IMPROVING RTISI PHASE ESTIMATION WITH ENERGY ORDER AND PHASE UNWRAPPING

Volker Gnann and Martin Spiertz

Institut für Nachrichtentechnik
RWTH Aachen University
Aachen, Germany
{gnann,spiertz}@ient.rwth-aachen.de

ABSTRACT

This paper presents two ways to improve the Real-Time Iterative Spectrogram Inversion (RTISI) algorithm. The standard RTISI phase estimator with look-ahead processes the buffered frames in reverse order. We show that better results are achieved by controlling this order according to frame energy. Another improvement is to initialize the last row of the phase estimator buffer by progressing the unwrapped phase difference of the previous frames. Furthermore, we extend these improvements to dual window length phase estimation and analyze the performance in SER with respect to different analysis window lengths.

1. INTRODUCTION

The goal of phase estimation is to complete a magnitude spectrogram with phase information so that a time-domain signal can be reconstructed. The magnitude spectrogram of the reconstructed signal should be as close as possible to the original magnitude spectrogram. One method for phase reconstruction is the Real-Time Iterative Magnitude Spectrogram Inversion with look-ahead (RTISI-LA) [1] which is real-time capable and delivers a high reconstruction quality. We will refer to this algorithm as RTISI in the following. This algorithm has also been extended to invert magnitude spectrograms with dual time/frequency resolution [2]. In the following, we will call this RTISI extension Dual-resolution RTISI. However, the iterative structure of RTISI and its variants leaves room for additional improvements.

One drawback of RTISI is the strict order the spectrogram frames are processed in. In onset situations, this leads to the paradox effect that the phase estimation for frames with high energy is determined by previous frames of low energy, whereas the other way round makes more sense. We show that controlling the order by energy improves the result of RTISI, but not of dual-resolution RTISI. Since RTISI is an iterative algorithm, its results depend heavily from the way the buffer frames are initialized. We show that an initialization with the phase progression of the instantaneous frequency of phase vocoder theory improves the results of RTISI and dual-resolution RTISI when it is applied on the long-window length frames only.

This paper is organized as follows. Section 2 gives a short overview over the general function of RTISI and dual-resolution RTISI. Section 3 explains the energy-based row update execution order. Section 4 shows how RTISI can be improved by the phase-unwrapping initialization. Section 5 evaluates these methods on the EBU-SQAM test set. This paper closes with the conclusions.

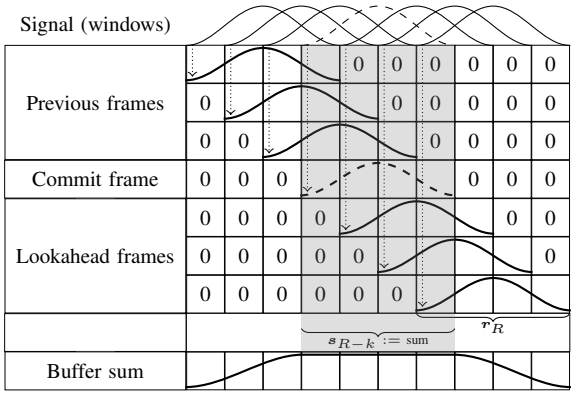

Figure 1: *Phase estimation buffer. Every sketched cell contains S elements, whereas S denotes the hop length between adjacent frames.*

2. STANDARD AND DUAL-RESOLUTION RTISI

The basic data structure of RTISI is a two-dimensional buffer which is illustrated in Figure 1. S determines the hop length between adjacent frames, L the window length, which is in our setup $4S$. The rows are described as vectors $\boldsymbol{r}_i$ where i denotes the row index. The corresponding sum is $\boldsymbol{s}_i$. The phase estimation starts with a buffer filled with zeros.

2.1. Standard RTISI Algorithm

Let us assume that the audio data for all rows except the last one have already been estimated. We can estimate the content for the last row (or improve its estimation) by the following procedure:

1. Initialize the last buffer row with zeros or with an application-given initial estimation. Alternatively, we can use the phase-unwrapping initialization of Section 4.
2. Calculate the sum of all buffer rows and limit it to the part covered by the last row.
3. This sum is implicitly windowed with a sum of overlapping analysis *and* synthesis windows $w[n]$, leading to inconsistencies between time and frequency representation (window sum error). Multiply this sum with $\frac{w[n]}{\sum_{m=1}^{L/S} w^2[n-mS]}$ to compensate the error, so that the sum is implicitly windowed with $w[n]$.
4. Calculate the phase spectrum $\angle S[k]$ of the sum $s[n]$ using an DFT.

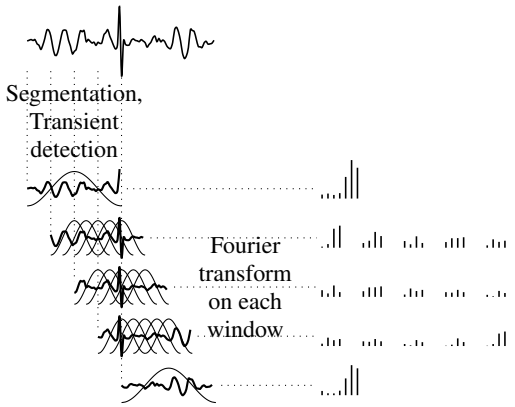

Figure 2: *Generation of spectrograms using window switching. The first step (segmentation) is equivalent to STFT windowing with a rectangle window. A transient detector decides if a segment is processed with one single long or with multiple overlapping short Hamming windows. After the actual windowing, each windowed segment is transformed into the frequency domain.*

5. Combine the phase spectrum $\angle S[k]$ with the corresponding magnitude of the magnitude spectrum $|X[k]|$ stored in the buffer using the formula

$$Y[k] = \frac{S[k] \cdot |X[k]|}{|S[k]|} \forall k \tag{1}$$

6. Transform $Y[k]$ into the time domain (using an inverse DFT), window the result and store it into the last row $\boldsymbol{r}_R$:

$$\boldsymbol{r}_R[n] = w[n] \cdot \sum_k Y[k] \mathrm{e}^{j2\pi kn/L} \tag{2}$$

7. Perform steps 2–7 on the second-last row $\boldsymbol{r}_{R-1}$, on the other lookahead-frame rows (in reverse order), and on the commit-frame row, respectively.
8. Repeat steps 2–8 a certain number I of additional iterations.
9. Commit the frame stored in the commit-frame row and synchronize the buffer to the next frame.

An overlap-add synthesis step assembles the final audio data from the committed frames.

2.2. Dual-resolution RTISI

An RTISI extension presented in [2] allows to estimate the phase for dual-resolution spectrograms. The generation and structure of these spectrograms is presented in Figure 2. As transient detector, we use the absolute discrete group delay method explained in [3].

The actual phase estimator uses two buffers, one for each window length, as presented in Figure 3. The phase estimation for a long-window frame works on the long-window-length buffer as described in Section 2.1. Every time when one short-window spectrum sequence is processed, the short-window-length buffer is synchronized to the long-window-length buffer. Then, the estimation (also see Sec. 2.1) is performed on the short-window buffer. The phase estimation result is transferred to the long-window buffer such that the overlap-add property is preserved.

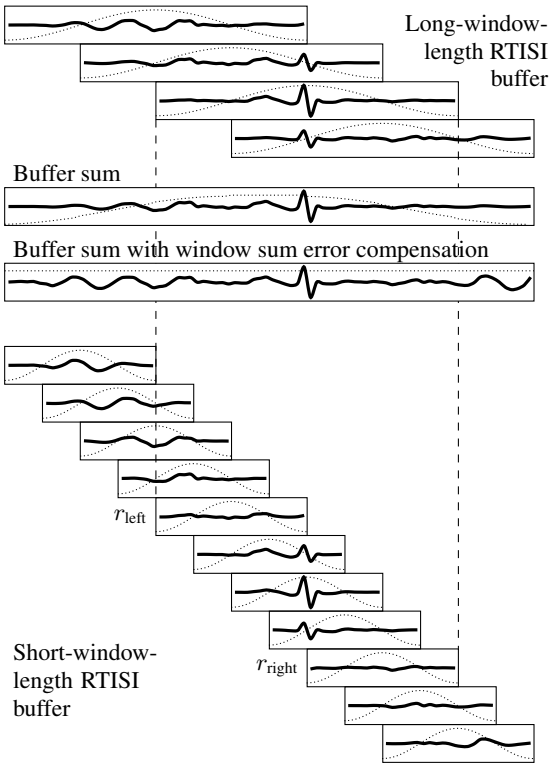

Figure 3: *Two RTISI buffers with different window lengths. To transport audio data between the buffers, the algorithm calculates the buffer sum, compensates the window sum error, and windows the result for each target buffer row. The dotted lines denote the window function the audio data in the buffer are implicitly multiplied with.*

3. ENERGY ORDER

RTISI uses a reverse linear order to process the frames in the buffer. First, the phase of the last buffer row is estimated, then the second last and so on, until the commit-frame row is reached. This order was developed from RTISI without look-ahead and gives good results. However, the phase estimation problem has no single optimal solution, there are many local optima. RTISI is an iterative algorithm, so the reconstruction quality depends on the row order and the initialization.

The linear order has especially the drawback that the phase estimation of loud segments depends on the estimation results of previous quiet segments. Instead, it is better to adjust the quiet frames to the loud ones, because with lower amplitude, an estimation error has a lower influence to the overall result. This holds especially for onset situations. This consideration leads to the idea that the process order should be controlled by the frame energy. Loud frames should be estimated first:

$$\text{order} = \text{argsort}_i \left(-\sum_n \left(\boldsymbol{r}_i[n] \right)^2 \right), \tag{3}$$

where the argsort function returns the sequence of indices i yielding the ascending sorted order of the argument.

On dual-resolution RTISI, this principle can not be applied on the short-window-length buffer for following reason: RTISI ex-

ploits the overlap of neighboring frames to estimate the phase for the recent one. When the hop length S is $\frac{1}{4}$ of the block length L (as usual for RTISI), the blocks r_{left} and r_{right} of the short-window length buffer in Figure 3 do not overlap. For larger window length ratios, the number of non-overlapping blocks even increases. As a result, a linear order from left to right is required in the first iteration to provide the initial estimation with overlapping. Further iterations can be processed in energy order.

4. PHASE UNWRAPPING INITIALIZATION

Since RTISI is an iterative algorithm, its result depends on the initialization of the last buffer row. In some applications, a buffer initialization can be retrieved from the application itself. When only the magnitude spectrogram is available, one possibility is to initialize the last buffer row with zeros. In the steady-state case, we can also prolong the phase difference of the preceding frames to the current frame, like in the phase vocoder [4]. The last frame $\boldsymbol{r}_R$ is initialized with

$$\boldsymbol{r}_R = A \cdot \mathrm{IDFT}\left\{|S_R[k]| \cdot \mathrm{e}^{j\left(\angle S_{R-1}[k] + \left(\angle S_{R-1}[k] - \angle S_{R-2}[k]\right)\right)}\right\} \tag{4}$$

$$= A \cdot \mathrm{IDFT}\left\{|S_R[k]| \cdot \mathrm{e}^{j\left(2\angle S_{R-1}[k] - \angle S_{R-2}[k]\right)}\right\} \tag{5}$$

$$= A \cdot \mathrm{IDFT}\left\{\frac{|S_R[k]| \cdot S_{R-1}^2[k] \cdot |S_{R-2}[k]|}{|S_{R-1}[k]|^2 \cdot S_{R-2}[k]}\right\}. \tag{6}$$

The gain factor A determines the volume of the initialization; $A = 0$ turns this initialitation off. This initialization improves the phase progression between neighboring frames and thus should theoretically deliver a better result in steady-state mode.

5. EXPERIMENTS AND RESULTS

In order to evaluate these improvements, we use a test set based on the Sound Quality Assessment Material (SQAM) from the EBU [5]. Our test set consists of 70 files containing speech, singing vocals, and instruments. The sampling frequency is 48 kHz. For spectrogram generation and phase estimation, we use an Hamming window with $L = 4S$, yielding an overlap of 75%. As evaluation measure, we use the mean signal-to-error ratio (in dB) of the magnitude spectrograms of the phase-reestimated signal versus the original, respectively:

$$\mathrm{SER} = 10 \log \frac{\sum_{m=-\infty}^{\infty} \sum_{k=0}^{L-1} |X[mS,k]|^2}{\sum_{m=-\infty}^{\infty} \sum_{k=0}^{L-1} (|X[mS,k]| - |X'[mS,k]|)^2} \tag{7}$$

This SER measure operates on STFT magnitudes and thus depends on its own STFT window length. This SER window length determines the operating point of the time-frequency resolution tradeoff. To analyze the phase estimation performance over the whole range of time/frequency resolutions, the SER values are plotted against this window length. A good phase reestimation should achieve high SER values for all window lengths. Low SER values for low analysis window lengths are a sign of a bad temporal resolution, leading to a smearing of transients. Low SER values for long analysis window lengths demonstrate a bad frequency resolution, resulting in a bad accuracy of low pitch.

5.1. Energy order

Figure 4 shows the average signal-to-error ratio on the EBU-SQAM test set for an RTISI with a block length of 1024 and 2048 samples. We can make following observations:

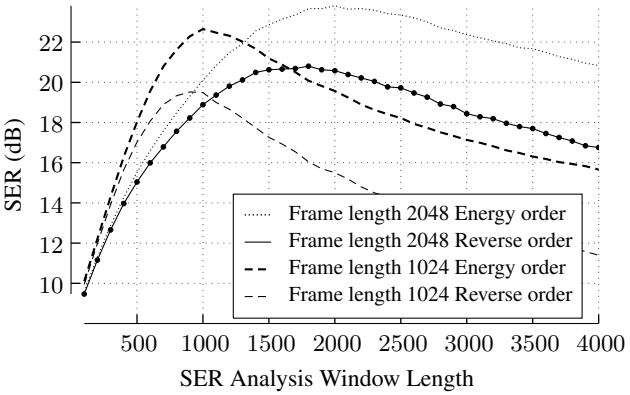

Figure 4: *SERs of fixed-resolution RTISI over the complete EBU test set. The bullet marks on the solid lines denote the measuring points of* all *curves.*

- The maximum SER peak is reached when the SER window length corresponds with the spectrogram window length the phases are estimated for. This is not surprising because RTISI minimizes the local mean-square error between the original and the reconstructed spectrogram, which is equivalent to a maximization of the SER at this window length. To reduce the direct influence of this optimization onto the result, in Figure 4 the SER is measured with a window length of 1000, 1100, ..., 2000 etc. , not at 1024 and 2048.
- The energy order leads to an SER gain up to ≈ 4 dB for single-resolution RTISI at the SER window length of 4000. At the optimal window length, this difference is ≈ 3 dB.
- In all cases, the decline of the SER for larger window lengths than optimum is lower than the rise below this length. One possible reason is that the magnitude spectrogram reflects slow changes very well, whereas transient behavior is mainly expressed by the phase spectrum. Since the phase estimator does not know the phase spectrum in advance, it can not reconstruct transient behavior accurately.

5.2. Phase Unwrapping Initialization

Now, we analyze the influence of phase unwrapping initialization in two steps. First, we try to find heuristically the optimal amplitude A (Eq. (4)) to initialize the phase estimator. Second, we measure the influence of phase unwrapping over the whole range of time/frequency resolutions on both RTISI and dual-resolution RTISI.

Figure 5 shows the SER of RTISI with block lengths of 1024 and 2048 samples and the SER of dual-resolution RTISI ($512/2048$ samples) as a function of the amplitude A of Eq. (4). The SER measure window length is 1024 on RTISI-1024, and 2048 on the others, since RTISI optimizes the SER exactly for these window lengths. As the main result, we can conclude that values of A around 0.3 are heuristically optimal.

Figure 6 shows the average signal-to-error ratio on the EBU-SQAM test set for an RTISI with a block length of 1024 and 2048

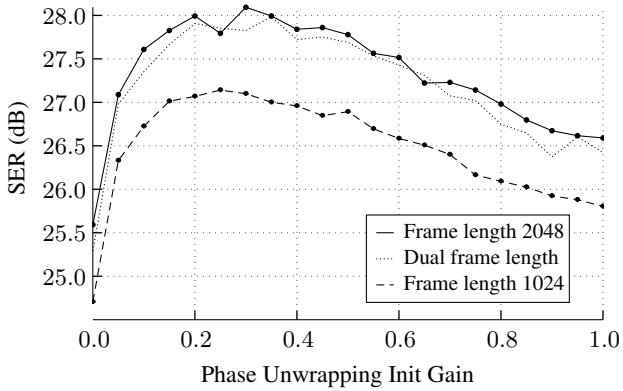

Figure 5: *SER vs phase unwrapping initialization gain. The SER analysis window length is 2048, with the exception of the RTISI-1024 curve, where the SER window length is also 1024.*

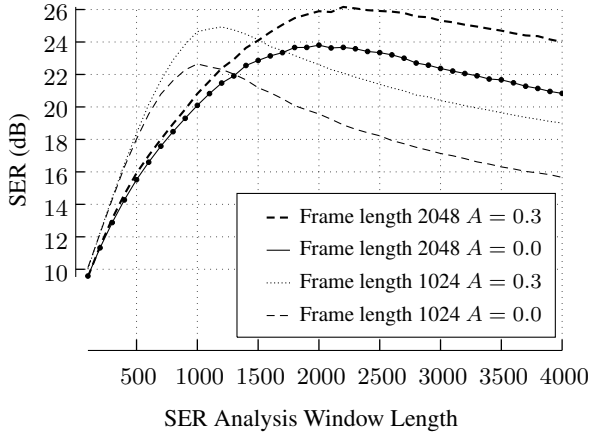

Figure 6: *SER of fixed-resolution RTISI with energy order, depending on phase unwrapping. A=0.0 (no phase unwrapping) and 0.3.*

samples. It corresponds to Figure 4, except that the reverse order experiments are replaced with the energy order, phase unwrapping experiments with $A = 0.3$. We can observe that phase unwrapping gains additional 3 dB with a reference SER window length of 4000, and 2 dB for the optimal window length.

5.3. Phase Unwrapping on Dual-Resolution RTISI

Figure 7 shows the influence of the improvements on dual-resolution RTISI. The block lengths are 2048 and 512 samples, respectively. As described in Section 2.2, the energy order is implemented as follows: For general (long frame) processing, the energy order works exactly like in single-resolution RTISI. When short frames are sub-processed in the short-window-length buffer, the first iteration is linear (from $\boldsymbol{r}_{\text{left}}$ to $\boldsymbol{r}_{\text{right}}$); in the following iterations the frames are processed in energy order. The influence of the energy order is comparable to single-resolution RTISI — about 3 dB. The phase unwrapping initialization leads to an additional SER gain of 3 dB.

The thick dashed line in Figure 7 shows the performance of

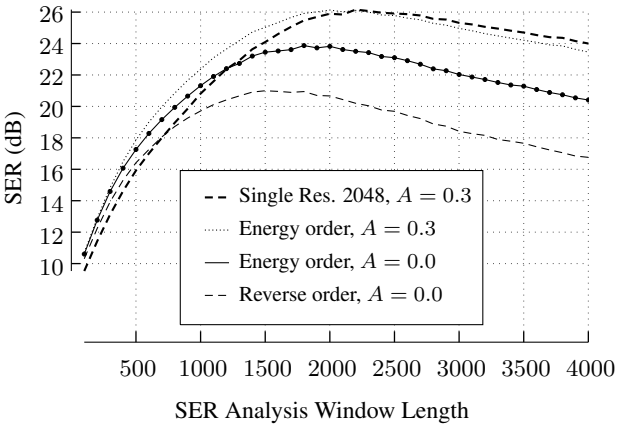

Figure 7: *Dual-resolution phase estimator results. The thick dashed line is the same as in Figure 6*

the single-resolution RTISI with a block length of 2048 samples, which equals the long-window length of the dual-resolution RTISI, both with energy order and $A = 0.3$. We can see that up to 2048 samples, the dual-resolution is superior, probably because of the better time resolution on transients. Above this point, single-resolution RTISI has a slight advantage.

5.4. Number of iterations

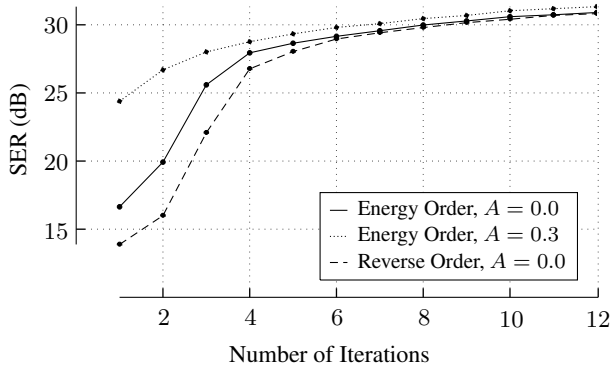

Figure 8: *SER values vs the number of iterations. Single-resolution RTISI, block size 2048 samples.*

On all previous experiments, the number of iterations was set to 3. This low number makes the differences clearly visible, but for a higher estimation quality, a higher number of iterations is preferred. Figure 8 shows how the influence of the initialization and processing order declines with an increasing number of iterations. The influence of the processing order becomes invisible, whereas the phase unwrapping initialization leads to improvements even at $I = 12$. This behavior can be explained: RTISI is based on the Griffin/Lim algorithm [6], which performs a local optimization. A different initialization leads to a better local optimum. In contrast, a different processing order scheme merely influences the speed of convergence, so with a high number of iterations, the same optimum is reached.

6. CONCLUSIONS

We have shown two possible improvements to RTISI phase estimation. We achieve better results by controlling the buffer row processing order according to the frame energies. Another improvement is to initialize the last phase estimator frame by progressing the unwrapped phase difference of the previous frames. Furthermore, we extended these improvements to dual window length phase estimation. The combination of both improvements leads to a mean SER gain of up to 6 dB for dual-resolution RTISI. Yet with an increasing number of iterations, this difference becomes smaller. Furthermore, we show that, generally, the SER error is smaller when the measure window length exceeds the processing window length than vice versa.

7. REFERENCES

[1] X. 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.

[2] V. Gnann and M. Spiertz, "Inversion of Magnitude Spectrograms with Adaptive Window Lengths," in *Proc. IEEE Int. Conference on Acoustic Speech and Signal Processing ICASSP '09*, 2009, pp. 325–328.

[3] V. Gnann and M. Spiertz, "Transient Detection with Absolute Discrete Group Delay," in *Proc. IEEE Int. Workshop on Intelligent Signal Processing and Communication Systems ISPACS '09*, 2009, pp. 311–314.

[4] U. Zölzer, *DAFX — Digital Audio Effects*, John Wiley & Sons, New York, USA, 2002.

[5] European Broadcasting Union, "Sound Quality Assessment Material," Tech 3253, 1988.

[6] D. Griffin and J. Lim, "Signal Estimation From Modified Short-Time Fourier Transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.