Single Pass Spectrogram Inversion

Gerald T. Beauregard
Cognosonic Pte Ltd
Singapore
g.beauregard@ieee.org

Mithila Harish
NUS Graduate School for Integrative Sciences & Engineering
National University of Singapore
Singapore
mithila.harish@u.nus.edu

Lonce Wyse
Communications and New Media
National University of Singapore
Singapore
lonce.wyse@nus.edu.sg

Abstract— **We present a computationally efficient real-time algorithm for constructing time-domain audio signals from spectrograms. The Single-Pass Spectrogram Inversion (SPSI) algorithm is similar to the synthesis step in phase-locked vocoders, but with phase rates at spectral peaks determined solely from the magnitude spectra using quadratic interpolation. The algorithm provides good quality results in a single fully-deterministic pass, and can also provide excellent initial phase estimates for improved results with iterative spectrogram inversion techniques such as the Griffin-Lim algorithm.**

Keywords—Magnitude-only reconstruction; signal estimation; spectrogram inversion; phase vocoder; phase reconstruction

I. INTRODUCTION

Magnitude spectra are widely used to represent, visualize, and perform operations on signals in the frequency domain. The spectrogram is a series of Short-Time Fourier Transform magnitudes. The equation of an STFT is given below:

$$X(mS,\varpi) = \sum_{n=-\infty}^{\infty} x(n)w(n-mS)e^{-j\varpi n} \quad (1)$$

and its magnitude, the STFTM can be represented by $|X(mS,\varpi)|$. Here w is the analysis window, S is the analysis hop size and m is the index of the frames of the STFT.

A spectrogram can be interpreted as a series of magnitude spectra representing a signal at sequential points in time, or alternatively as a multi-channel envelope representation of a signal [1]. Magnitude spectra (and spectrograms) do not have an explicit representation of the phase of the constituent frequency components. Spectrogram Inversion is the process of recovering a time-domain signal from its spectrogram representation and requires an estimate of phases for frequency components.

In many applications, the analysis and modification of the Short-Time Fourier Transform (STFT) and the Short-Time Fourier Transform Magnitude (STFTM) of speech and audio signals are necessary. These applications include, but are not limited to audio enhancement, reverberation analysis, time and pitch modification, and noise cancellation. Phases are either lost, become meaningless as the spectral representations are manipulated, or simply do not exist for artificially constructed spectrograms. The objective, then, is to use these spectral representations to generate a real-valued signal that corresponds as closely as possible to the original spectrograms.

Griffin and Lim [2] developed an iterative algorithm to estimate a signal based on its STFTM. We will refer to this algorithm henceforth as 'G&L' in this paper. A drawback of G&L is that the phase estimate for a given frame is dependent on all future and all past frames in the original signal, as noted in [3]. Thus, this method is inherently non-real time. Further, for G&L, it is important to choose appropriate initial estimates, since different initial estimates will yield different results, and there is no guarantee that the globally optimum solution will be reached [4].

The Single-Pass Spectrogram Inversion (SPSI) method we propose herein is akin to that of phase vocoders that incorporate phase-locking around peak bins as described by Laroche [5] and Puckette [6] but in which the phase rate at the peak bins is estimated by interpolation of the magnitude spectrum as described by Smith [7]. Since the phase rate of the peaks is estimated directly from the magnitude spectrum, and the phases of non-peak bins are locked to those of the peaks, the method can efficiently transform a magnitude-only spectrogram into a time-domain signal in a single deterministic pass.

The SPSI method provides a good estimate for phases in terms of an error measure used for comparison with a known time-domain signal. The method can also provide excellent initial phases that improve the performance of iterative spectrogram inversion techniques such as G&L [2] and others derived from it, such as RTISI and RTISI-LA [3][8].

In this paper we present the SPSI algorithm and evaluate its performance, and demonstrate its efficacy as an initial phase estimate for G&L reconstruction.

II. MATHEMATICAL DESCRIPTION

The SPSI method begins with a series of magnitude spectra assumed to represent overlapping windowed frames of an actual or hypothetical time-domain signal. The frames are assumed to be windowed using a Hanning window, with an analysis step size S which is one quarter the frame length L.

The steps of the algorithm for each frame are outlined below:

From the magnitude spectrum, we identify the bins that represent peaks in the spectrum by comparing the magnitude of each bin j with that of neighbors, $j+1$ and $j-1$. Thus if $|X(mS,\omega_j)| > |X(mS,\omega_{j-1})|$ and $|X(mS,\omega_j)| > |X(mS,\omega_{j+1})|$, then bin j is considered a peak. Here m is the time index and $\omega_j = \frac{2\pi j}{L}$ is the frequency of bin j and N refers to the size of the Fourier Transform. For simplicity, we shall use the following Greek letters to depict these parameters: $\alpha = |X(mS,\omega_{j-1})|$, $\beta = |X(mS,\omega_j)|$ and $\gamma = |X(mS,\omega_{j+1})|$.

Quadratic interpolation is then used as in [7] to identify the true peak position based on the magnitude of the peak bin and its neighbors using the formula:

$$p = 0.5\frac{\alpha-\gamma}{\alpha-2\beta+\gamma} \tag{3}$$

The value p has a value in the range [-0.5, 0.5]. It represents the deviation from the true peak from the peak bin as a proportion of the bin size, as shown in Figure 1. This is important since each peak corresponds to a sinusoid whose frequency does not necessarily line up precisely with a bin center frequency. This interpolation yields a better estimate for the true peak frequency. If the denominator in (3) is 0, then the true peak is exactly aligned with the bin frequency.

Next, the frequency at the true peak is calculated using the formula given below, as noted in [9], where j refers to the peak bin position, and p is as calculated as in (3):

$$\omega_j = \frac{2\pi(j+p)}{L} \tag{4}$$

where ω_j is the adjusted phase rate associated with the peak bin.

Using [5], we have a phase accumulator ϕ which represents the phase values to use for bin j at the current frame m,

$$\phi_{m,j} = \phi_{(m-1),j} + S\omega_j \tag{5}$$

where S refers to the synthesis step size (which is the same as the analysis step size). The phase accumulator is updated according to this formula only for the peak bins.

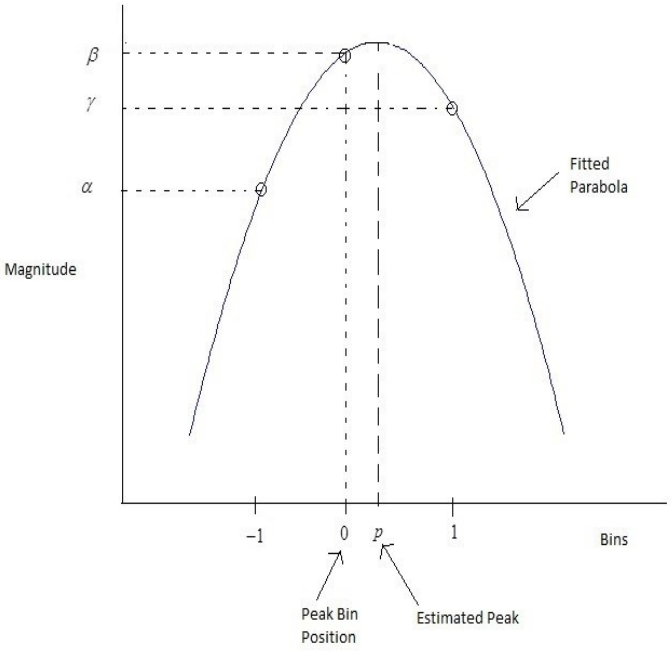

Figure 1. Quadratic Interpolation to Estimate the Value of the True Peak

Now that the phase at the peaks has been determined, the phases at the remaining bin positions can be determined depending on the sign of p. The Pi-phase alternation strategy suggested by Puckette [6] is used which has two cases: $p < 0$ and $p \geq 0$ in either case, the two neighboring bins will take the phase of the peak bin shifted by pi. For all the bins other than the peak bin and its immediate neighboring bins, the following applies:

Case (i): $p < 0$,
If $p < 0$, All the bins to the right of the peak bin will have the phase of the peak bin with a shift of pi until the next 'trough' or 'valley.' Beyond the trough, the bins are locked to the next peak. The bins to the left of the peak bin will have the same phase as that of the peak bin. Again this value is propagated until a trough is encountered.

Case (ii): $p \geq 0$,
If $p \geq 0$, the same procedure is followed as above, but in reverse. The bins to the right of the peak bin will take the phase of the peak bin, and the bins to the left of the peak bin will take the phase of the peak bin with a shift of pi. In both cases, as before, this is carried out only as far as the next trough.

Now that phases for every bin have been computed, they can be combined with the frequency component magnitudes providing the information necessary to reconstruct the time domain signal. The final step is to compute an Inverse Fast Fourier Transform which is Hanning windowed to yield the output frames which are then overlap-added to create a real valued audio signal. A flowchart summary of the SPSI process is shown in Figure 2.

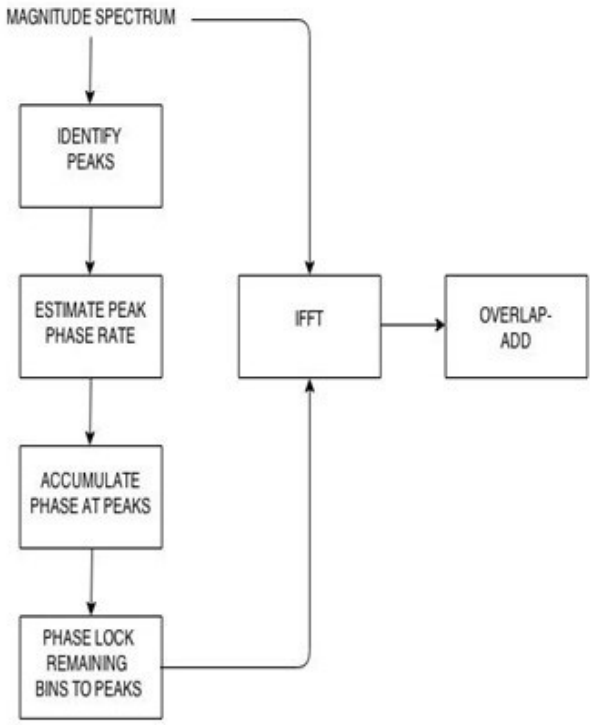

Figure 2. Flowchart of the SPSI reconstruction process

III. EVALUATION

To evaluate the effectiveness of the algorithm, we compute a spectrogram from a known time-domain audio signal. We then run the SPSI algorithm to estimate the phases for the series of magnitude spectra which we compare to the original using a measure knows as the Signal-to-Error Ratio (SER). The SER is defined as in [10]:

$$SER = 10\log \frac{\sum_{m=-\infty}^{\infty} \frac{1}{2\pi} \int_{\omega=-\pi}^{\pi} X(mS,\omega)^2 d\omega}{\sum_{m=-\infty}^{\infty} \frac{1}{2\pi} \int_{\omega=-\pi}^{\pi} [|X(mS,\omega)| - |X'(mS,\omega)|]^2 d\omega} \quad (6)$$

Here X is STFTM of the original signal, X' is the STFTM of the signal reconstructed using SPSI, m refers to the time index of the frames of the STFT and ω refers to the frequency index. A high SER indicates better fit between the two spectra being compared.

We evaluate the performance of the SPSI Algorithm on its own, as well as by using it as an initial estimate of the G&L, which for this purpose is described briefly below.

A. Description of the G&L Algorithm.

Starting with an initial estimate $X^0(n)$ of the original time-domain signal x(n), each iteration of the G&L algorithm iteratively renews the estimate:

$$x^{i+1}(n) = \frac{\sum_{m=-\infty}^{\infty} w(n-mS) \frac{1}{2\pi} \int_{\omega=-\pi}^{\pi} \hat{X}^i(mS,\omega) e^{-j\omega n} d\omega}{\sum_{m=-\infty}^{\infty} w^2(n-mS)} \quad (7)$$

In (7), $|X(mS,\omega)|$ is the STFTM of the original signal $x(n)$, $|X^i(mS,\omega)|$ is the STFTM of the ith estimate $x^i(n)$ and S refers to the step size of the analysis window.

$\hat{X}^i(mS,\omega)$ is the STFT of $x^i(n)$ with the following magnitude constraint:

$$\hat{X}^i(mS,\omega) = X^i(mS,\omega) \frac{|X(mS,\omega)|}{|X^i(mS,\omega)|} \quad (8)$$

That is, each iteration yields a new set of phases that are combined with the original magnitude spectrum for the next iteration. The distance measure is calculated as the squared error between the original and reconstructed signal spectrograms using the SER. One of the key results from Griffin and Lim [3] is that the algorithm monotonically reduces the SER with each iteration.

B. Performance of the SPSI Algorithm.

We use four audio samples for analysis: two speech signals and two musical signals. The speech signals consist of one male and one female recording and the musical signals consist of samples of two songs with percussion. Each signal has a duration of 4 seconds and is sampled at a sampling rate of 44100 Hz. We use a window length of 2048 and 75% frame overlap for all the signals. The table below illustrates the SER (in dB) of the speech and music signals processed using the SPSI Algorithm in a single pass.

TABLE I. ANALYSIS OF THE SPSI ALGORITHM

SAMPLE	SPSI SER
Male Speaker	20.82
Female Speaker	24.25
Music #1	26.67
Music #2	23.89

For comparison, we compute SER values for the G&L algorithm using a) a single iteration, and b) 10 iterations – in each case starting with an initial zero-phase estimate across all frequencies (a frequently used alternative is to use random phases).

TABLE II. ANALYSIS OF THE CONVENTIONAL G&L ALGORITHM

SAMPLE	SER after 1 ITERATION	SER after 10 ITERATIONS
Male Speaker	14.40	23.50
Female Speaker	17.04	30.02
Music #1	15.16	30.35
Music #2	15.60	27.54

As can be seen by comparing Table I and Table II, the SPSI algorithm yields an SER significantly greater (better) than single-pass G&L algorithm. It is lower, though quite close to the SER measured for the G&L algorithm after 10 iterations for each of the four audio signals.

Although the G&L algorithm guarantees that the SER monotonically improves with each iteration, it does not guarantee convergence to a globally optimal solution, and depends upon the initial conditions. For this reason, we next tested the G&L algorithm using the SPSI-derived phases as an initial condition.

TABLE III. ANALYSIS USING THE SPSI ALGORITHM AS AN INTIAL ESTIMATE FOR THE G&L ALGORITHM

SAMPLE	1 ITERATION	10 ITERATIONS
Male Speaker	24.38	25.89
Female Speaker	29.77	40.64
Music #1	35.05	46.17
Music #2	30.93	42.52

It can be seen that after only 1 G&L iteration starting from the SPSI-derived initial condition, the SER is comparable to the 10-iteration zero-phase G&L condition. After 10 iterations, the G&L algorithm with SPSI derived initial phases produces SER measures that are significantly better than 10-iteration G&L algorithm run starting from zero phase. The SPSI+G&L reconstruction is close to the original signal in the time domain as well as is qualitatively shown in Figure 3.

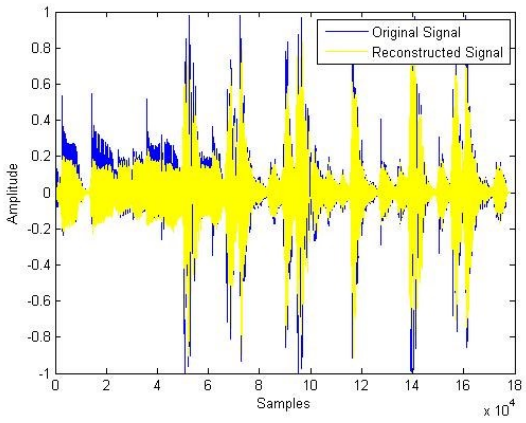

Figure 3. Plots of a Musical Signal and its Reconstructed Version using SPSI derived initial phases for G&L.

IV. CONCLUSION

The Single Pass Spectrogram Inversion (SPSI) algorithm was developed which estimates phases for spectrogram reconstruction in a single pass. The advantage of this method over other phase reconstruction methods is the computational simplicity and increased efficiency. The phase reconstruction performed by this algorithm involves the calculation of peaks in the magnitude spectra of the signal, and the subsequent phase increment at the peaks by *pi* shifts. Looking at this from the phase vocoder perspective, it possesses distinct advantages over other methods since no phase unwrapping and arctan calculations are necessary. SPSI also compares favorably with spectrogram inversion algorithms that necessitate many frequency transform iterations. Future work could include detailed quality analysis using both quantitative and subjective methods. Further analysis could also be conducted on its use as an initial estimate for the Griffin and Lim Algorithm and its variants. We also expect that additional improvement in quality could come from allowing non-peak frequency bins to be influenced by more than one neighboring peak.

V. ON-LINE

All SPSI software is open-source and available on-line at http://anclab.org/software/phaserecon. All of the sound examples used in this paper can be heard and run at the same URL.

REFERENCES

[1] R. Decorsiere, P. L. Sondergaard, E. N. MacDonald, and T. Dau, "Inversion of Auditory Spectrograms, Traditional Spectrograms, and Other Envelope Representations," *Audio Speech Lang. Process. IEEEACM Trans. On*, vol. 23, no. 1, pp. 46–56, 2015.

[2] D. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," *Acoust. Speech Signal Process. IEEE Trans. On*, vol. 32, no. 2, pp. 236–243, 1984.

[3] Xinglei Zhu, G. T. Beauregard, and L. Wyse, "Real-Time Iterative Spectrum Inversion with Look-Ahead," *Multimed. Expo 2006 IEEE Int. Conf. On*, pp. 229–232, 2006.

[4] D. L. Sun and J. O. Smith III, "Estimating a Signal from a Magnitude Spectrogram via Convex Optimization," in *Audio Engineering Society Convention 133*, 2012.

[5] J. Laroche and M. Dolson, "Phase-vocoder: about this phasiness business," *Appl. Signal Process. Audio Acoust. 1997 1997 IEEE ASSP Workshop On*, Oct. 1997.

[6] M. Puckette, "Phase-locked vocoder," *Appl. Signal Process. Audio Acoust. 1995 IEEE ASSP Workshop On*, pp. 222–225, 1995.

[7] Mototsugo Abe and Julius O.Smith III, "Design Criteria for the Quadratically Interpolated FFT Method (I): Bias due to Interpolation," *CCRMA STAN-M-114*, 2004.

[8] Xinglei Zhu, G. T. Beauregard, and L. Wyse, "Real-Time Signal Estimation From Modified Short-Time Fourier Transform Magnitude Spectra," *Audio Speech Lang. Process. IEEE Trans. On*, vol. 15, no. 5, pp. 1645–1653, 2007.

[9] Yi-Wen Liu and Julius O Smith, "Audio Watermarking through Deterministic plus Stochastic Signal Decomposition," *EURASIP J. Inf. Secur.*, vol. 2007, Jan. 2007.

[10] D. Griffin, D. Deadrick, and J. S. Lim, "Speech synthesis from short-time Fourier transform magnitude and its application to speech processing," *Acoust. Speech Signal Process. IEEE Int. Conf. ICASSP 84*, vol. 9, pp. 61–64, 1984.