SIGNAL RECONSTRUCTION FROM STFT MAGNITUDE: A STATE OF THE ART

Nicolas Sturmel, Laurent Daudet*

Institut Langevin
ESPCI CNRS UMR 7587
10 rue Vauquelin 75005 Paris, France
firstname.lastname@espci.fr

ABSTRACT

This paper presents a review on techniques for signal reconstruction without phase, i.e. when only the spectrogram (the squared magnitude of the Short Time Fourier Transform) of the signal is known. The now standard Griffin and Lim algorithm will be presented, and compared to more recent blind techniques. Two important issues are raised and discussed: first, the definition of relevant criteria to evaluate the performances of different algorithms, and second the question of the unicity of the solution. Some ways of reducing the complexity of the problem are presented with the injection of additional information in the reconstruction. Finally, issues that prevents optimal reconstruction are examined, leading to a discussion on what seem the most promising approaches for future research.

1. INTRODUCTION

The ubiquitous Short Time Fourier Transform (STFT) is a very efficient and simple tool for audio signal processing, with a representation of the signal that simultaneously displays both its time and frequency content. The STFT computation is perfectly invertible, fast (based on the Fast Fourier Transform (FFT)), and provides a linear framework well suited for signal transformation. However, a majority of these modifications act on the magnitude of the STFT ; in this case phase information is lost, or at least corrupted. Source separation, for instance, is often based on the estimation of the time-frequency local energy of the sources, and the isolated sources are usually recovered trough Wiener filtering [1], *i.e.* with the phase of the original mixture. Other cases of adaptive filtering, like denoising [2], usually perform subtraction in the amplitude domain, once again not taking account of the phase of the signal. Signal modifications, such as time-stretching or pitch shifting [3], may also involve changes on the magnitude of the STFT (adding/removing frames, moving bins) without perfect knowledge of the expected structure of the phase. Although phase vocoder [4] brings some answers to the problem, the overall quality of the modification is still perfectible.

Furthermore, accurate reconstruction of a signal from its magnitude STFT is also of paramount importance in the domain of signal representation. Many works are addressing the relation between magnitude and phase of a Discrete Fourier Transform (DFT) [5, 6, 7]. Therefore, solving convergence issues of existing algorithms could also give ways of solving the problem of phase and magnitude dependency in the time-frequency domain. In short, being able to reconstruct a signal while only knowing its magnitude could bring significant improvements in many situations from source separation to signal modification.

Here, the key point is that the STFT has an important property: redundancy of the information. For a real signal, each length-N analysis window provides $N/2 + 1$ independent complex coefficients (keeping only components corresponding to positive frequencies), and with the additional constraint that the coefficients at frequencies 0 and $N/2$ are real by construction, this amounts to N real coefficients (in other words, the Discrete Fourier Transform is an orthogonal transform). However, with the STFT the analysis is always carried out with an overlap between adjacent analysis windows. In the case of minimal overlap of 50%, a real input signal of length N provides $2N$ real coefficients (neglecting here boundary effects). In the common case where the overlap is higher than 50%, this redundancy of information gets even higher. Similarly, the FFT can be oversampled in frequency (with zero-padding in time), providing more coefficients per frame.

This brings an important point: the STFT has to verify a so-called "consistency criterion" [8]. In other words, the set of complex STFT coefficients lives within a subset of the space $\mathbb{C}^{N\times M}$, but is not isomorphous to it: in general, an array of complex coefficients does not correspond to the STFT of a signal. Now, when keeping only the magnitude of the STFT, a real input signal of length N provides $N + 1$ real coefficients (with 50% overlap): phase reconstruction from magnitude-only spectrograms *may* still be possible [3]. The main issue is whether some crucial information has been lost by taking the magnitude, bringing ambiguities and/or ill-posedness issues. In the case of source separation, for instance, Gunawan showed [9] that phase reconstruction improved the quality of the separation. In the case of adaptive filtering, Le Roux showed [10] that the inconsistency criterion led to an improved estimation of the Wiener filter.

The goal of this article is to provide a state of the art in the problem of signal reconstruction from spectrograms (the squared magnitude of the STFT). Its goal is not only to review the benefits and drawbacks of each of the published methods, but also to discuss fundamental and sometimes open issues that make this problem still very active after decades of intense research. The article is organized as follows: the framework of the STFT will be presented in section 2, and the unicity of the representation will be discussed in section 3. The baseline technique for phase reconstruction, the so-called Griffin and Lim algorithm, will be presented in section 4 and quantification of the convergence will be discussed in section 5. Then, more recent reconstruction techniques will be presented: blind reconstructions in section 6 and informed ones in section 7. Finally, issues that arise when trying to achieve perfect reconstruc-

* This work was supported by the DReaM project (ANR-09-CORD-006) of the French National Research Agency CONTINT program.

tion of the signal will be discussed in section 8 and applications of such phase estimation to digital audio processing in section 9 prior to the conclusion of the document in section 10.

2. SHORT TIME FOURIER TRANSFORM

Let $x \in l_2(\mathbb{R})$ be a real, discrete signal, of finite support. On this support, we define the $STFT$ operator such that $S(n,m) = STFT[x]$ computed with an analysis window w of length N and an overlap $N-R$ (i.e., a hop size of R samples between consecutive analysis windows):

$$S(n,m) = \sum_{k=0}^{N-1} e^{-i2\pi\frac{kn}{N}} w(k)x(k+Rm) \quad (1)$$

Here, n is the frequency index, and m the time index. Inversion of this $STFT$ is achieved by the synthesis operator $STFT^{-1}$ described in equation (2) using the synthesis window s which gives the signal $\tilde{x}$:

$$\tilde{x}(l) = \sum_m s(l-mR) \sum_n S(n,m) e^{i2\pi n\frac{l-mR}{N}} \quad (2)$$

If the synthesis and analysis windows verify the energy-complementary constraint:

$$\sum_m w(l+mR)s(l+mR) = 1$$

then perfect reconstruction is achieved: $\tilde{x} = x$.

However, one might want to have more freedom in the choice of analysis / synthesis windows, and therefore the $STFT^{-1}$ operator must include a window ponderation such that $\tilde{x}(l) = \frac{1}{\tilde{s}(l)} STFT^{-1}[S]$, where $\tilde{s}(l) = \sum_m w(l+mR)s(l+mR)$ which is equivalent, up to boundary effects, to constraining the synthesis window to $\frac{s(l)}{\tilde{s}(l)}$. In [11], the inverse STFT is also described with the use of a vector formulation.

The different domains involved and the functions used to pass from one to another are presented on figure 1. The spectrogram W is the squared magnitude[1] of S and is given by $W = SS^*$ where S^* is the complex conjugate of S. Note that the spectrogram of a signal is also its autocorrelation and can be used as such for the interpolation of signals [12]. W is a set of real non-negative numbers $\in \mathbb{R}_+^{N\times M}$. The goal of the reconstruction is then to estimate $\tilde{S}(n,m)$ such that $\tilde{S} \in \mathbb{S}_{N,M}$, where $\mathbb{S}_{N,M}$ is the subset of $N \times M$ complex arrays representing co-called "consistent" STFTs, while keeping $\tilde{S}\tilde{S}^* = W$. Consistency of S is provided by the constraint $\mathcal{I}(S) = 0$, where $\mathcal{I}$ is defined by:

$$\mathcal{I}(S) = S - STFT[STFT^{-1}[S]] \quad (3)$$

In many applications such as the ones mentioned in the introduction, the array W used for reconstruction might not itself belong to the set of "consistent spectrograms" (the image of $\mathbb{S}_{N,M}$ by the operator $M \to |M|^2$). This might be due to the fact that the estimation of W is corrupted by noise (for denoising), or the cross-talk of other sources (for source separation), or because W is obtained through an imperfect interpolation algorithm (for time-stretching). In this case, there is no signal x that exactly verifies

[1] It should be noticed that some authors alternatively refer to spectrogram as the set S, i.e. the complex STFT coefficients

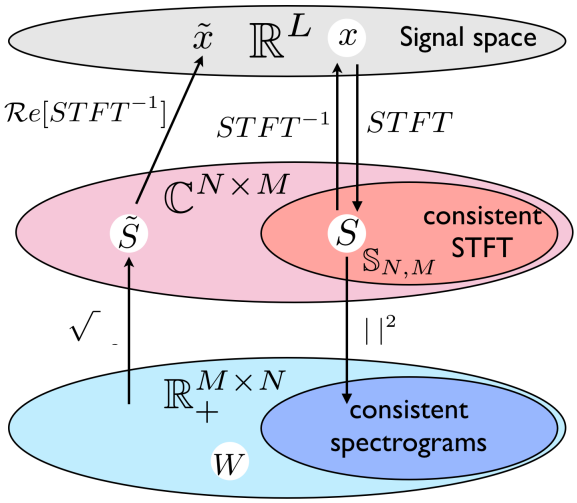

Figure 1: Domains involved when processing STFTs and spectrograms (expanded from [8]).

$S_x S_x^* = W$. There, the goal is to find the closest approximation, that minimizes the norm of $\mathcal{I}(S)$ (for some matrix norm, usually the Froebenius norm). In other words, one looks for the set $\mathbb{S}_{N,M}$ of consistent STFTs that verify $SS^* = W$.

Because we are specifically addressing a problem that uses compact STFTs, we discard techniques involving oversampling of each DFT [13, 14]: oversampling the DFT, while retaining the overlapping of the frames, introduces a redundancy of information that is too large to be handled in most practical cases. Signal reconstruction in those conditions can be considered solved by the previous studies even in the case of an isolated frame [15]. In this review, we will focus on techniques that, on the contrary, do not require specific constraints on the window design, the DFT oversampling, or hop size (we just assume that the STFT and inverse STFT are fixed and well-defined).

When trying to estimate the phase of an STFT from its magnitude only, some problems arise: the unicity of the representation [16, 12] discussed in section 3, how to quantify the convergence of the reconstruction (section 5), but also the tendency of reconstruction algorithms to catch local, non optimal, minima. A notable issue preventing optimal convergence is the so-called stagnation of the optimization [17] and will be discussed in section 8.

3. UNICITY OF THE REPRESENTATION

When addressing the problem of perfect reconstruction of a signal from its spectrogram, the first question that comes in mind is the unicity of the representation: can two different signals provide the same spectrogram ? The work of Nawab [12] produced some practical answers to the problem while only providing sufficient but not mandatory conditions to guarantee the unicity of x represented by $W(n,m)$. Some other works, such as [16] addressed signal uniqueness with the use of asymmetric windows ($w(n) \neq w(N-n)$), but such window is not suited for analysis of the spectrogram for the sake of phase linearity amongst other causes.

3.1. Sign indetermination

Some simple examples can be given to prove that unicity is not always verified. This is caused by the sign indetermination $|STFT[x]| = |STFT[-x]|$. Take for instance two signals x_1 and x_2 such as they do not overlap: $x_1 = 0$ outside $[N_{1A}; N_{1B}]$ and $x_2 = 0$ outside $[N_{2A}; N_{2B}]$ with $N_{1B} + N < N_{2A}$, then $x_1 - x_2$ and $x_1 + x_2$ have the same STFT $S(n, m)$.

Therefore, there are at least two signals x and $-x$ verifying the spectrogram W and the solution can only be unique under some constraints such as positivity of the signal (for instance in the case of image processing). But when this sign indetermination happens between big chunks of an audio signal, this case is either perceptually insignificant or can be countered by some simple knowledge on the signal.

However, it will be shown that this sign problem can happen locally in the reconstructed signal and regardless of its structure, this phenomenon is called *stagnation* by Fienup et al. [17] and will be discussed in section 8.

3.2. Conditions for the unicity of the reconstruction

The important conditions providing unicity in the case of a partial overlap, that is when hope size is $R > 1$, are given by Nawab [12]:

1. Known window function $w(n)$
2. Overlap of at least 50% ($R \leq \frac{N}{2}$).
3. Non zero window coefficients within the interval $[0; N]$
4. One sided signal, to define at least one boundary
5. Knowing R consecutive samples of the signal to be reconstructed starting from the first non-zero sample.
6. Less than R consecutive zeros samples in the signal.

Condition 1 of knowing $w(n)$ can be simply explained. This was illustrated by Le Roux in [18], with the example of designing an inconsistent STFT $H \in \mathbb{C}^{N \times M}$ so that $\sum |H| > 0$ but $STFT(STFT^{-1}(H)) = 0$ only for a given analysis/synthesis window pair. Since each analysis window has a different time-frequency smearing (see figure 6, in section 6), the information contained in the spectrogram is directly linked to w. This is especially true for inconsistent STFTs, of which the spectrogram is a particular case.

Condition 2 suggests that the amount of data contained by $|S(n, m)|$ is superior or equal to the one originally present in x, while condition 3 prevents missing informations due to zeros in w. Without any a priori on the signal, necessity of those two conditions seems rather natural. Enforcing regularity on the signal (like the techniques discussed in section 7) can lower those specific conditions.

Condition 4 imposes boundaries to the signal, allowing injection of some informations for the reconstruction, similar to the approach of Hayes and Quatiery in [19]. These boundaries were also used by Fienup et al. [17] but the support of an audio signal is too big in regard to the analysis window in order for such condition to be efficient. In fact, much more happens between the boundaries.

Since Nawab's work was based on successive interpolation of the signal, conditions 5 and 6 were established in order to know precisely the first R samples of the signal and continuously interpolating the signal without gaps. We feel that condition 5 is not always necessary, but condition 6 prevents sign indetermination problems like illustrated in section 3.1.

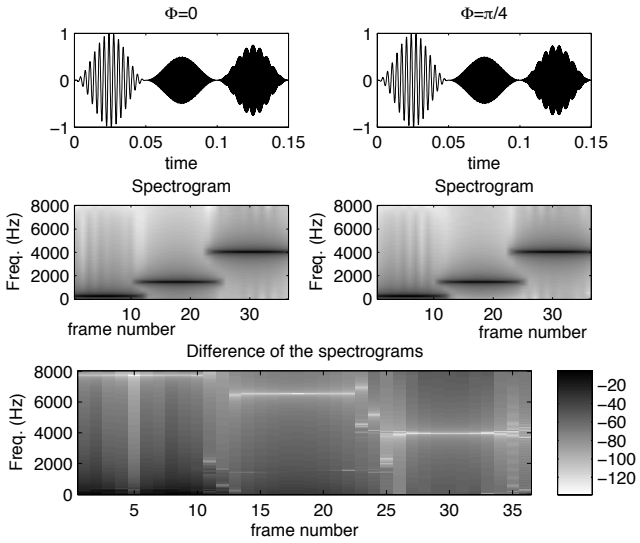

Figure 2: Spectrogram differences between two simple signals x_0 and $x_{\pi/4}$.

Some examples will be given throughout the paper in order to show that if the signal is not unique, it often comes down to the duality of the sign indetermination. We will also show in section 8 that greater issues are preventing the reconstruction and that unicity of the solution can be overlooked until those issues are solved. However those issues will often be linked to the unicity problem.

3.3. Phase rotation and spectrogram invariance

One common misconception about spectrogram is that it is phase invariant. Of course, if one were to work with complex signals, this phase invariance would be verified, but whether this still holds for real signals (whatever this means) is not so obvious.

For real signals, the only way to appropriately define the phase of the signals is within the framework of analytic signals. Let us assume that the signal x under study is the real part of a mono-component analytic signal H with slowly-varying amplitude $A(t)$: $x(t) = \mathcal{R}e(H) = A(t)\cos(\omega t)$, and let us construct the families of functions x_Φ for the same amplitude A and frequency ω, but with varying absolute phase Φ: $x_\Phi = \mathcal{R}e[He^{i\Phi}]$. If phase invariance were to hold, the spectrogram $|S_\Phi|^2$ of x_Φ would be the same as $|S_0|^2$ for any value of Φ.

Figure 2 shows the signal, spectrogram and absolute spectrogram difference of x_Φ for $\Phi = 0$ (left) and $\Phi = \frac{\pi}{4}$ (right) for three frequencies (300, 1500 and 4050Hz) at 16kHz sampling frequency and for an envelop A in the shape of a Hanning window with three different amplitudes (1, $\frac{1}{2}$ and $\frac{3}{4}$). The difference is computed as $|\ |S_0| - |S_{\pi/4}|\ |^2$. As one can see, this difference has an energy far from negligible.

Two interesting remarks can be made: first, the error is spread throughout the spectrum and not only in the vicinity of the signal's frequency. Second, this error is not either concentrated in time around the onset or offset of the tones: it can be shown as well that there is a similar error even when the amplitude of the signal stays constant.

Figure 3, shows the average spectrogram difference $\mathcal{C}(S_0, S_\Phi)$

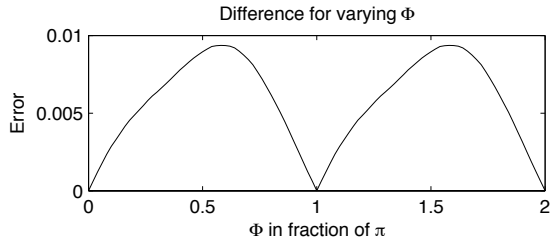

Figure 3: Spectrogram differences (equation (4)) for varying Φ in x_Φ.

as defined by

$$\mathcal{C}(S_0, S_\Phi) = \sqrt{\frac{\sum_{n,m} |\ |S_0(n,m)| - |S_\Phi(n,m)|\ |^2}{\sum_{n,m} |S_0(n,m)|^2}} \quad (4)$$

for varying Φ from 0 to 2π. One can see that the difference is π periodic due to the sign indetermination of $|S_\Phi(n,m)|$ and that most of the time it is inferior to 0.01 (i.e. -20dB).

This small experiment leads to the following rule of thumb: strictly speaking, the STFT is *not* phase invariant. However, when the computation is only made with low precision (less than 20 dB), the standard error criteria on the spectrogram don't "see" the phase. When minimizing this error, it appears that the original signal is indeed the true minimum but within a very flat surface. However, this fact that STFT is not strictly invariant to phase is good news: phase information seems to be present to some extent in the amplitude, but as a second-order effect. We shall see that this observation is the basis for discussion on the main issues making phase reconstruction such an intricate problem.

3.4. Perfect reconstruction

While the signal to be reconstructed from W is not necessarily unique, our goal is to find the most accurate reconstruction in regard to the original signal x. We call perfect reconstruction the estimation of the signal $\tilde{x}$ with an error of at most the measure error on x. If x is 16bits sampled, then the error power to achieve is approximatively equal to the quantification error power, that is to say approx. -90dB.

Moreover, we will consider perfect reconstruction as the estimation of x or $-x$. That is, we are implicitly discarding the global sign problem in the determination of x. We will show in section 8 that local indetermination of this sign can cause convergence issues.

4. ITERATIVE RECONSTRUCTION OF THE SIGNAL: THE GRIFFIN AND LIM BASELINE ALGORITHM

Based on the Gerchberg and Saxon algorithm [20], Griffin and Lim proposed the first global approach to solve the problem of signal reconstruction from spectrograms [3]. Due to the good perceptual results despite its simplicity for a basic implementation, this reconstruction algorithm remains the baseline for all subsequent work. Note that, as in the case of Gerchberg and Saxon reconstruction of the phase, uniqueness of the reconstruction is not guaranteed.

The approach from Griffin and Lim relies on a two-domain constraint, similar to the work of Hayes [15]. Before reconstruction, the spectrogram W of the STFT S is known but the phase

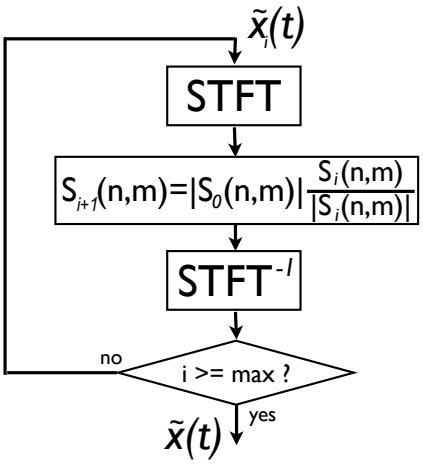

Figure 4: The iterative framework of Griffin and Lim [3]

$\angle S$ is unknown and can be initialized to 0 or at random values. In the spectral domain, absolute values of the estimated STFT $\tilde{S}_i$ are constrained to $|S_0| = \sqrt{W}$ at each iteration i, while the temporal coherence (as defined by equation (3)) of the signal is enforced by the operator $STFT[STFT^{-1}]]$.

The algorithm is presented on figure 4. First, it is initialized with $S_0 = \sqrt{W}$. At iteration i, the estimated STFT $\tilde{S}_i$ is computed and $\angle \tilde{S}_i$ is given to the original spectrogram so that the resulting time domain signal x_i is computed by inverse STFT of $|S_0| \frac{\tilde{S}_i}{|\tilde{S}_i|}$. In [3] it is shown that the mean square error between the STFT of the signal x_i and the estimated STFT of amplitude$|S_0|$ can be expressed as a distance:

$$d(S_0, \tilde{S}_i) = \sum_{n,m} |\ |S_0| \frac{\tilde{S}_i}{|\tilde{S}_i|} - \tilde{S}_i\ |^2 \quad (5)$$

and can be reduced to:

$$d(S_0, \tilde{S}_i) = \sum_{n,m} |\ |S_0| - |\tilde{S}_i|\ |^2 \quad (6)$$

It is also demonstrated that the gradient of d verifies $\Delta d(S_0, \tilde{S}_i) \leq 0$ and that this technique therefore reduces the distance d at each iteration.

This algorithm presents three main drawbacks:

1. First, its computation requires offline processing, as it involves computation of the whole signal at each iteration, and computation of both an STFT and an inverse STFT.
2. Second, convergence can be very slow, both in terms of computation time per iteration and by the number of iterations before convergence.
3. Finally, the algorithm does not perform local optimization to improve signal consistency, neither does it provide a consistent initialization of the phase from frame to frame.

Griffin and Lim's algorithm often provides time-domain signals that sound perceptively close to the original. However, depending on the sound material and the STFT parameters, some artifacts can be perceived: extra reverberation, phasiness, pre-echo... Indeed, while looking at the temporal structure of the reconstructed signals, we can see that they are often far enough from

the original to produce RMS error above 0dB. Although the corresponding sound quality may be sufficient in many cases, there are some application scenario where this may be a severe limitation. For instance, in the context of audio source separation, one may want to listen to the residual signal without the estimated source (karaoke effect): obviously a badly estimated time-domain signal prevents a correct source subtraction from the mix.

5. CONVERGENCE CRITERION

In order to assess the performance of the reconstruction, different criteria have been proposed. The most common ones are:

1. The spectral convergence $\mathcal{C}$, expressed as the mean difference of the spectrogram W with the absolute value of the reconstructed STFT $\tilde{S}$ as expressed by:

$$\mathcal{C} = \sqrt{\frac{\sum_{n,m} |\sqrt{W(n,m)} - \sqrt{\tilde{S}(n,m)\tilde{S}^*(n,m)}|^2}{\sum_{n,m} W(n,m)}} \tag{7}$$

   The convergence criterion $\mathcal{C}$ relates directly to the minimization process of Griffin and Lim's technique (equation (6)). This is the distance between the current coherent spectrogram and the target spectrogram. Then, when $\mathcal{C} = 0$, perfect reconstruction is achieved modulo unicity of the solutions.

2. The consistency $\mathcal{I}$ of the estimated STFT $\tilde{S}$ as given in equation (3). Again, $\mathcal{I} = 0$ means an accurate reconstruction, up to invariants.

3. The signal x to reconstruction $\tilde{x}$ root mean square error power:

$$\mathcal{R} = \sqrt{\frac{\sum (x(n) - \tilde{x}(n))^2}{\sum x(n)^2}} \tag{8}$$

   This criterion, analogous to the inverse of the signal-to-noise ratio, gives a better view of the reconstruction quality (we chose error over signal-to-noise ratio in order to observe the variations of $\mathcal{C}$ and $\mathcal{R}$ in the same direction). Note that the computation of $\mathcal{R}$ requires the knowledge of the original signal x. Therefore, it can only be used in (oracle) benchmarking experiments, and not in (blind) practical estimation. In this case, when $\mathcal{R} = 0$ the reconstruction is strictly equal to the original.

Obviously, the choice of the convergence criterion will have an effect on the discussion of the results obtained by each method. Even if $\mathcal{R} = 0$ is equivalent to $\mathcal{C} = 0$, one can easily find very small values of $\mathcal{C}$ associated with high values of $\mathcal{R}$.

Such issue is illustrated on figure 5 in the simple case of the DFT. The signal x used to compute figure 5 is a speech signal sampled at 16kHz and quantized on 16 bits. A random phase delay $\Phi(n)$ is computed, respecting the Hermitian symmetry ($\Phi(-n) = -\Phi(n)$), and making sure that this delay is always an integer in samples $\forall n$. Then, the phase of the DFT of x is shifted by $\Phi(n)$, multiplied with an integer factor k, with k ranging from 1 to 20. This is done through

$$\tilde{X}_k(n) = X_k(n) e^{ik\Phi(n)}$$

The resulting time-domain signal is called $\tilde{x}_k$. The two signals x and $\tilde{x}_k$ have the same energy ($XX^* = \tilde{X}_k\tilde{X}_k^*$), but are randomly delayed across frequencies. The figure displays the convergence criterion ($20\log\mathcal{C}$) and the reconstruction error ($20\log\mathcal{R}$) both in dB between signals x and $\tilde{x}_k$. Since there are two possible solutions (x and $-x$), $\mathcal{R}$ displayed on figure 7 is computed as $min(\mathcal{R}|_x, \mathcal{R}|_{-x})$. In this figure, one can see that the two criterions evolve separately. While $\mathcal{C}$ is staying at approx. $-14dB$, $\mathcal{R}$ is slowing rising to values above 0dB. This illustrates the fact that $\mathcal{C}$ may not be a good indicator of the reconstruction quality, with respect to the original signal.

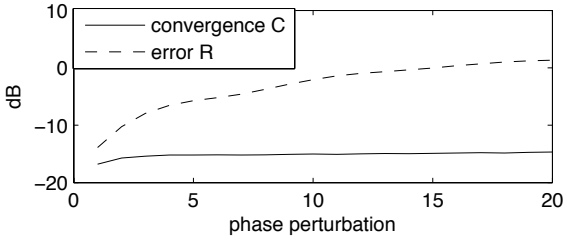

Figure 5: Difference between the $\mathcal{C}$ and $\mathcal{R}$ criteria used to evaluate the signal reconstruction, as a function of the amplitude of a random delay (integer in samples) on the DFT spectrum.

6. BLIND TECHNIQUES FOR SIGNAL RECONSTRUCTION

In this section, we review recent techniques that have been designed to improve Griffin and Lim's algorithm.

6.1. STFT consistency

STFT consistency of equation (3) can lead to the spectral domain only formulation of Griffin and Lim's least square estimation of the signal. In [8], an extensive work is presented to show how equation (3) can be used for the estimation of the phase of the corresponding coherent STFT. For instance, equation (10) gives a phase estimate Φ at each coordinate (n, m) of the STFT:

$$\mathcal{I}(n,m) = S(n,m) - STFT[STFT^{-1}[S(n,m)]]$$

$$\mathcal{I}(n,m) = \sum_{p=-\frac{N}{2}}^{\frac{N}{2}-1} \sum_{q=1-Q}^{Q-1} e^{i2\pi\frac{qn}{Q}} \alpha(p,q) S(n-p, m-q) \tag{9}$$

$$\Phi(n,m) = \angle\left(S(n,m) + \mathcal{I}(n,m)\right) \tag{10}$$

with $\alpha(p,q) = -\frac{1}{N}\sum_k \frac{w(k)s(k)}{\tilde{s}(k)} e^{-i2\pi p\frac{k+qR}{N}} + \delta_p\delta_q$

The term $\alpha(p,q)$ is the convolutive kernel applied to the STFT, that ensures both time domain (coordinate q) and frequency domain (coordinate p) coherence of the representation (this is the equivalent of the so-called "reproducing kernel" in wavelet analysis). This kernel is directly computed with the analysis and synthesis windows, and is invariant for the whole STFT. The shape of different kernels $\alpha(p,q)$ is given on figure 6 for four different window functions. The temporal dispersion of the kernel has a weak dependency on the window shape, but the frequency distribution is in direct relation to the spectral leakage of the window function [21].

The expression of $\mathcal{I}(n,m)$ given by equation (9) makes explicit the consistency criterion given in equation (3). This criterion is particularly efficient to provide information on the local coherence of the STFT as the phase correction depends directly on the value of $\frac{|\mathcal{I}(n,m)|}{|S(n,m)|}$. Equation 10 is also the direct application of

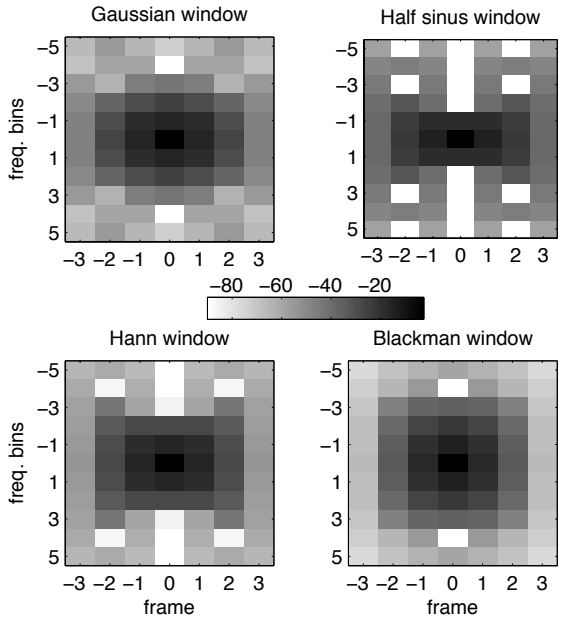

Figure 6: The influence of different windows on the STFT representation for an overlap of 75%. (amplitudes in dB)

Griffin and Lim optimization and follows the convergence of distance d defined in equation (6).

Additional studies in the same line [8] proposed solutions to lower the computation time while keeping a similar convergence speed. First, limiting the frequency domain span of the window α drastically lowers computation time while introducing only minimum error. When using analysis windows with low spectral leakage, one can reduce the term p of equation (9) to, for instance, the range $[-2; 2]$. This simplification significantly reduces the computation time, at the cost of a small error typically below 0.1%. Figure 6 presents some shapes of $\alpha(p, q)$ for different analysis and synthesis windows. We can see that the energy is concentrated around $(0, 0)$ especially for the half sinus window (used for the experiments in [8, 22]), allowing further approximation to the frequency bins around 0.

The second simplification is the use of sparseness of the signal in the time-frequency domain, in order to only update the bins of high energy. At each iteration, bins of lower and lower energy are updated. Empirical results shows that such simplification does not significantly modify the reconstructed signal $\tilde{x}$ at convergence, while drastically lowering computation time.

When using both simplifications, computation times given in [8] show a reduction by a factor 10 to 40 over the original Griffin and Lim iterative STFT reconstruction. This method improves convergence speed but does not significantly improve the final quality of the reconstruction. Note that both the computation time and the framework of this technique allow for real-time implementation, with minimal delay.

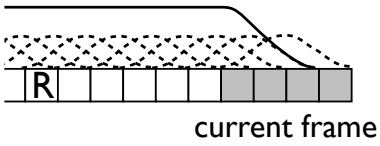

Figure 7: Real Time Iterative Spectrogram Inversion for an overlap of 75%.

6.2. Real-Time Iterative Spectrogram Inversion

The main drawback of Griffin and Lim's reconstruction algorithm is the processing of the whole signal for each iteration, preventing any use for on-line processing. Zhu and al. [23] proposed two implementations of the reconstruction, starting with a constraint of a real-time implementation.

First, the baseline Real Time Iterative Spectrogram Inversion (RTISI [24]) technique is based on the coherence of preceding reconstructions in regards to the frame begin reconstructed. This technique illustrated on figure 7, can be decomposed into two steps:

1. Consider the m-th frame S^m of the STFT $S(n, m)$ with its window function w_m and the signal $\tilde{x}_m$ which contains the weighted sum (equation (2)) of formerly processed frames. Then, $\angle S_0^m$ is initialized so that:

$$\angle \tilde{S}_0^m = \angle DFT[w_m \tilde{x}_m]$$

2. Then, the iterations are done as in Griffin and Lim, but restrained to frame m. At each step:

$$\begin{aligned} \angle \tilde{S}_i^m &= \angle DFT[w_m \tilde{x}_m + w_m \tilde{x}_{i-1}^m] \\ \tilde{x}_i^m(l) &= s(l) DFT^{-1}[\tilde{S}_i^m] \end{aligned}$$

This method is especially suited for multiple window length STFT, in a similar way to the window-switching method of MPEG 2/4 AAC coding [25]. However, RTISI offers results somewhat lower than Griffin and Lim's, mainly caused by the lack of look-ahead and optimization toward the *future* of the signal.

Therefore, a second method, RTISI with Look-Ahead (RTISI-LA [26]) was proposed. It is described by the scheme of figure 8. This method performs phase estimation of RTISI on k frames after the current one, ensuring that the estimated phase for the frame soon to be committed in the resulting signal $\tilde{s}$ is both in agreement with the past and future evolutions of the signal.

Convergence values $\mathcal{C}$ obtained for the RTISI-LA algorithm are usually better than the ones obtained with Griffin and Lim, but only in the order of 6dB of improvement. This improvement is mainly based on the emphasis on time coherence of the signal, as construction is done in both ways (forward and backward). Additional work from Gnann et al. [27] has focused on the phase initialization and processing order of the reconstruction. By processing the frame according to their energy and initializing the phase with unwrapping, one can improve the convergence of the reconstruction by 1 to 5dB.

Additional work from Le Roux [22] showed the same tendency when adding the phase initialization of RTISI-LA to the STFT consistency-based reconstruction.

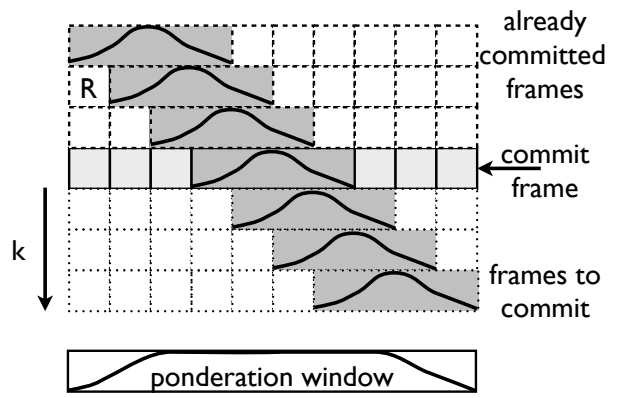

Figure 8: Real Time Iterative Spectrogram Inversion with Look Ahead of $k = 3$ and 75% of overlap.

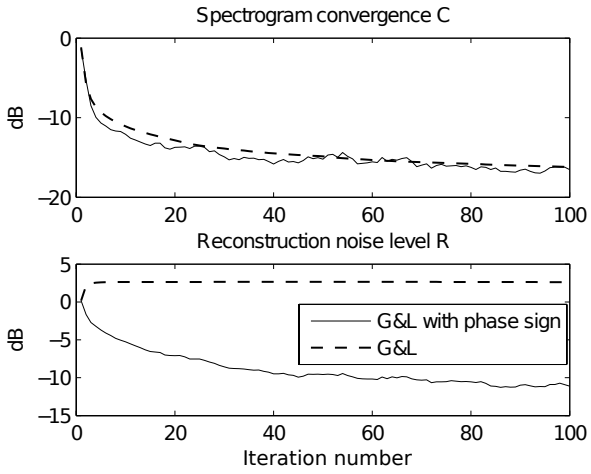

Figure 9: Convergence and reconstruction noise level for Griffin and Lim's method, with and without knowledge of the sign of the original phase.

6.3. Summary on existing techniques

Existing techniques are gradually introducing more and more constraints in the time domain, compared to the first approach of Griffin and Lim. They are still providing results that are close to the original spectrogram (convergence in the $\mathcal{C}$ criterium) but far from the original time domain signal. This tendency to generate incoherent signals in the time domain will be explained in the section 8 addressing fundamental issues shared by these current approaches.

Informal experiments were done using the initialization proposed in condition 5 and 6 of section 3 (knowledge of first samples of the signal) using the RTISI-LA technique. Unfortunately, this condition was not able to improve the reconstruction quality. Indeed, these conditions are neither necessary nor sufficient to perform perfect signal reconstruction, with both STFT coherence or real time spectrogram inversion.

7. INJECTING ADDITIONNAL INFORMATIONS

The three algorithms presented before do not show high accuracy in the reconstruction of the signal. Reconstruction errors $\mathcal{R}$ are often above zero, and rarely below $-6dB$. Therefore, injecting additional information on the signal could be a possible way to achieve a better reconstruction.

As perfect signal reconstruction involves very small variations on the spectrograms, much lower than the convergence values $\mathcal{C}$ usually obtained with the previous methods, one solution is to inject additional information during reconstruction. This information can be a prior on the shape of the signal, local phase information or shape criterion.

7.1. Additional knowledge on the signal spectrum

Alsteris and al. [5] have proposed an extended study on the possibility of reconstructing the signal while only knowing partial information of the spectrum and especially the knowledge of phase sign, phase delay or group delay. Moreover, when a prior on the position of the poles and zeros of the z-transform of each frame of the STFT is known, reconstruction can be made using the known relations between amplitude and phase of a DFT [7].

Phase sign, alternatively, has been shown to be a powerful addition to the spectrogram [28] in order to achieve a reconstruction of good quality for a very small amount of extra information (only one bit per bin). However, such information is not always available, especially in the case of blind source separation when the signal to be reconstructed is not known well enough. New approaches such as informed source separation could however benefit from the information of phase sign.

On figure 9, both convergence $\mathcal{C}$ and reconstruction noise $\mathcal{R}$ are shown for the Griffin and Lim reconstruction (512 samples half sinus window with 75% overlap) with or without knowledge of the phase sign. The test signal is a music sample of 2 seconds, sampled at 44.1kHz. One can see that phase sign does not improve the convergence speed of the algorithm in terms of $\mathcal{C}$, but dramatically enhances the quality of the reconstruction, as $\mathcal{C}$ and $\mathcal{R}$ become strongly correlated. Perceptively, transients are better reconstructed with less smearing and artifacts.

However, as shown with this example, sign information does not seem sufficient to achieve perfect reconstruction in practice, as the reconstruction noise levels $\mathcal{R}$ remain high even after 100 iterations. However, convergence could probably be faster while using this prior on phase sign for proper initialization of the algorithm.

7.2. Probabilistic inference

Another idea that has been explored is to use some statistical properties of the signal. The work proposed by Achan [29] uses an autoregressive model of the speech signal to be reconstructed, in order to improve the convergence of the algorithm. As mentioned in the article, the proposed method performs only slightly better than the classic Griffin and Lim (approx. 2 to 4dB depending on the model) and resorts to a posteriori regularization of the signal. This can however be an interesting approach when the class of signals to be recovered is well defined. Also, the idea beneath this technique is interesting, as concurrent optimization is done both in the time and STFT domain, whereas blind techniques only constrain the STFT domain.

7.3. Local observations

Spectrograms also possess local properties that can be extracted with or without a prior in order to recover the original signal:

Nouvel [30] proposed the iterative estimation of local patterns of the time-frequency diagram, patterns based on a polynomial expression of the phase, for instance. The algorithm proposed performs better than Griffin and Lim only when there is no overlap. Missing information is then brought to the reconstruction by the prior learning of the polynomial coefficients.

Another approach is the Max-Gabor analysis of spectrograms from Ezzat et al. [31]. It uses local patch of the spectrogram where local amplitude, frequency, orientation and phases are estimated. The information are used in order to synthesize the time-domain signal with Gabor functions. Unfortunately this study does not address the quality of the reconstruction by comparing it to Griffin and Lim as it was not aimed originally at the task of phase recovery.

7.4. Conclusion: usefulness of additional information

In this section we presented some recent techniques that perform signal reconstruction from spectrogram while having additional informations on the signal to reconstruct. We saw that despite some advanced models, the proposed algorithms are only slightly better than the original framework from Griffin and Lim, especially in terms of the time-domain $\mathcal{R}$ error criterium.

Even when using the sign of the STFTs, Griffin and Lim algorithm does not convergence faster, nor better: only the quality (SNR) improves. This proves that most of the work to improve the convergencel has to be done on the reconstruction algorithms themselves, as additional information only serves at improving the final quality. The issues that are preventing the convergence despite the additional information are discussed in the next section.

8. OVERLOOKED ISSUES

As far as the state of the art goes, a number of issues regarding signal reconstruction from spectrograms seem overlooked. One of them is the use of the convergence criterion $\mathcal{C}$ which requires extremely high convergence (difference of approx. -90dB) in order to achieve a perfect reconstruction of the signal. Other issues are caused by the way information is spread in the spectrogram or by the minimization technique of the reconstruction itself.

8.1. Phase information and spectrogram

The first major issue of signal reconstruction from spectrograms is the effect of phase information in the modulus of the STFT. Because the STFT is obtained via windowing, one can find at bin n the contribution of many spectral components added to one another, thus forming a linear system [13, 14, 32]. However, such system only finds a suitable solution under three precautions:

1. The analysis window has to produce a lot of spectral leakage. The Gaussian window is a good example of such window and is often used.
2. The overlap has to be very high, in order to provide as little time downsampling as possible in every frequency channel.
3. Usually DFT are oversampled, bringing yet another layer of redundancy in the STFT

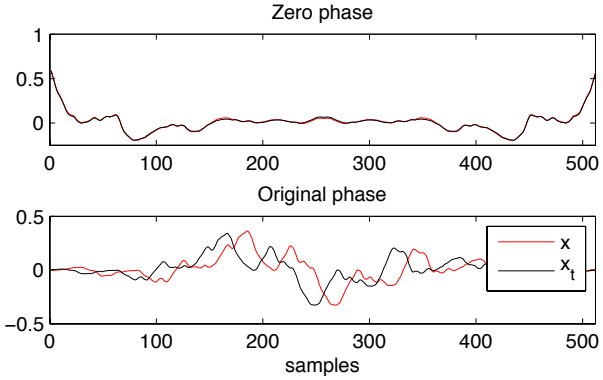

Figure 10: Spectrogram amplitude difference with or without phase for two signal x and x_t. x_t is the signal x translated 20 samples to the left. Spectral difference $\mathcal{C}$ between the two frames is -25dB.

In real analysis conditions, when using windows with a low spectral leakage and a rather low overlap (usually 50% or 75%), such an analytic resolution of the system is not possible, mainly due to the precision of both the data contained in the STFT and the complexity of the system to solve.

One example is given on figure 10 where the same frame of two different STFTs of a speech signal sampled at 16kHz and quantized on 16 bits are displayed: in red, the frame inverse DFT of a frame of the STFT of original signal x and in black the same frame of the STFT of x_t, the signal x shifted by 20 samples to the left. On the top row, the inverse DFTs are presented with zero phase (magnitude only) and on the bottom row the time-domain inverse DFTs with the original phase information are given. Despite the vast difference between the two frames, the zero phase responses are very similar (differences are barely visible around samples 160 and 350). Difference $\mathcal{C}$ of the two signals on the top row of figure 10 is -25dB, approx. the convergence limit of Griffin and Lim's technique. Although this figure is a good example of the poor effect of phase on the magnitude of the STFT, it will also serve well the illustration of stagnation by translation given later.

8.2. Stagnation caused by sign indetermination

Fienup et al. [17] proposed an interesting study on the problems preventing iterative algorithms such as Griffin and Lim's to converge toward a unique solution. It described this issues as *stagnation*, a self explanatory term that illustrates the inability of the algorithm to converge toward an optimal solution because it reached a local minima of optimization. Although Fienup's work was based on image processing, two of the three stagnations described in [17] can very well be observed on one-dimensional signals.

The first stagnation is linked to the sign indetermination illustrated in section 3. During reconstruction, the algorithm can be stuck between a mix of the two possible solutions x and $-x$, because it converged toward features of both signals. This phenomenon is illustrated on figure 11. On this figure, one can see that at the beginning the estimated signal $\tilde{x}$ is in phase with x whereas at the end it is in phase with $-x$. On the middle on the figure, one can see a characteristic point when $\tilde{x}$ gets closer to zero, illustrating an inflection point from one frame to another. Note that the

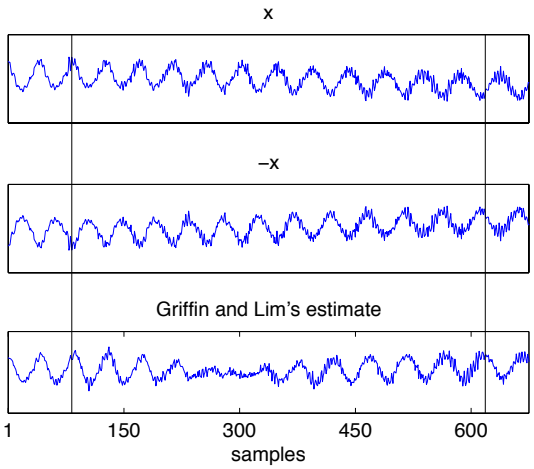

Figure 11: The first stagnation: the algorithm estimation (bottom) is stuck between a mix of x (top) and $-x$ (middle). Estimation with an half sinus window of 512 sample long, overlap of 75%.

difference between the two local minima is approximately equal to the window size. Such stagnation is also observed on signals reconstructed with RTISI-LA and the STFT coherence. Moreover, this stagnation is not consistent along the frequency axis: a closer look to the signal presented on figure 11 shows that phase coherence toward x or $-x$ is only true for the first harmonic.

This first stagnation is countered by the knowledge of the sign of the STFT presented in section 7 and is the main cause of the very high noise estimation levels $\mathcal{R}$ observed when reconstructing a signal with either of the three method presented in section 6. Basically, knowing the sign of the STFT causes the uniqueness of the solution to be true, avoiding a lot of local minima during minimization.

An other stagnation, also explained by the sign indetermination is what Fienup called "fringes". Sadly, this observation is hard to make on audio signals but is still present during the reconstruction. Because of this sign indetermination $|DFT[x(-n)]| = |DFT[x(n)]|$, frames happen to be estimated in the wrong time direction. Most of the times, overlap is enough to prevent such stagnation, which is then the most unlikely to happen.

Solutions to overcome these stagnations proposed in [17] do not apply well to signal processing, as they were designed for image processing. However, the idea of Monte Carlo method and artificial boundaries of the reconstruction seem interesting and easily transposable to the signal domain.

8.3. Stagnation caused by translation

The third stagnation is the translation of the signal. Because the TFD operator is circular, translation of the signal does not always drastically change the magnitude of the transform (figure 10) despite the windowing. Therefore, convergence can happen to a translated version of the original signal: like in figure 12 where a signal and its reconstruction with Griffin and Lim's technique are presented. This problem can be linked to the phase rotation problem addressed in section 3 but on local portions of the signal.

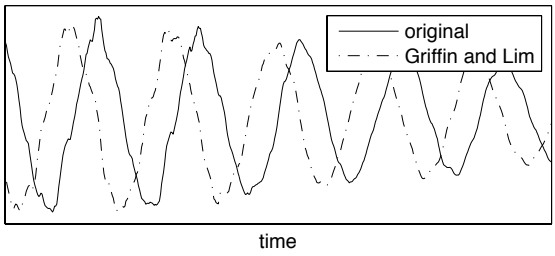

Figure 12: Stagnation by translation for a Griffin and Lim reconstruction (half sinus 512 sample window, 75% overlap, 200 iterations)

8.4. Different stagnation per frequency band

An other issue of the stagnation is that it happens at different levels on different frequency bands. Because the coherence of the STFT is limited (surfaces of figure 6) in both time and frequency, a gap in energy can cause different patchs of the reconstructed STFT to present different kinds of stagnation. As music signal often presents harmonic structure or colored noise, localized energy is very common.

An illustration of this phenomenon is given on figure 13 where a speech signal (the original, 16bits and 16kHz, at approx. 200Hz fundamental) and its reconstruction from its spectrogram (Griffin and Lim, 512 sample half sinus window, 200 iterations) are showed for different frequency band. The filter bank presents a passing band of 400Hz and a zero phase to prevent delay to be inserted between observations.

On the two first bands, from 1600 to 4600Hz, the signal is well reconstructed and is mainly presenting a small stagnation by translation. However, the direction of the translation is not the same for the two bands.

On the bands three to five, one can mainly see a stagnation by sign indetermination with characteristic inflection points based around samples 2300 and 2460 for band 3, 2375 for band 4 and 2325, 2495 for band 5. Once again, even if the bands are presenting the same type of stagnation, their evolution is different, mainly dependent on the local frequency.

It can be noted that, as expected, this stagnation issue gets more and more problematic as frequency increases. At low frequencies, the overlap between adjacent windows represents a smaller phase increment than at high frequencies. This may give an insight on why standard phase reconstruction offers a rather good sound quality despite a low SNR: at high frequencies, the ear is not so sensitive to the phase but rather to the general energy in the frequency bands. It may also indicate that algorithms based in the injection of additional information should have different tradeoffs in terms of precision versus amount of extra information, in different frequency bands.

9. APPLICATIONS TO DIGITAL AUDIO PROCESSING

In the case of source separation in a linear instantaneous stationary mixture, one often knows partial information on the source to be reconstructed, such as its spectrogram (or corrupted spectrogram). In this case, Gunawan [9] proposed a framework in order to use the information contained in the mixture M_x of N sources to help the

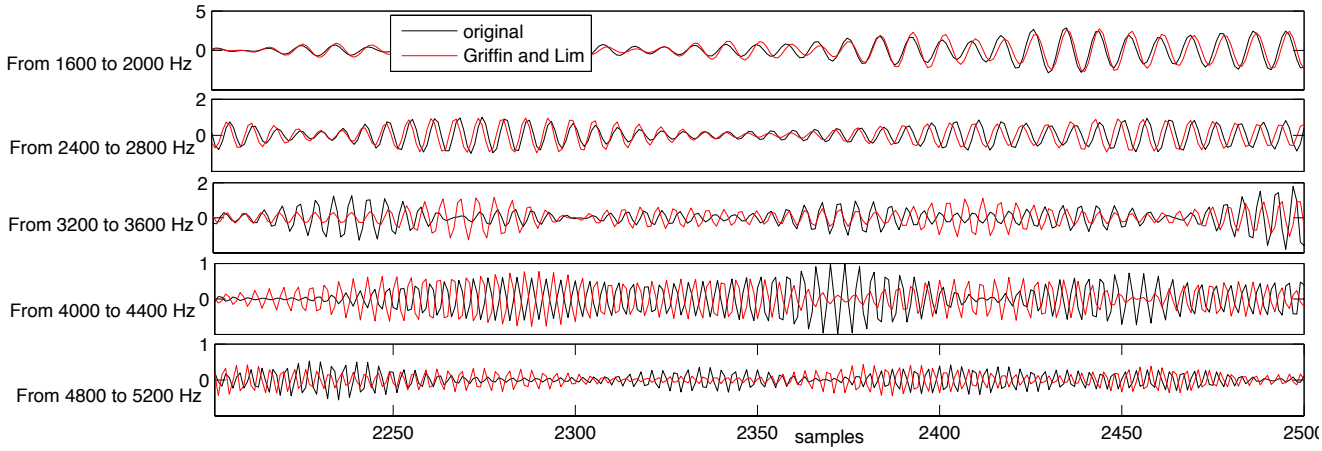

Figure 13: Signal Comparison (original in black, Griffin and Lim's reconstruction in red) for different frequency bands (zero phase filter bank). Stagnation are not consistent across frequency

phase estimation. While constraining the spectrogram W_j of the j-th source, one can reconstruct its phase with the following steps:

$$\hat{S}_j^{k+1} = \sqrt{W_j} e^{i\angle STFT(STFT^{-1}(S_j^k + \frac{e^k}{N}))} \quad (11)$$

$$e^{k+1} = M_x - \sum_j \hat{S}^{k+1} \quad (12)$$

This way, stagnations such as sign indetermination or translation are automatically compensated by the error computed on equation (12). The phase of the mixture is used as an additional information to constrain the reconstruction. Of course, this study provides the best results when the target spectrogram of each source W_j is perfectly known, while in practice the target spectrogram is only an estimate. Results are also conditioned by the number N of sources, with the best results for only 2 sources.

An other study [10] proposed by Le Roux used the spectrogram consistency (the fact that $S = STFT(STFT^{-1}S)$) as a constraint for the maximum likelihood estimation of a Wiener filter α_j for the j-th source. Such filters are used to perform adaptive filtering (for instance, in denoising) but usually rely on the energy ratio between the sources:

$$\alpha_j(n,m) = \frac{W_j(n,m)}{\sum_k W_k(n,m)} \quad (13)$$

$$\hat{S} = \alpha M \quad (14)$$

By explicitly adding the constraint that

$$\hat{S}_j - STFT(STFT^{-1}\hat{S}_j) = 0$$

into the equation, results show an improvement in SNR of around 3dB when applied on speech denoising.

10. SUMMARY AND CONCLUSION

In this paper we presented a state of the art on the question of signal reconstruction from spectrogram. We especially addressed the problem of perfect reconstruction and the issues preventing existing algorithms from converging to one (or one of the possible) solution.

Unicity is an important question to be asked in this case, but ordinary conditions are sufficient to guarantee that there is no more than two possible solutions for the reconstruction, given by the sign indetermination of the magnitude operator. Still, we saw that duplicity of the solution is the cause of the stagnation of the minimization by sign indetermination.

The three current techniques of blind reconstruction (Griffin and Lim, RTISI-LA and STFT coherence) have been described and discussed. Although there has been more than 20 years between Griffin and Lim's and the two other techniques, overall reconstruction quality has not significantly improved. Of course, computation time and implementation (especially in the case of real-time processing) have been a huge development part of such techniques, but we feel that most of the work has yet to be focused on the actual process leading to the optimal convergence of the algorithm in order to get better than just perceptively close reconstructions.

Given the amount of information present in the spectrogram, especially with the typical value of 75% overlap, perfect reconstruction (i.e. reconstructing x from $|STFT[x]|$ with error inferior the measure error on x itself) should be possible. We raised however a number of issues preventing convergence of the reconstruction toward the absolute minima. Those issues, called stagnation by Fienup [17] are configurations that prevent further minimization of the error. Stagnation presented are of two types: stagnation by sign indetermination (time inversion and signal inversion) and stagnation by translation. Because music signals are not evenly distributed on the time-frequency plan, stagnation can occur independently on local patches of the spectrogram both in time and frequency and is therefore difficult to correct.

Future work should then emphasize the resolution of the stagnation problems highlighted in this article, either with side information or using blind reconstruction. Whereas solving the problem of sign indetermination should be rather simple as one can observe sign coherent patches in the reconstructed STFT, phase translation if more problematic as it produces time delay that varies for the whole time-frequency domain.

11. REFERENCES

[1] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 191 – 199, 2006.

[2] Guoshen Yu, S. Mallat, and E. Bacry, "Audio denoising by time-frequency block thresholding," *IEEE Transactions on Signal Processing*, vol. 56, no. 5, pp. 1830 –1839, may 2008.

[3] D.W. Griffin and J.S. Lim, "Signal reconstruction from short-time fourier transform magnitude," *IEEE Trans. Acoust., Speech, and Signal Proc.*, vol. 32(2), pp. 236–243, 1984.

[4] M. Portnoff, "Implementation of the digital phase vocoder using the fast fourier transform," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, no. 3, pp. 243 – 248, June 1976.

[5] Leigh D. Alsteris and Kuldip K. Paliwal, "Iterative reconstruction of speech from short-time fourier transform phase and magnitude spectra," *Computer Speech & Language*, vol. 21, no. 1, pp. 174 – 186, 2007.

[6] K.K. Paliwal and L.D. Alsteris, "On the usefulness of stft phase spectrum in human listening tests," *Speech Communication*, vol. 45 (2), pp. 153–170, 2005.

[7] B. Yegnanarayana, D. Saikia, and T. Krishnan, "Significance of group delay functions in signal reconstruction from spectral magnitude or phase," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 3, pp. 610 – 623, June 1984.

[8] Jonathan Le Roux, Hirokazu Kameoka, Nobutaka Ono, and Shigeki Sagayama, "Fast signal reconstruction from magnitude stft spectrogram based on spectrogram consistency," *proceedings of DAFx'10*, 2010.

[9] D. Gunawan and D. Sen, "Iterative phase estimation for the synthesis of separated sources from single-channel mixtures," *IEEE Signal Processing Letters*, vol. 17, no. 5, pp. 421 –424, may 2010.

[10] Jonathan Le Roux, Emmanuel Vincent, Yuu Mizuno, Hirokazu Kameoka, Nobutaka Ono, and Shigeki Sagayama, "Consistent Wiener filtering: Generalized time-frequency masking respecting spectrogram consistency," in *Proc. 9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA 2010)*, Sept. 2010, pp. 89–96.

[11] Bin Yang, "A study of inverse short-time fourier transform," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2008*, 4 2008, pp. 3541 –3544.

[12] S. Nawab, T. Quatieri, and Jae Lim, "Signal reconstruction from short-time fourier transform magnitude," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 31, no. 4, pp. 986 – 998, Aug. 1983.

[13] Jake Bouvrie and Tony Ezzat, "An incremental algorithm for signal reconstruction from short-time fourier transform magnitude," *proceedings of ICSLP'06*, 2006.

[14] Radu Balan, "On signal reconstruction from its spectrogram," *proceedings of the Conference on Information and Sciences Systems*, 2010.

[15] M. Hayes, Jae Lim, and A. Oppenheim, "Signal reconstruction from phase or magnitude," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, pp. 672–680, 1980.

[16] W. Kim and M.H. Hayes, "Phase retrieval using a window function," *IEEE Transactions on Signal Processing*, vol. 41, no. 3, pp. 1409 –1412, Mar. 1993.

[17] J. R. Fienup and C. C. Wackerman, "Phase-retrieval stagnation problems and solutions," *J. Opt. Soc. Am. A*, vol. 3, pp. 1897–1907, 1986.

[18] Jonathan Le Roux, Nobutaka Ono, and Shigeki Sagayama, "Explicit consistency constraints for stft spectrograms and their application to phase reconstruction," *proceedings of SAPA'08*, 2008.

[19] M. Hayes and T. Quatieri, "The importance of boundary conditions in the phase retrieval problem," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '82.*, May 1982, vol. 7, pp. 1545 – 1548.

[20] R. W. Gerchberg and W. O. Saxton, "A practical algorithm for the determination of the phase from image and diffraction plane pictures," *Optik*, vol. 35, pp. 237–246, 1972.

[21] F. J. Harris, "On the use of windows for harmonic analysis with the discrete fourier transform," *proc. of the IEEE*, pp. 51–83, 1978.

[22] Jonathan Le Roux, Hirokazu Kameoka, Nobutaka Ono, and Shigeki Sagayama, "Phase initialization schemes for faster spectrogram-consistency-based signal reconstruction," in *Proceedings of the Acoustical Society of Japan Autumn Meeting*, Mar. 2010, number 3-10-3.

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

[24] G. T. Beauregard, X. Zhu, and L. Wyse, "An efficient algorithm for real-time spectrogram inversion," in *Proceedings of the 8th International Conference on Digital Audio Effects (DAFx-05)*, Sept. 2005, pp. 116–118.

[25] Volker Gnann and Martin Spiertz, "Inversion of short-time fourier transform magnitude spectrograms with adaptive window lengths," *proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 325–328, 2009.

[26] X. Zhu, G. Beauregard, and L. Wyse, "Real-time iterative spectrum inversion with look-ahead," *proceedings of IEEE International Conference on Multimedia and Expo*, pp. 229–232, 2006.

[27] Volker Gnann and Martin Spiertz, "Improving rtisi phase estimation with energy order and phase unwrappin," *proceedings of DAFx'10, Gratz, Austria*, 2010.

[28] P. Van Hove, M. Hayes, Jae Lim, and A. Oppenheim, "Signal reconstruction from signed fourier transform magnitude," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 31, no. 5, pp. 1286 – 1293, Oct. 1983.

[29] Kannan Achan, Sam T. Roweis, and Brendan J. Frey, "Probabilistic inference of speech signals from phaseless spectrograms," in *In Neural Information Processing Systems 16*. 2003, pp. 1393–1400, MIT Press.

[30] Bertrand Nouvel, "A study of a local-features-aware model for the problem of phase reconstruction from the magnitude spectrogram," *proceedings of ICASSP'10*, pp. 4026–4029, 2010.

[31] Tony Ezzat, Jake Bouvrie, and Tomaso Poggio, "Max-gabor analysis and synthesis of spectrograms," *proceedings of ICSLP'06*, 2006.

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