A STUDY OF A LOCAL-FEATURES-AWARE MODEL FOR THE PROBLEM OF PHASE RECONSTRUCTION FROM THE MAGNITUDE SPECTROGRAM

Bertrand Nouvel

Japanese French Laboratory for Informatics - LIA CNRS
National Institute of Informatics,
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430
nouvel@nii.ac.jp

ABSTRACT

This paper introduces a new algorithm for phase reconstruction from a modified magnitude spectrogram. Conventional methods are based on specific optimization process and classical overlap-add procedure. By referring to previous works, it is observed that use of local patterns may also provide good reconstructions. This leads to a reconstruction technique based on iterative inference of local patterns. The resulting method is suited for incremental and real-time processing. It may be used directly, or as initial-guess to an optimization algorithm. The paper explores variants of models based local patterns. The results of these variants are compared.

Index Terms— Phase reconstruction, Fourier spectroscopy, Cellular automata

1. INTRODUCTION

In our everyday life, real signals keep changing over time. In order to represent their evolution though time, time-frequency analysis, i.e. the study of the spectrum of a signal evolution through the time, is done by computing successively the magnitude of the Short Time Fourier Transform (STFT). The resulting sequence is a spectrogram.

The spectrograms are the keystone element of a wide set of sound processing algorithms, including speech recognition and synthesis, or instrument separation, many sound effects such pitch-manipulation effects, time-stretching and voice correction techniques... These algorithms proceed by analysis and sometimes by interpolating in-between the different signals on the spectrogram, and then by reconstructing the signal from the modified magnitude spectrogram. Since the phase is discarded from the (magnitude) spectrogram, its reconstruction requires a specific algorithm. The problem of reconstructing the missing phase information of a modified spectrogram has thus become a classical problem of sound processing. This problem is generally complicated by the fact that artificially modified spectrograms are generally invalid spectrograms, in the meaning that there exists no signal that has the specified magnitude spectrogram.

The paradigmatic approach from Griffin and Lim [1] is to consider the phase-reconstruction problem as a generic inverse problem that can be solved through optimization process: One is looking for a signal that would give the same magnitude spectrogram as the original one. Although magnitude spectrograms do not contain the phase information explicitly encoded, it has been noticed that the overlaps in-between the different magnitudes of the STFT at different times provide information about the phase information, see [2].

This topic has been studied for thirty years, but there still remain interrogations on how to get good quality quickly, or when usual conditions are not satisfied. Griffin and Lim Algorithm [1] does converge to an optimal result, it requires at least 20 iterations to converge to acceptable solution and then is relatively slow. Moreover, current phase-reconstruction algorithms are based on information that they can infer from the overlap in time of the successive window where the STFT is computed. Hence, they perform quite poorly when the frequency at which the different window of the spectrogram are computed is low with respect to the analysis window size. So forth, nowadays, in order to achieve good signal reconstruction, we still have to handle relatively huge spectrogram data. Alternative methods may be good candidates to solve this issue.

In this paper, we explore the use of local patches of the spectrogram to reconstruct the phase by statistical means. Our approach aims at being similar to super-resolution or in-painting in image processing. Also, rather than using belief propagation, we will use cellular-automata-like process since it allows to apply the algorithm in real-time, and to have lighter computational cost. This approach shall work even when there is no overlap in-between windows. Hence, it shall not have the same weaknesses of previously mentioned methods.

Our paper is structured as follow: At first, in the Section 2, we review prior methods that have been suggested to solve the problem, then in the Section 3, we present the principle and the basics of our method. In the Section 4, we present an evaluation of our results. We finish with a brief discussion about the perspectives opened by this work.

2. STATE OF THE ART

The most simple way to regenerate sound from a spectrogram is to add a random phase to the signal. The resulting signal is generally very different from the original one. A little bit cleaner result is obtained by computing the spectrogram on different windows by interpolating in-between the different signals (overlap-add).

The Griffin and Lim algorithm [1] is based on an optimization process and on the overlap-add procedure. Griffin and Lim optimizes the phase function so that the reconstruction of the spectrogram from the signal is as close as possible from the input spectrogram. The algorithm reconstructs the signal by iteratively re-enforcing the magnitude information. It is to be noticed that the reconstruction of a signal from an invalid phase will give a different signal that will not have the same STFT. The resulting process converges to the optimal solution according to convex optimization. The RTIS algorithm (Real-Time Iterative Spectrogram Inversion Method) is a variant of Griffin and Lim that works in real-time. The algorithm operates with a sliding window, and combines its result using a classical overlap-
add scheme. RTISI has received other improvements, recently in [3], a version with look-ahead has been suggested, RTISI-LA, and is able to provide better results. Also, there exists extensions to support dynamic buffer for compatibility with AAC [4].

Among other approaches, we noted especially the work of Bouvrie and Ezzat [5] that suggests a conceptually different method: Instead of convex optimization, they use numerical methods for finding roots of non-linear equation systems. Their approach also involves an explicit smoothness hypothesis. This algorithm performs much better than previous method on reconstruction of signal from slowly sampled spectrograms.

3. OUR MODEL

3.1. Preliminary

The successive computation at frequency $S$ of the windowed Short-Time Fourier Transform is defined by:

$$X(S n, \omega) = \sum_{u=-\infty}^{\infty} x(u)w(u - S n)e^{-i \omega u}$$

The (magnitude) spectrogram for a signal in discrete time is defined by $A(S n, \omega) = |X(S n, \omega)|$. The phase spectrogram is defined similarly, $\phi(S n, \omega) = X(S n, \omega)/A(S n, \omega)$, assuming that $A(S n, \omega) \neq 0$. If $A(S n, \omega) = 0$, then, we can set $\phi(S n, \omega)$ to any arbitrary real value, since it has no effect on further discussion. Because the phase of continuous wave signal is continuously evolving through time, $\phi(S n, \omega)$ is generally easier to studied when derived through time. We denote as $\phi'$ the function $\phi(S n, \omega) - \phi(S n - 1, \omega)$. In later discussion, it is actually this function that we will attempt to recover. Theoretically, on a sine wave signal and without window function, $\phi'$ behaves linearly along $\omega$ and shall remain constant through the time. However, due to the window function $\phi'$ will be subject to spectral leakage, and thus the resulting function will not be linear: The constant value $\phi'(n - 1) - \phi'(n)$ on non-evolving signals will have in practice to be computed for each $\omega$ with respect to the window that we are using.

Then, in order to reconstruct $\phi$ from $\phi'$, one needs to know the initial phase of the spectrogram, however, according to our experiments, normal sounds and usual musics are generally not perceptually deteriorated when the initial phase is initialized randomly.

3.2. Theoretical Background

Portnoff [2] clearly establishes that: Assuming that we are using a Gaussian-window-based STFT, “given a valid magnitude (or phase) function, the pair of first order equations can be integrated to determine the corresponding phase (or magnitude) function”. Thus, [2] establishes that it is possible though some analytical equations and perfect STFT to reconstruct the phase from a spectrogram. The equation (12b) of [2] is the key equation to link phase and magnitude of a spectrogram under the assumptions taken by Portnoff in his article:

$$\frac{\partial \phi}{\partial t} - \frac{1}{\sigma^2} \frac{\partial \hat{A}}{\partial \omega} = 0$$

where, with respect to to original paper notations, $\hat{A}$ denotes the log-magnitude, and where $\sigma$ the standard deviation of the Gaussian window used in the analysis. So, it sounds great? Unfortunately, it is a little bit too nice to be as easy as this. These results stand essentially for the continuous spectrograms issued from STFT based on Gaussian window kernel, and it would be slightly different if computed for Blackman or Hamming window. Moreover they are not applicable when the window slides by more than one sample per hop. Especially, since we have only discrete observation, it is difficult to evaluate number of periods that have occurred within hop since the phase information is provided modulo $2\pi$. So forth, the attempts to use directly algebraic results on real signals will not give satisfactory results.

In spite of this last remark, Portnoff [2] opened the way to the idea that it may be possible to reconstruct the phase of the spectrogram from observation of local patches of the spectrogram: Since there exists an equation relating local features of the continuous spectrogram to its evolution, similar relation should hold on the discrete spectrograms. Thus, observing the variations of magnitude of the coefficients through frequencies and iterating local computations shall be meaningful way to reconstruct a signal.

3.3. Principle

In our algorithm, the phase will be inferred iteratively by observing and reconstructing local patterns of the time-frequency diagram. The reconstruction of the phase is done by computing a descriptor made of surrounding (local) phase and log-magnitude coefficients inside of the diagram.

The phase reconstructed at each position of the diagram depends of the descriptor computed at each position of the diagram. The exact nature of the descriptor we are using is described in Section 3.4. The function that links a descriptor with a result will be learned by different strategies (see Section 3.5). To provide more continuity in the representational space, the phase is encoded as a two-components vector.

3.3.1. Expected Strengths of the Model

Let $L$ denote the size of the interval where the windowing function $w$ is non-zero. If $L = S$ then there is no overlap in-between consecutive STFT in the spectrogram. Thus, there is no direct mathematical constraints to be exploited to reconstructed the phase. However, with normal signals, statistical approaches may still give results: The fact that an unexpected strong variation of the phase occurs exactly in-between the sampling of each window sounds is quite unlikely. So, given the initial phase of a signal, it is a reasonable assumption to say a signal should continue on the same phase in-between each frame.

3.3.2. Description of the Algorithm

In this algorithm, $S$ denotes the time-shift operator. The algorithm is based on the assumption that a spectrogram has been computed by windowed STFT (line 2) on a signal $x$, and that we have no phase information (line 3), and a process may later have modified the spectrogram (line 5), then from (line 7) the algorithm really starts, we proceed iteratively to each line through time-increasing direction, we compute a local observation (line 8), according a process described in Section 3.4, and from this observation and from a previously trained model we compute the expected result for the phase, (line 9). This last result will appear in later observations of the algorithm (line 10), and it corresponds to our prediction for the phase of the signal. Once all the phases have been computed, we re-agglomerate the magnitude and the predicted phase, and we invert the STFT with a process similar to [1]. The signal may be smoothed further by a few iterations of Griffin and the Lim Algorithm [1].
3.4. Local Observation Descriptors

Since Portnoff equations involved only partial differential equations of functions that were not shifted, we may conclude that robust computation of partial differentials is the key ingredient required to have successful results. However, since we work with discretized observations, things do not work that well, and some errors that used to be negligible in the continuous framework are no more negligible in the discrete framework. Also, in order to be more robust to noise, it is meaningful to add some prior filtering on the signal within the descriptor computation.

We define a descriptor function \( O(n, \omega) \) of the spectrogram \( X(Sn, \omega) \), as a function of that is computed according to local observation:

\[
O(n, \omega) = f((\log(A(Sn', \omega', r))))_{n', \omega \in B(Sn, \omega, r)} + (\phi(Sn', \omega', r))_{n', \omega \in B(Sn, \omega) \setminus U(Sn, \omega)}
\]

Where \( f \) is any function providing a meaningful representation of the local patch of information, and where \( B(Sn, \omega, r) \) denotes a topological ball centered in \( (Sn, \omega) \) and with an arbitrary radius \( r \), and where \( U(Sn, \omega) \) represents the set of not yet computed elements. Note that, since \( A \) is completely known, it is possible to include information about the future of signal in the descriptor.

Large patterns allow us to have more redundancy and robustness. The exact nature of this descriptor depends on the model used to do the phase observation.

3.5. Model of the Phase from Observation

3.5.1. Multinomial Models

Since the theoretical phase and the log magnitude are to be related by some PDE, one may expects that the discrete models to have a multi-linear form:

\[
\phi(t, \omega) = \sum_{w=w_\phi \cdot \sigma \omega} \sum_{u=t_\phi \cdot \tau \omega} \alpha_{\phi, u} \phi(u, w) + \sum_{w=w_\phi \cdot \sigma \omega} \sum_{u=t_\phi \cdot \tau \omega} \alpha_{A, u} \log(A)(u, w)
\]

Here, we assume that the phase may be estimated from surrounding log-magnitude, and previously-computed phase information. This model can be refined by using uniform degree multinomial model. We studied multinomial models up to the degree 3. Also, since their may be already numerous variable in the observation vector, the number of multinomials is quickly increasing.

The multinomials coefficients \( \alpha_{\phi, u} \) and \( \alpha_{A, u} \) are to be fitted by training the model over the training set.

3.5.2. K-nearest Neighbors

Multinomials models still have strong complexity limitations, \( k \)-nn based models allow to pass though this limitation.

Assuming that there exists a reasonably smooth function that maps \( O \) to a value, then the similar images should produce the similar effects. Learning, for each input, the associated output is another good way to learn. It is generally less restrictive than previous model. Although, it may be subject to over-fitting.

In order to perform well, we had to put in our training set only the vectors that contain meaningful enough information: we have removed all the observations of the phase that were related to low magnitude coefficients, since for those one the phase is very likely to be erroneous. These last meaningless vectors represent about 90% of the spectrum data this remark is crucial in order to get good performances.

In order to use \( k \)-nearest neighbors efficiently, a fast nearest neighbors algorithm, or a approximate nearest neighbors algorithm, is a must. Our implementation of the algorithm relies on SASH. SASH is randomized iterative approximative nearest-neighbor algorithm based on metric information only. One advantage of SASH, is that since the search structure is constructed from distance information only, the data structure does not suffer from the curse of dimensionality. Practically, the SASH is a hierarchical structure that approximates a nearest-neighbor search by pre-computing some of the near-neighbors of each node. All details about SASH may be found in [6]. The distance we used for SASH computation is simply a Euclidean distance in-between weighted feature vectors describing the pattern.

4. EVALUATION

4.1. Principle of the Evaluation

We evaluate our algorithm with the signal-noise ration (SNR):

\[
SNR(S_1, S_2) = 20 \log_{10} \left( \frac{|s_1|}{\sqrt{|s_1 - s_2|^2}} \right)
\]

For realizing this evaluation, we used various short sound samples. They were sampled at 44.1Khz in mono.

We evaluated the performance of our algorithm for reconstruction in different settings. We have tried different windows type of Windows: Hamming, Hann, Blackman, Gaussian and Lanczoc. We have also tried different window size: 128, 256, 512, 2048, with different overlap ratio ranging from 25 \( \rightarrow \) (very large overlap) to 0 (no overlap). We also had two different tasks: first to know whether machine learning are effectively able to learn something from the training set, we did measure cross-validation results on this set. Secondly, we used the algorithm to reconstruct some signals.

In general, good results have been obtained for hamming windows, short window size, high overlap.
4.2. Learning Independent Features

The first things one may want to ensure is that the model that are used are powerful enough to predict the data that we expect. Thus, the first thing we had to pay attention was the results on cross-validation on training-set.

This results were obtained using k-nearest-neighbors. Multinomial models generally gives less satisfactory models.

When the coefficients $\alpha_0$ are large in front of the coefficients $\alpha_A$, i.e. when the previous phase information is considered to be preeminent over the magnitude information, our method achieved an SNR of 19.55 dB (for windows of size 128 and overlap 15/16), 16.88 dB (for windows of size 128 and overlap 11/12), 14.55 dB (for windows of size 128 and overlap 9/10).

Conversely, when the coefficients $\alpha_A$ are large with respect to $\alpha_0$, we manage to achieve up 18 dB on the imaginary part and 15 dB on the real part reconstruction of the derived phase. When there is no overlap, the method on 128 frames only achieves 2 dB for a window size of 128 samples, but similarly if the $S$ is lowered, quality improves, for $S = 16$, 25 dB score is achieved. This means that our algorithm works also with assumptions that are different to those standard algorithms.

4.3. Reconstructing the Signal

Once the model has been trained, and that it has provided good enough results on each component, we used the model to reconstruct the signal. We discuss these results.

When $\alpha_0$ are larger $\alpha_A$, the coefficients depends too much on previously computed phase, and the process does not learn from the signal. So, the model may not be apply as it to iterated signal reconstruction. In the case where the coefficients $\alpha_A$ are larger $\alpha_0$, the acceleration of the phase is inferred from the log-magnitude spectrogram, and the results are consistent with those observed during cross-validation on training set.

Although the SNR is close to 10, it is insufficient compared to other algorithms. Griffin and Lim algorithm, when computed for a reasonable amount of time is generally evaluated to around 20 dB. RTISI [3], due to real-time settings performs generally a little bit worst with an SNR of 16 dB. RTSI-LA provides up quality improvement up to 23.12 dB. Our results are obviously very far from Bouvrie and Ezzat [5] who achieved up to 125 dB on some samples.

However, there remain a few case where we are the better than existing algorithms: For small window sizes, when there is no overlap, our algorithm has a positive SNR (Figure 1), while other algorithms have negative SNR, see [5].

5. PERSPECTIVES

The algorithm for phase reconstruction from modified magnitude spectrograms presented in this paper is far not able to reconstruct the phase as well as the Griffin and Lim algorithm does. However, it theoretically requires less constraints on the input spectrogram. Also, the algorithm returns better solutions than other ones, when there is no overlap in-between consecutive frames; at least when the window size is not too large (close to 128).

In further development, we have to explore more features, and other learning techniques, and to find new observations that can be used related to the evolution of the phase through time. The exact tuning for optimal sound quality has to be clarified.

We decided to use a cellular-automata-like process in order to rebuild the phase the magnitude. One other way to reconstruct the phase would have been to consider the phase spectrogram as a Markovian field on which one may run Gibbs sampler or belief propagation algorithm in order to produce the result. However, the computational costs of these alternative solutions is significantly higher. Another way of improving the quality may be to alter the order and the direction in which the computations are performed.

6. REFERENCES


