad$/\pi$ (modulo 1) for varying time hop size a (b) (c) (d).

Algorithm 2: Algorithm initialization for partially known phase

```
Additional input: Set of indices of coefficients $\mathcal{M}$ with
                  known phase $\phi(m,n)_{(m,n)\in\mathcal{M}}$.
1  $\widehat{\phi}(m,n) \leftarrow \phi(m,n)$ for $(m,n) \in \mathcal{M}$;
2  for $(m,n) \in \mathcal{M} \cap \mathcal{I}$ do
3      if $(m+1,n) \notin \mathcal{M}$ or $(m-1,n) \notin \mathcal{M}$ or
       $(m,n+1) \notin \mathcal{M}$ or $(m,n-1) \notin \mathcal{M}$ then
4          Add $(m,n)$ to the heap;
5      end
6  end
```

Note that the phase of the border coefficients can be used directly (that is not unwrapped). Depending on the situation, the phase might be propagated from more than one border coefficient but the phases coming from distinct sources are never combined.

C. Connections to Phase Vocoder

In this section we discuss some connections between the present algorithm and the phase vocoder [17] and consequently algorithms presented in [16] and [18].

The basic phase vocoder employs different analysis and synthesis steps in order to change the signal duration. The pitch change can be achieved by playing the signal at the sampling rate adjusted by the ratio of the steps. In the synthesis, the phase must be kept *consistent* in order not to introduce artifacts. In the phase reconstruction task, the original phase is not available, but the basic phase behavior can be exploited nevertheless. E.g. it is known that for a sinusoidal component with constant frequency the phase grows linearly in time for all frequency channels the component covers in the spectrogram. For these coefficients, the instantaneous frequency (STFT phase derivative with respect to time (22)) is constant and the local group delay (STFT phase derivative with respect to frequency (21)) is equal to zero.

In the aforementioned papers, the instantaneous frequency is estimated in each column (time frame) from the magnitude by peak picking and interpolation. The instantaneous frequency determines phase increments for each frequency channel m such that

$$\phi(m,n) = \phi(m,n-1) + 2\pi a m_0/M, \tag{34}$$

where m_0 is the estimated, possibly non-integer instantaneous frequency from the interval $\big(0, \lfloor M/2 \rfloor\big)$. This is exactly what the present algorithm does in case of constant sinusoidal components, except the instantaneous frequency is determined from the DGT log-magnitude. The integration in Alg. 1 does nothing else than a cumulative sum of the instantaneous frequency in the time direction.

The algorithm from [18] goes further and employs an impulse model. The situation is reciprocal to sinusoidal components such that the phase changes linearly in frequency for all coefficients belonging to an impulse component but the rate is only constant for fixed n and it is inversely proportional to the local group delay $n_0 - n$ (STFT phase derivative with respect to frequency (21)) such as

$$\phi(m,n) = \phi(m-1,n) + 2\pi a(n-n_0)/M, \tag{35}$$

where an_0 is the time instant of the impulse occurrence. Again, this is what the proposed algorithm does for coefficients belonging to impulses.

The advantage of the proposed algorithm over the other two is that the phase gradient is computed from the DGT log-magnitude such that it is available anywhere without even analysing the spectrogram content. This allows an arbitrary integration path which combines both the instantaneous frequency and the local group delay according to the magnitude ridge orientation. In the other approaches, the phase time derivative can be only estimated in a vicinity of sinusoidal components and, vice versa, the frequency derivative only in a vicinity of impulse-like components. Obviously, such approaches will not cope well with deviations from the model assumptions although careful implementation can handle multiple sinusoidal components with slowly varying inst. frequencies and impulses with frequency varying onsets. The difficulty of algorithm [18] lies in detecting the onsets in the spectrogram and separating the coefficients belonging to the impulse-like component from the coefficients belonging to sinusoidal components.

Figure 2 shows phase errors achieved by algorithms from [16], [18] and by the proposed algorithm. The phase difference at the transient coefficients is somewhat smoother for Alg. [18] when compared to [16] because of the impulse model. The present algorithm produces almost constant phase difference due to the adaptive integration direction. The setup used in the example is the following: the length of the signal is $L =$ 8192 samples, time hop size $a = 16$, number of channels $M = 2048$, time-frequency ratio of the Gaussian window is $\lambda = aM/L$. Only the values for the 50 dB range of the highest coefficients are shown.

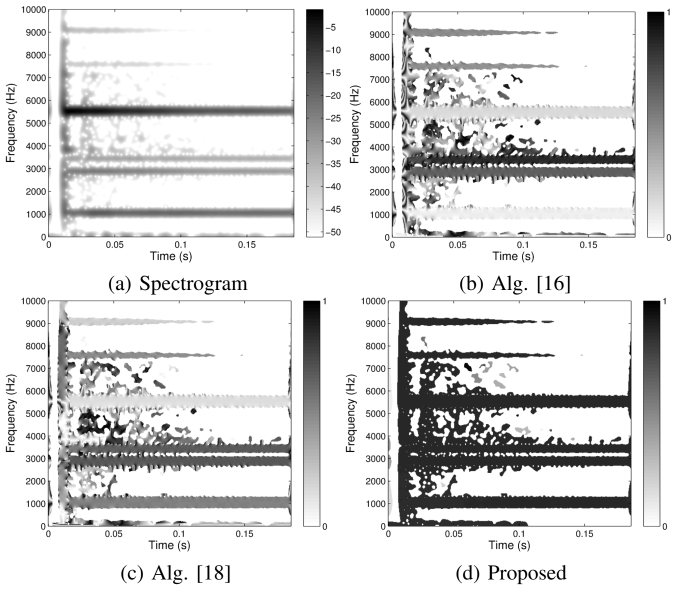


Fig. 2: Spectrogram of an excerpt form the *glockenspiel* signal and the absolute phase differences in rad$/\pi$ (modulo 1) for three different algorithms.

V. Experiments

In the experiments, we use the following equation to measure the error

$$E(x,y) = \frac{\|x-y\|_{\text{fro}}}{\|x\|_{\text{fro}}},\ E_{\text{dB}}(x,y) = 20\log_{10} E(x,y), \tag{36}$$

where x, y are either vectors or matrices. In [6] the *spectral convergence* is defined as

$$\mathcal{C} = E\left(s, |P\widehat{c}|\right), \tag{37}$$

where $P = F_g^* F_{\widetilde{g}}$. In [4] the authors proposed a slightly different measure $E(\widehat{c}, P\widehat{c})^2$ called *normalised inconsistency measure* defining normalised energy lost by the reconstruction/projection. Such measures clearly do not accurately reflect the actual signal reconstruction error $E(f, \widehat{f})$, but they are independent of the phase shift. Some other papers evaluate the algorithms using the signal to noise ratio, which is equal to $SNR(x,y) = 1/E(x,y)$ and $SNR_{\text{dB}}(x,y) = -E_{\text{dB}}(x,y)$ respectively.

Unfortunately, as the phase difference plots in Fig. 1 and Fig. 2 show, the phase difference is usually far from being constant when subsampling is involved (this holds for any algorithm, even the iterative ones) so the frames and even each frequency bin within the frame might have a different phase shift rendering the error $E(f, \widehat{f})$ to be very high even when the other error measures are low and the actual perceived quality is good.

The testing was performed on the speech corpus database MOCHA-TIMIT [29] consisting of recordings of 1 male

TABLE I: Comparison with SPSI, MOCHA-TIMIT database

	Gauss	Trunc. Gauss	Hann	Hamming
SPSI	-16.75	-16.75	-14.53	-14.02
Proposed	**-27.36**	**-27.37**	**-25.09**	**-24.87**

TABLE II: Comparison with SPSI, EBU SQAM database

	Gauss	Trunc. Gauss	Hann	Hamming
SPSI	-18.01	-18.19	-17.09	-16.79
Proposed	**-24.79**	**-24.71**	**-23.14**	**-22.66**

and 1 female English speakers each of which performing 460 sentences. The total duration of the recordings is 61 minutes and 5 seconds. The sampling rate of all recordings is 16 kHz. The Gabor system parameters used with this database (Table I and Fig. 3) were the following: number of channels $M = 1024$, hop factor $a = 128$, time-frequency ratio of the Gaussian window $\lambda = aM/L$, time support of the truncated Gaussian window and the other compactly supported windows was M samples.

Next, we used the EBU SQAM database of 70 test sound samples [30] recorded at 44.1 kHz. Only the first 10 seconds of the first channel was used from the stereophonic recordings to reduce the execution time to a reasonable value. The Gabor system parameters used with this database (Table II and Fig. 4) were the following: number of channels $M = 2048$, hop factor $a = 256$, time-frequency ratio of the Gaussian window $\lambda = aM/L$, time support of the truncated Gaussian window and the other compactly supported windows was M samples.

A. Performance Of The Proposed Algorithm

In this section, we evaluate the performance of the proposed algorithm and compare it to results obtained by the Single Pass Spectrogram Inversion algorithm [16] (SPSI). Unfortunately, we were not able to get good results with the algorithm from [18] consistently due to the imperfect onset detection and due to the limitation of the impulse model.

The implementation of SPSI was taken from http://anclab.org/software/phaserecon/ and it was modified to fit our framework. The most prominent change was the removal of the alternating π and 0 phase modulation in the frequency direction which is not present when computing the transform according to (5).

The results for the proposed algorithm were computed via a two step procedure. In the first step Alg. 1 with $tol = 10^{-1}$ was used, and in the second step, the algorithm was run again with $tol = 10^{-10}$ initialized with the result from the previous step according to Alg. 2. This approach avoids error spreading during the numerical integration and improves the result considerably when compared to a single run with either of the thresholds.

Tables I and II show the average $\mathcal{C}$ in dB over whole databases for the SPSI and the proposed algorithm. The proposed algorithm clearly outperforms the SPSI algorithm by a large margin. The performance of the proposed algorithm further depends on the choice of the window. While the Gaussian window truncation introduces only a negligible performance degradation, the choice of Hann or Hamming windows increase the error by about 2 dB. For a detailed comparison, please find the scores and sound examples for the individual files from the EBU SQAM database using the Gaussian window at the accompanying web page.

We can only provide a rough timing for the algorithms as the actual execution time is highly signal dependent and our implementations might be suboptimal. In the setup used in the tests, the runtime of the proposed algorithm is about 6-8 times longer than of the implementation of the SPSI algorithm. In particular, the current implementation of the proposed algorithm is very slow for noise-only signals.

B. Comparison With The State-of-the-art

We further compare the present algorithm with the following iterative algorithms:

- The Griffin-Lim algorithm [2] (GLA) as the baseline.
- A combination of Le Roux's modifications of GLA [4] and the fast version of GLA [5] with constant $\alpha = 0.99$ (FleGLA). More precisely from [4] we use the modification called *on-the-fly truncated modified update* which was reported to perform the best. The order of the on-the-fly phase updates is in the natural order of frames starting with the zero frequency bin within each frame. The projection kernel was always truncated to size $2M/a - 1$ in both directions.
  This combination outperformed both algorithms [4] and [5] when used individually.
- The gradient descend-like algorithm from [7] (lBFGS) with the refined objective function ($p = 2/3$). Unfortunately, the lBFGS implementation we use ([31]) fails in some cases.
- The real-time iterative spectrogram inversion algorithm with look-ahead (RTISI-LA). The algorithm was published in [3], but we implemented a refined version using the truncated projection kernel from [4] as proposed in [32]. The amount of the look-ahead frames was always $M/a - 1$ and an asymmetric analysis window was used for the latest look-ahead frame. Due to the nature of the algorithm, its performance can only be evaluated at M/a multiples of per-frame iterations.

As all the iterative algorithms optimize a non-convex objective function, the result depends strongly on the initial phase estimate. In addition to the zero phase initialization, we also evaluate performance of the algorithms initialized with the phase computed using the present algorithm. We will denote such initialization as *warm-start* (ws). Unfortunately, our tests showed that the RTISI-LA algorithm does not benefit from the warm-starting as it does its own initial phase guess from the partially reconstructed signal.

Figures 3 and 4 show average $\mathcal{C}$ in dB over the MOCHA-TIMIT and EBU SQAM databases respectively depending on the number of iterations without (solid lines) or with (dashed lines) the warm start. The horizontal dashed line is the average $\mathcal{C}$ in dB achieved by the proposed algorithm (values from Tables I and II). In addition, the scores and sound examples for individual files from the EBU SQAM database using the Gaussian window can be found at the accompanying web page.

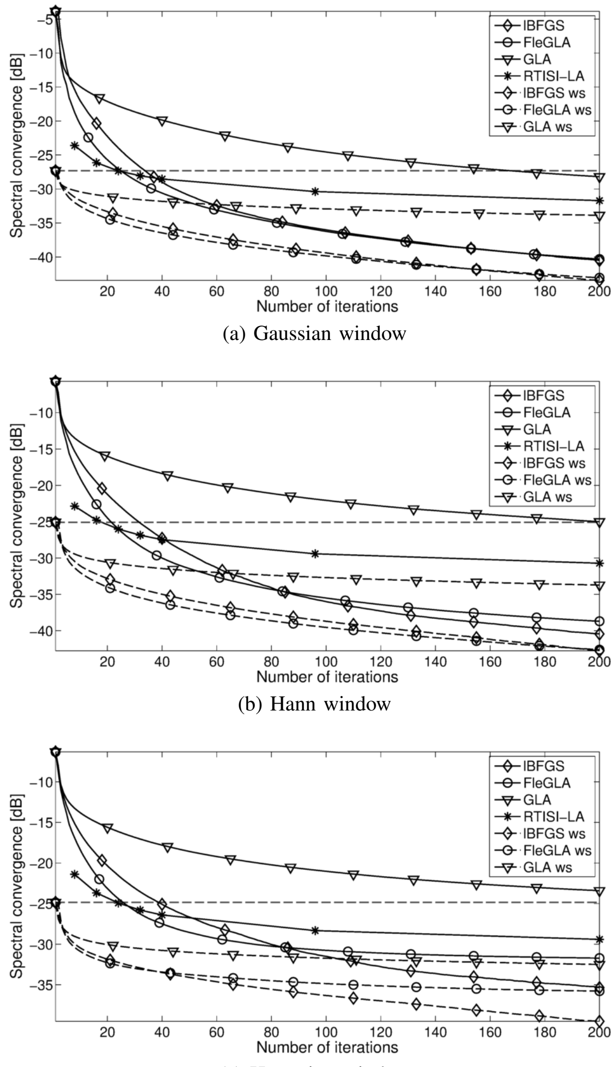
Graphs for the truncated Gaussian window are not shown because they exhibit no visual difference from the graphs for the full-length Gaussian window. Further, the lBFGS algorithm was excluded from the comparison using the EBU-SQAM database (Fig. 4); it failed to finish for a considerable number of the samples.

The graphs show that the proposed algorithm provides a suitable initial phase for the iterative algorithms i.e. the best overall results are obtained when combined with the FleGLA and lBFGS algorithms. When considering the iterative algorithms without warm-starting, the crossing points indicating number of iterations necessary to achieve the performance of the proposed algorithm can be clearly identified (if present). For the MOCHA TIMIT database (Fig. 3), the crossing point of the best algorithms is at about 20 iterations and the behavior is consistent for all the windows. The results for the EBU SQAM database (Fig. 4) are more erratic. First of all, the GLA algorithm never reaches the performance of the proposed algorithm even after 200 iterations. The crossing point of the FleGLA algorithm varies from 45 to 170 iterations depending on the window used. On the other hand, the RTISI-LA algorithm gets close to the line only for 8 per-frame iterations using the Gaussian window (Fig. 4a) and it performs better than the proposed algorithms in all the other cases. The RTISI-LA algorithm however “fails” for sound excerpts like castanets (crossing point at 80–120 it.), drums, cymbals and glockenspiel (crossing points 20–40 it.).

In the tests, the execution time of the proposed algorithm was equivalent to the execution time of 2–4 iterations of the GLA algorithm with the Gaussian window and to the execution time of 4–10 iterations for the compactly supported windows.

C. Modified Spectrograms

The main application area of the phase reconstruction algorithms is the reconstruction from the modified spectrograms. In general, the modified spectrogram is no longer a valid spectrogram (there is no signal having such spectrogram) and therefore the task is to construct rather than reconstruct a suitable phase. Unfortunately, it is not clear for which spectrogram modifications the equations (21) and (22) still hold nor how does it affect the performance if they don’t. Nevertheless, in order to get the idea of the performance of the proposed algorithm acting on modified spectrograms, we implemented phase vocoder-like pitch shifting (up and down by 6 semitones) via changing the hop factor ([17], [33]) using all the algorithms to rebuild the phase. The synthesis hop factor $a = 256$ was fixed and the analysis hop factor was changed accordingly to achieve the desired effect. Sound examples for the EBU SQAM database along with Matlab/GNU Octave script generating them can be found at the accompanying web page. According to our informal listening tests there is a little perceivable difference between the algorithms with the exception of SPSI. As expected, the SPSI algorithm introduces disturbing “echo-like” effects to sounds not conforming to the model assumption.



(a) Gaussian window

(b) Hann window

(c) Hamming window

Fig. 3: Comparison with the iterative algorithms, speech database.

VI. Conclusion

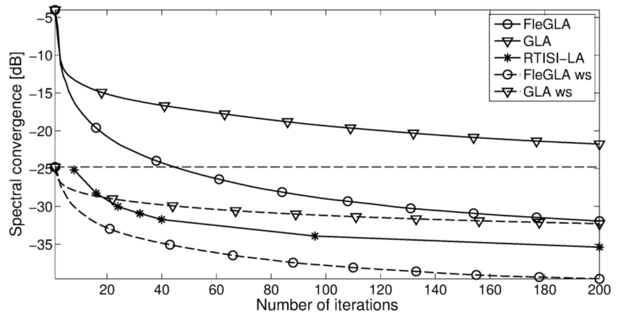
As a future work it would be interesting to investigate whether (simple) equations similar to (22) and (21) could be found for non-Gaussian windows. Moreover, the effect of the aliasing and spectrogram modifications on the phase-magnitude relationship should be systematically explored.

From the practical point of view, a big drawback of the proposed algorithm is the inability to run in the real-time setting i.e. to process streams of audio data in a frame by frame manner. Clearly the way how the phase is spread among the coefficients would have to be adjusted. This was done in [34] where we presented a version of the algorithm introducing one or even zero frame delay.

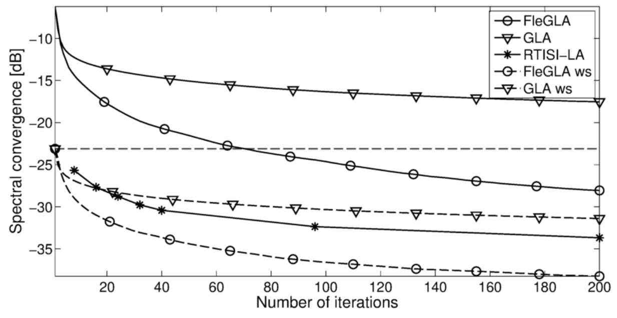
References

[1] K. Gröchenig, *Foundations of time-frequency analysis*, ser. Applied and numerical harmonic analysis. Boston, Basel, Berlin: Birkhäuser, 2001.

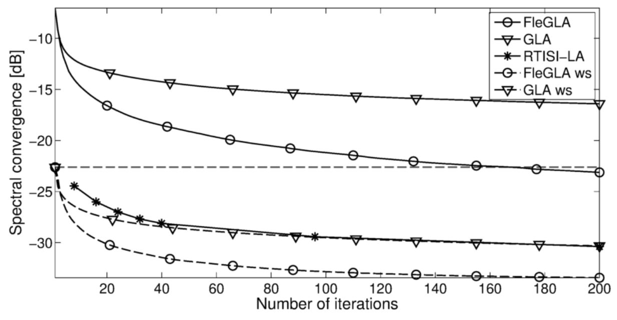
[2] D. Griffin and J. Lim, “Signal estimation from modified short-time Fourier transform,” *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 32, no. 2, pp. 236–243, Apr 1984.



(a) Gaussian window



(b) Hann window



(c) Hamming window

Fig. 4: Comparison with the iterative algorithms, SQAM database.

[3] X. Zhu, G. T. Beauregard, and L. Wyse, "Real-time signal estimation from modified short-time Fourier transform magnitude spectra," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 5, pp. 1645–1653, July 2007.

[4] J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama, "Fast signal reconstruction from magnitude STFT spectrogram based on spectrogram consistency," in *Proc. 13th International Conference on Digital Audio Effects (DAFx-10)*, Sep. 2010, pp. 397–403.

[5] N. Perraudin, P. Balazs, and P. Søndergaard, "A fast Griffin-Lim algorithm," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*, Oct 2013, pp. 1–4.

[6] N. Sturmel and L. Daudet, "Signal reconstruction from STFT magnitude: A state of the art," *Proc. Int. Conf. Digital Audio Effects DAFx*, vol. 2012, pp. 375–386, 2011.

[7] R. Decorsiere, P. Søndergaard, E. MacDonald, and T. Dau, "Inversion of auditory spectrograms, traditional spectrograms, and other envelope representations," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 23, no. 1, pp. 46–56, Jan 2015.

[8] I. Waldspurger, A. d'Aspremont, and S. Mallat, "Phase recovery, maxcut and complex semidefinite programming," *Math. Program.*, vol. 149, no. 1-2, pp. 47–81, 2015.

[9] E. J. Candes, Y. C. Eldar, T. Strohmer, and V. Voroninski, "Phase retrieval via matrix completion," *SIAM Review*, vol. 57, no. 2, pp. 225–251, 2015.

[10] D. L. Sun and J. O. Smith, III, "Estimating a signal from a magnitude spectrogram via convex optimization," in *Audio Engineering Society Convention 133*, Oct 2012.

[11] R. Balan, "On signal reconstruction from its spectrogram," in *Information Sciences and Systems (CISS), 2010 44th Annual Conference on*, March 2010, pp. 1–4.

[12] Y. Eldar, P. Sidorenko, D. Mixon, S. Barel, and O. Cohen, "Sparse phase retrieval from short-time Fourier measurements," *Signal Processing Letters, IEEE*, vol. 22, no. 5, pp. 638–642, May 2015.

[13] J. V. Bouvrie and T. Ezzat, "An incremental algorithm for signal reconstruction from short-time Fourier transform magnitude." in *INTERSPEECH*. ISCA, 2006.

[14] M. R. Portnoff, "Magnitude-phase relationships for short-time Fourier transforms based on Gaussian analysis windows," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '79.*, vol. 4, Apr 1979, pp. 186–189.

[15] B. Nouvel, "A study of a local-features-aware model for the problem of phase reconstruction from the magnitude spectrogram," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, March 2010, pp. 4026–4029.

[16] G. T. Beauregard, M. Harish, and L. Wyse, "Single pass spectrogram inversion," in *Digital Signal Processing (DSP), 2015 IEEE International Conference on*, July 2015, pp. 427–431.

[17] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *Speech and Audio Processing, IEEE Transactions on*, vol. 7, no. 3, pp. 323–332, May 1999.

[18] P. Margon, R. Badeau, and B. David, "Phase reconstruction of spectrograms with linear unwrapping: application to audio signal restoration," in *Proceedings of EUSIPCO*, Aug 2015.

[19] J. W. Eaton, D. Bateman, S. Hauberg, and R. Wehbring, *GNU Octave version 4.0.0 manual: a high-level interactive language for numerical computations*, 2015. [Online]. Available: http://www.gnu.org/software/octave/doc/interpreter

[20] P. L. Søndergaard, B. Torrésani, and P. Balazs, "The Linear Time Frequency Analysis Toolbox," *International Journal of Wavelets, Multiresolution Analysis and Information Processing*, vol. 10, no. 4, 2012.

[21] Z. Průša, P. L. Søndergaard, N. Holighaus, C. Wiesmeyr, and P. Balazs, "The Large Time-Frequency Analysis Toolbox 2.0," in *Sound, Music, and Motion*, ser. Lecture Notes in Computer Science. Springer International Publishing, 2014, pp. 419–442.

[22] P. L. Søndergaard, "Gabor frames by Sampling and Periodization," *Adv. Comput. Math.*, vol. 27, no. 4, pp. 355 –373, 2007.

[23] ——, "Finite discrete gabor analysis," Ph.D. dissertation, Technical University of Denmark, available from: http://ltfat.github.io/notes/ltfatnote003.pdf.

[24] ——, "Efficient Algorithms for the Discrete Gabor Transform with a long FIR window," *J. Fourier Anal. Appl.*, vol. 18, no. 3, pp. 456–470, 2012.

[25] T. J. Gardner and M. O. Magnasco, "Sparse time-frequency representations," *Proceedings of the National Academy of Sciences*, vol. 103, no. 16, pp. 6094–6099, 2006.

[26] F. Auger, E. Chassande-Mottin, and P. Flandrin, "On phase-magnitude relationships in the short-time Fourier transform," *Signal Processing Letters, IEEE*, vol. 19, no. 5, pp. 267–270, May 2012.

[27] E. Chassande-Mottin, I. Daubechies, F. Auger, and P. Flandrin, "Differential reassignment," *Signal Processing Letters, IEEE*, vol. 4, no. 10, pp. 293–294, 1997.

[28] J. W. J. Williams, "Algorithm 232: Heapsort," *Communications of the ACM*, vol. 7, no. 6, p. 347348, 1964.

[29] A. Wrench, "MOCHA-TIMIT: Multichannel articulatory database," 1999. [Online]. Available: http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html

[30] "Tech 3253: Sound Quality Assessment Material recordings for subjective tests," The European Broadcasting Union, Geneva, Tech. Rep., Sept. 2008. [Online]. Available: https://tech.ebu.ch/docs/tech/tech3253.pdf

[31] M. Schmidt, "minFunc: unconstrained differentiable multivariate optimization in Matlab," 2005, http://www.cs.ubc.ca/%7Eschmidtm/Software/minFunc.html.

[32] J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama, "Phase initialization schemes for faster spectrogram-consistency-based signal reconstruction," in *Proceedings of the Acoustical Society of Japan Autumn Meeting*, no. 3-10-3, Mar. 2010.

[33] U. Zoelzer, Ed., *Dafx: Digital Audio Effects*. New York, NY, USA: John Wiley & Sons, Inc., 2002.

[34] Z. Průša and P. L. Søndergaard, "Real-Time Spectrogram Inversion Using Phase Gradient Heap Integration," in *Proc. Int. Conf. Digital Audio Effects (DAFx-16)*, Sep 2016, to appear.