3-10-3

Phase initialization schemes for faster spectrogram-consistency-based signal reconstruction *

◎ Jonathan Le Roux, Hirokazu Kameoka (NTT CS Labs),
Nobutaka Ono and Shigeki Sagayama (The University of Tokyo)

1 Introduction

In previous contributions [1, 2], we presented a fast algorithm for the reconstruction of a time-domain signal from a magnitude short-time Fourier transform (STFT) spectrogram, and showed experimentally that it could lead to substantial reduction in computation time compared to previous work by Griffin and Lim [3] when both algorithms are initialized using zero phase. We study here other strategies for phase initialization, compare the performance of our method with the state-of-the-art RTISI-LA algorithm [4], and consider several extensions and combinations of the two approaches.

2 Signal reconstruction from magnitude spectrograms

Let $(A_{m,n})_{0\leq m\leq M-1, 0\leq n\leq N-1} \in \mathbb{C}^{MN}$ denote an array of real non-negative numbers, where m will correspond to the frame index and n to the frequency band index. This array, for example obtained through modifications of the magnitude spectrogram of a sound, is assumed to be the magnitude part of a hypothetical STFT spectrogram. The problem of signal reconstruction is to find the “most relevant” phase for A.

Griffin and Lim [3] formulated the problem as that of estimating a real-valued time-domain signal x such that the magnitude of its STFT is closest to the magnitude spectrogram A in a least-squares sense. To solve it, they proposed the iterative STFT algorithm, which consists in iteratively updating the phase $\phi^{(k)}$ at step k by replacing it with the phase of the STFT of its inverse STFT while keeping A fixed. Alternatively, we showed in [1] that this problem can also be formulated as that of estimating a phase ϕ such that $H = Ae^{j\phi}$ is “as consistent as possible”, where we call “consistent spectrograms” the elements of $\mathbb{C}^{MN}$ which can be obtained as the STFT spectrogram of a time-domain signal. The lack of consistency, or inconsistency, of any array H is numerically characterized by the L^2 norm of $\mathcal{I}(H) = \|\mathcal{G}(H) - H\|^2$, where $\mathcal{G}(H) = \text{STFT}(\text{iSTFT}(W))$ denotes the STFT of the inverse STFT of H. The two formulations can be shown to be equivalent near convergence, and we can actually interpret the iterative STFT as the application of the auxiliary function method to the minimization of $\tilde{\mathcal{I}}$. Our consistency-based formulation has the advantage to be defined directly in the time-frequency (T-F) domain, to allow for fast computation of $\mathcal{G}(H)$ thanks to local approximations, and to justify the use of partial updates limited to a subset of T-F bins. Based on it, we developed a signal reconstruction algorithm which can achieve the same reduction of inconsistency 10 to 60 times faster than the iterative STFT algorithm. We shall refer to [2] for more details, and only give here the update equations: at iteration k, we update the phase $\phi_{m,n}$ of all bins whose magnitude is larger than c_k, where c_k is exponentially decaying with k, through

$$\phi_{m,n} \leftarrow \angle\Big(\sum_{\substack{(p,q)\neq(0,0)\\ |p|\leq l}} e^{j2\pi\frac{qR}{N}n}\alpha_{q,p}H_{m-q,n-p}\Big), \quad (1)$$

where the $\alpha_{q,p}$ are weights depending on the analysis window of the STFT, N the window length, R the window shift, l a truncation order typically set to 5, and H is the current spectrogram estimate.

3 RTISI-LA

In our previous papers, we performed all comparisons by simply initializing the algorithms with zero phase. Our intent was to compare the performances of Griffin and Lim’s method (“G&L”) and our method under the same conditions, leaving the question of using better initial phase estimates to future works. In order to investigate further the applicability of our approach, we consider here its comparison with a state-of-the-art method relying on better initial phase estimates, Zhu et al.’s real-time iterative spectrogram inversion with look-ahead (RTISI-LA) [4], as well as several extensions and combinations of the two approaches.

RTISI-LA is based on a sliding-block approach: some processing is performed on the block constituted of frames m to $m+k$; when it is over, frame m exits the block and is no longer updated, while a new frame with index $m+k+1$ is inserted in the block. The processing performed by RTISI-LA consists in: 1) computing the spectrum of the short-term signal obtained by overlap-adding the contributions at frame $m+k$ of the previous frames m to $m+k-1$ overlapping with frame $m+k$, using the best available phase estimates for these frames. The initial phase of frame $m+k$ is determined as the phase part of this spectrum; 2) performing p iterative STFT updates on the current block, taking into account all past frames but discarding the contribution of future frames with index larger than or equal to $m+k+1$. The rationale for discarding future frames is that the absence of a reliable phase estimate for them is likely to make their contribution more of a disturbance than a useful clue. In order to partially compensate for the above asymmetry between previous and future frames, Zhu et al. propose to use asymmetric analysis windows with a reverse effect when analyzing frame $m+k$. Although the above procedure relies on heuristic considerations, the authors show that it leads to much better performance than G&L for a given number of iterations per block p (note that, as RTISI-LA is an online algorithm discarding future frames, the result of the algorithm with a given number of iterations $p+1$ cannot be deduced from the result for p).

4 Further developments

Other than just comparing our previously proposed method with RTISI-LA, we shall also consider new developments stemming from insights gained from both approaches: combined usage of initial phase estimation through RTISI-LA, followed by our method (“RTISI-LA Then Proposed”); introduction of a consistency-based RTISI-LA-like initial phase estimation in our method (“No future initialization”); design of a consistency-based version of RTISI-LA directly in the time-frequency domain (“TF-RTISI-LA”). The first development is self-explanatory, and we shall only explain the two others.

In RTISI-LA, the first phase estimate for a new frame is obtained by discarding the contributions of that frame and of future ones. This amounts to considering that the phase values for these frames are uniformly distributed, and that the expectation

* スペクトログラム無矛盾性に基づく位相付加アルゴリズムにおける位相初期値選択による高速化の検討、ルルー・ジョナトン、亀岡弘和 (NTT)、小野順貴、嵯峨山茂樹 (東大情報理工)

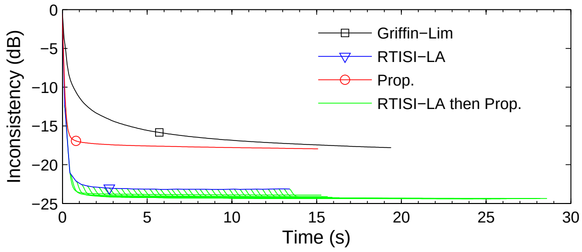

Fig. 1 *Evolution of* $\mathcal{C}$ *(music,* $f = 0.7$*).*

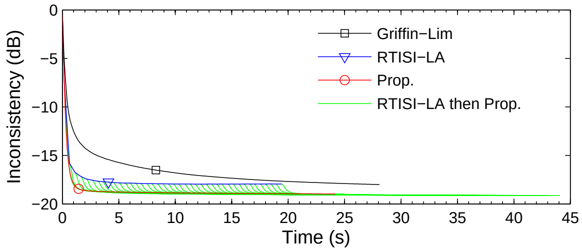

Fig. 2 *Evolution of* $\mathcal{C}$ *(speech,* $f = 0.7$*).*

of the complex value of the corresponding T-F bins is thus zero. This can be incorporated in our framework by performing a first swipe from left to right over the spectrogram, and limiting the sum in (1) to $q > 0$, thus considering only the bins which are in the frames to the left. We can also consider using the already initialized bins in the same frame, i.e., include the terms with $q = 0$ and $p > 0$. Once all bins have been initialized, we use the classical consistency-based updates.

As our approach consists in replacing computations involving round trips between the T-F domain and the time domain by local computations in the T-F domain, one can easily design a T-F domain equivalent of RTISI-LA. The computation of $\mathcal{G}$ as STFT $\circ$ iSTFT is simply replaced by its corresponding local computations similar to (1). The asymmetric windows can even be introduced by computing specific sets of weight parameters for the updates of the newest frame. These weights are computed in the same way as the ones used in (1), only replacing the analysis window by the asymmetric one. We shall however leave the evaluation of this method to future works.

5 Experimental evaluation

We investigate the performances of the various proposed methods and compare them with that of RTISI-LA and G&L in terms of the decrease of a normalized inconsistency measure defined as $\mathcal{C}(H) = 10 \log_{10} \frac{\mathcal{I}(H)}{\|H\|^2}$ (i.e., the objective function $\mathcal{I}(H)$ represented in decibels with the total energy of the original spectrogram as a reference) w.r.t. the computation time. The task we consider is that of the estimation of the most consistent phase for a given magnitude spectrogram for the purpose of time-scale modification. All methods were implemented in C. As the data, we used speech from male and female speakers divided in 10 files of average length 33 s, and 8 excerpts of music pieces [5], each 23 s long. The sampling rate was 16 kHz. The sine window was used as analysis and synthesis window. The time-scaled spectrograms are built through a sliding-block analysis technique described in [1]. The time-scale modification factors f used were 0.7 (slow down by 30 %), 1 (no time-scale modification), and 1.3 (pace up by 30 %). We used a frame length of 1024 samples and two settings for the frame overlap, namely 50 % and 75 %, although we shall show here only the results for 75% overlap. G&L and the previously proposed method use the magnitude of the time-scaled spectrograms as initial point, and are stopped after 200 iterations. The sparseness threshold at iteration k is set to $ae^{-bk^c}E$, where E denotes the mean magnitude of the given magnitude spectrogram. The parameters were set to $a = 100$, $b = 0.1$ and $c = 1$ based on a preliminary experiment.

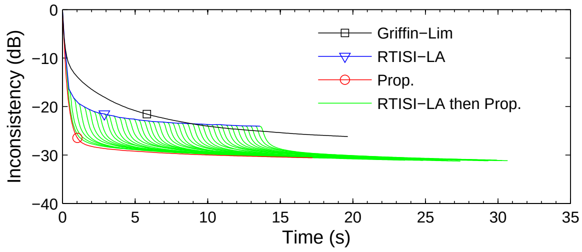

Fig. 3 *Evolution of* $\mathcal{C}$ *(speech,* $f = 1$*).*

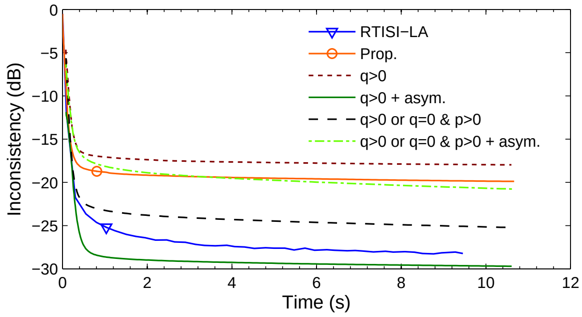

Fig. 4 *"No future" initialization (music,* $f = 0.7$*).*

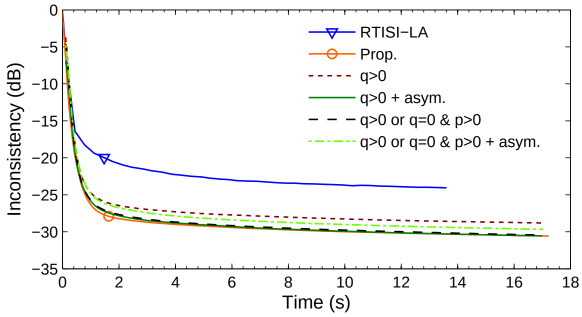

Fig. 5 *"No future" initialization (speech,* $f = 0.7$*).*

We can see in Fig. 1 that our previously proposed method is outperformed by RTISI-LA on the music data for 75 % overlap. This is however not the case on the speech data, as shown in Figs. 2 and 3, as well as for 50 % overlap (omitted). In all cases, using RTISI-LA as an initialization step for our method leads to the most reliable results. We think that the difference of behavior on speech and music data is due to the greater stability of sounds in music data, as starting from incoherent neighboring bins is then more likely to lead to local minima.

Figures 4 and 5 show the evolution of $\mathcal{C}$ for various kinds of "No future" initialization steps on music and speech data respectively, for $f = 0.7$. The best results are obtained for an initialization involving only the bins for $q > 0$ and using weights α computed using an asymmetric window. Similar results were obtained for 50 % overlap and other values of f.

6 Conclusion

We compared the consistency-based method for signal reconstruction that we previously proposed to the RTISI-LA method of Zhu et al., and considered several combinations of those methods, leading to yet faster reconstruction algorithms outperforming the state-of-the-art methods.

References

[1] Le Roux et al., Proc. SAPA, 23–28, 2008.

[2] Le Roux et al., Proc. DAFx-10, 2010.

[3] Griffin and Lim, IEEE Trans. ASSP, 32(2), 236–243, 1984.

[4] Zhu et al., IEEE Trans. ASLP, 15(5), 1645–1653, 2007.

[5] Goto, Proc. ICA, 553–556, 2004.