Multimedia Tools and Applications
https://doi.org/10.1007/s11042-022-12421-1

Audio signal reconstruction using phase retrieval: Implementation and evaluation

Raja Abdelmalek[1] · **Zied Mnasri**[2,3] · **Faouzi Benzarti**[1]

Received: 25 January 2021 / Revised: 25 June 2021 / Accepted: 25 January 2022

Abstract

Phase retrieval has been theoretically proved to be an efficient method for signal reconstruction given only the magnitude spectrum of short time Fourier transform (STFT). Recently, this topic has regained increasing interest for its usefulness in several applications such as compressive sensing, speech synthesis, speech enhancement, source separation, etc. Therefore this paper presents an efficient algorithm for audio signal reconstruction using phase retrieval from the STFT magnitude spectrum, based on an explicit relationship between STFT magnitude and phase. First, the performance of the proposed algorithm is studied for different types of audio signals, i.e. monophonic (speech) and polyphonic (music), in order to tune its parameters. Then, a detailed comparison with the state-of-the-art phase retrieval algorithms is presented. Thus, two types of evaluation are carried out: (a) An objective evaluation is performed using the standard metrics in signal reconstruction, i.e. time-domain segmental signal-to-noise ratio (segSNR), time-frequency domain signal-to-error ratio (SER), and cepstrum-related distance measures, namely log-likelihood ratio (LLR), Itakura-Saito distorsion (IS) and cepstrum distance. Such an evaluation was performed first for the proposed algorithm alone, and then in comparison to state-of-the art methods; (b) a subjective evaluation is conducted with a series of listening tests commonly used in audio quality rating, namely Mean Opinion Score (MOS), Degradation Mean Opinion Score (DMOS) and preference tests. The results of both evaluation protocols confirm the improvement brought by the proposed approach.

A full description of this work with Matlab code is available here (or at https://github.com/zied-mnasri/phase-retrieval)

✉ Zied Mnasri
zied.mnasri@enit.utm.tn

Raja Abdelmalek
raja.abdelmalek@enit.utm.tn

Faouzi Benzarti
faouzi.benzarti@ensit.rnu.tn

[1] SITI Laboratory, ENIT, University of Tunis El Manar, Tunis, Tunisia

[2] Electrical Engineering Department, ENIT, University of Tunis El Manar, Tunis, Tunisia

[3] DIBRIS, University of Genoa, Genova, Italy

Published online: 02 March 2022

Keywords Signal reconstruction · Short-time Fourier transform (STFT) · Spectrogram inversion · Phase retrieval · Objective evaluation · Subjective evaluation

1 Introduction

1.1 Addressed problem

In signal reconstruction problems, it is a common practice to use only the STFT magnitude spectrum and its time sequences whatever phase information is available or not [56]. The problem of phase retrieval, namely signal reconstruction given only the STFT magnitude spectrum, is encountered in several signal processing applications such as signal reconstruction [29], compressive sensing [36, 41], speech synthesis [52], speech enhancement [3, 30, 39] and source separation [19, 35]. Besides, the importance of phase in audio signal perception is still an open research topic. Hence several works have been investigating the effect of phase retrieval on the quality of the reconstructed sound [59].

1.2 Feasibility of phase retrieval

Actually, one of the questions which has always been debated in signal processing, and particularly audio and speech processing, is about the relevance of each of the STFT magnitude and phase spectra in the signal reconstruction process [28]. Whereas in [55] Van Hove et al. argued that only phase spectrum is necessary to reconstruct a signal with satisfactory quality, Griffin and Lim replied that magnitude spectrum is sufficient to reconstruct a signal, through what was called spectrogram inversion, or phase retrieval, i.e. a new phase synthesis given only the magnitude spectrum [16]. However, state-of-the-art phase retrieval algorithms cannot always achieve the totality of the necessary requirements at the same time [51]. For instance, in some algorithms, reconstructing an audio signal given only the magnitude spectrum requires a large number of iterations to reach a satisfactory result [16, 29].

To overcome this problem, some non-iterative algorithms were developed such as the single-pass-spectrogram-inversion algorithm (SPSI), proposed in 2015 by Beauregard et al. [7]. Recently, the phase retrieval problem has been studied for specific signals. Thus, in 2017, Iwen et al. proposed a two-stage sparse phase retrieval strategy that uses a near-optimal number of measurements, which was proved to be computationally efficient and robust to measurement noise [27]. Furthermore, the problem of multi-source phase retrieval from single-channel mixed phaseless STFT measurement was investigated in [20]. Thus, a new model was proposed for the problem of multi-source phase retrieval from a mixed phaseless STFT measurement [9].

1.3 Usefulness of phase retrieval in audio and speech processing

Since the very beginning of audio and speech processing, phase retrieval has been explored to provide alternative solutions to classical signal reconstruction problems. Thus, it was first applied to speech reconstruction, through the use of the phase vocoder [44], and then it was extended to a broad range of applications such as automatic speech recognition [4], speech synthesis [12, 48], speech enhancement [37], single-channel source separation [34, 37], and more recently for low-latency source separation for hearing aids [32]. Also, the importance of phase in audio and speech perception has been the topic of several research works, in

order to understand the effect of phase on the mechanism of hearing. Hence, the analysis of phase was proved to be efficient for onset detection [10] and beat tracking [23] in music signal. In speech signal, sensitivity to phase was revealed important to detect speech polarity [49] and speech imposture [11]. Also, in the study of [43] about applying to speech coding, it was revealed that the human ear is more sensitive to phase than what was believed, as reported by [38].

1.4 Relationship between STFT magnitude and phase

Another approach to resolve the phase retrieval problem is to investigate the relationship between STFT magnitude and phase spectra. Actually, since 1979 the existence of a relationship between the first-order derivatives of STFT magnitude and phase under specified conditions has been proved by Portnoff [45]. Later, in 1984, Yegnanarayana et al. [61] showed an explicit relationship between the amplitude and phase through cepstrum coefficients. However, this relationship holds only in the case in which the minimum or maximum phase signals are applied.

As an alternative, some researchers have returned since a few years to investigate the existence (and the underlying conditions) of such a relationship, partly due to the inability of experimental algorithms to improve the state-of-the-art results. In fact some interaction between the STFT magnitude and phase has already been revealed since 1979 [45]. To continue the development of these early results, Auger proposed an evaluation of the derivatives of the phase and the logarithm of the magnitude of a given STFT with a specific window in 2012 [5]. Recently, in 2017, Shimauchi et al. proved a theoretical basis for the relationship between STFT of magnitude and phase spectra, thoroughly detailed in [50]. In particular, the latter work introduced a set of relationships between the STFT magnitude and phase, which was considered as a particular case in which the Gaussian window is chosen for the STFT operation. Thus, Shimauchi et al. showed that the group delay and the instantaneous frequency, which are calculated as partial derivatives of the phase spectrum, can be explicitly linked to the magnitude spectrum. Consequently, the magnitude and the phase are also directly linked through the group delay or the instantaneous frequency [50].

1.5 Contribution and outline

This article is a continuation of our work that was published in [1]. In that first publication, we were aiming at evaluating the optimal conditions for phase retrieval, based on some state-of-the-art algorithms such as real-time iterative spectrogram inversion (RTISI) algorithm [8]. Thus, we determined the optimal reconstruction parameters necessary to yield better quality of the reconstructed signal. Hence, in that publication the state-of-the-art methods have been reviewed; whereas the present paper aims at proposing a novel signal reconstruction algorithm based on an explicit relationship between the amplitude spectrum and the phase spectrum of the STFT. It should be noted that (a) this theoretic relationship is quite novel, as it has been published in [50], and (b) to the best of our knowledge, this is the first work that utilises this relationship in phase retrieval for audio and speech reconstruction.

The main contribution of the work described in this paper consists in implementing a speech signal reconstruction algorithm using the theoretic relationship between the STFT magnitude and phase established in [50]. This implementation is based on an iterative approach, where a set of parameters are experimentally tuned to improve the quality of the reconstructed speech signal. It should be noted that the preliminary results of this work were published in [2]. However, this work contains the comparison with similar

state-of-the-art algorithms, i.e. based on the relationship between STFT phase and magnitude, and the results of subjective evaluation.

The rest of the paper is organised as follows: Section 2 presents the state-of-the-art signal reconstruction algorithms; Section 3 explains the theoretic relationship between STFT magnitude and phase on which this work is based, and presents the proposed algorithm; Then, Sections 4 and 5 detail the experimental protocol, including the test audio data, the standard metrics and the results of objective and subjective evaluation, respectively. Finally, comments and outlook ideas are drawn in the conclusion.

2 Related work

The problem of signal reconstruction from partial spectral data, i.e. phase retrieval from the spectrogram named also spectrogram inversion, has been addressed since the very beginning of digital signal processing. Since then, a variety of approaches and techniques have been developed.

2.1 Background theory

Since 1980, the problem of signal reconstruction has been studied by Hayes et al. [22] in MIT labs. Their goal was to determine the minimal conditions theoretically required to reconstruct a signal from partial spectral data. Then two theorems appeared:

Theorem 1 *A sequence which is known to be zero outside the interval [0,N-1] is uniquely specified to within a scale factor by (N-1) distinct samples of its phase spectrum in the interval* $0 < \omega < p$ *if it has a z-transform with no zeros on the unit circle or in conjugate reciprocal pairs* [22].

Theorem 2 *Let* $x(n)$ *and* $y(n)$ *be two real, causal, and finite extent sequences with z-transforms which have no zeros on the unit circle.*
If $A_x(x : x_0) = A_y(x : x_0)$ *for all* x *then* $x(n) = y(n)$*, where*

$$A_s(\omega : \omega_0) = \begin{cases} |S_s(\omega)| & if \quad -\omega_0 < \Phi_s(\omega) < \omega_0 + \pi \\ -|S_s(\omega)| & otherwise \end{cases} \tag{1}$$

and $S_s(\omega)$*,* $\Phi_s(\omega)$ *are respectively the magnitude and the phase spectra of the signal* $s(n)$*,* $\omega = 2\pi f$ *where* f *is the frequency.* $A_s(\omega : \omega_0)$ *is then called the signed spectrum of* $s(n)$ [55].

Theorem 1 proves that under a certain condition of stability, a signal can be recuperated to a scale factor given its phase spectrum only, whereas Theorem 2 stipulates that under similar conditions, a signal can be identified to another one if their signed magnitude spectra, defined in (1), are equal. However, the required stability conditions are too tight to be applied to a large category of signals, including audio and speech signals.

2.2 Phase retrieval methods

Since signal reconstruction is an estimation problem, the developed methods can be split into iterative and non-iterative ones. In fact, the iterative methods are based on convergence

criteria, whereas the non-iterative approaches are based on a single-pass process aiming to estimate the missing phase from the original STFT magnitude spectrum, so that the signal can be reconstructed by inverse STFT.

2.2.1 Iterative methods

Typically, signal reconstruction using phase retrieval uses iterative methods to generate the phase spectrum from the STFT magnitude spectrum, and then update the estimation of the signal frame at each iteration until the convergence criteria are met.

Griffin-and-Lim algorithm The iterative Griffin-and-Lim algorithm (GLA) [16] is the most generic phase retrieval-based signal reconstruction algorithm . It proceeds by applying an inverse Fourier transform to converge toward a time-domain signal with the desired spectrum. The distance measure used in this algorithm between the STFT magnitude of both the original signal and the reconstructed one, respectively denoted $|X_\omega(mS,\omega)|$ and $|Y_\omega(mS,\omega)|$ in (2), is minimized at each iteration:

$$D = \sum_{m=-\infty}^{+\infty} \frac{1}{2\pi} \int_{-\pi}^{\pi} (|X_\omega(mS,\omega)| - |Y_\omega(mS,\omega)|)^2 d\omega, \tag{2}$$

where S is the synthesis step size, m is the index of the frame and ω is the Fourier angular frequency.

Let $x^i(n)$ denote the estimated signal $x(n)$ after the i^{th} iteration. The $(i+1)^{th}$ estimate $x^{i+1}(n)$ is obtained by taking the STFT of $x^i(n)$ replacing the magnitude of $X^i_\omega(mS,\omega)$ with the given magnitude $|Y_\omega(mS,\omega)|$ and then finding the signal which STFT is as close as possible to this modified STFT. The estimated signal is updated at each iteration as:

$$x^{i+1}(n) = \frac{\sum_{n=-\infty}^{+\infty} W(ms-n) \int_{\omega=-\pi}^{\pi} \hat{X^i_\omega}(mS,n) e^{-jn\omega} d\omega}{\sum_{n=-\infty}^{+\infty} W^2(mS-n)}, \tag{3}$$

where

$$\hat{X^i_\omega}(mS,n) = |Y_\omega(mS,\omega)| \frac{X^i_\omega(mS,\omega)}{|X^i_\omega(mS,\omega)|}. \tag{4}$$

RTISI and RTISI-LA algorithms Beauregard et al. proposed two real-time algorithms based on the idea of GLA, namely the real-time iterative spectrogram inversion (RTISI) algorithm [8], and the RTISI with look-ahead (RTISI-LA) algorithm [62].

In the RTISI strategy, the signal is reconstructed according to a time-sequential order (frame-by-frame), in opposition to GLA algorithm where all the frames are updated concurrently (cf. Fig. 1). Then the number of transform iterations should be kept to a minimum [8]. The aim of this algorithm is to reconstruct the frame m at each step. The first $(m-1)$ frames of the signal are supposed as already generated and denoted as $y_i(n)$ for $i = 1, ..., m-1$. The m^{th} partial frame comes from the overlap-added results of the estimation of frames $(m-1)$, $(m-2)$ and $(m-3)$ of $y(n)$ while the fourth quarter of the frame m is all zero, given that the synthesis window overlap is fixed at $3/4$ in order to estimate the frame m (for $m > 1$). Then, the update function of the GLA algorithm, cf. (3) and (4), is used at each step to compute the estimate of the current frame m only, instead of updating the estimate of the whole signal $y(n)$ [8].

Beauregard et al. used a zero-initial-phase estimate with the target magnitude spectrum to generate the first frame of $y(n)$ for the estimation of the first frame of the signal. The

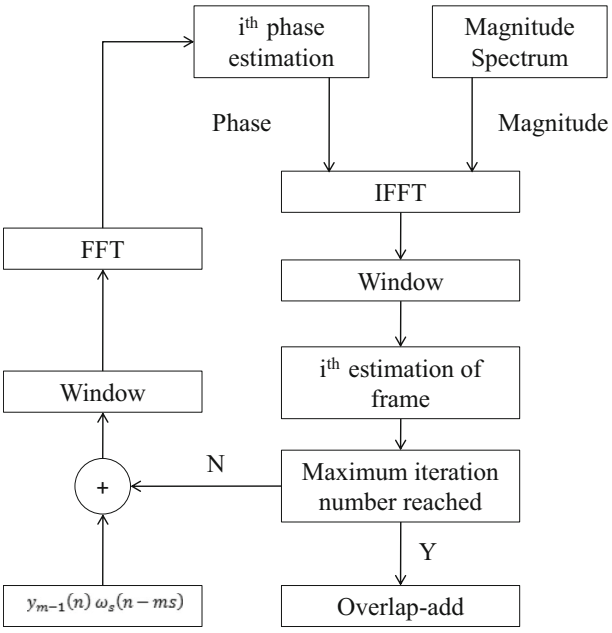


Fig. 1 Real-Time Iterative Spectrogram Inversion (RTISI) algorithm [8]

frame m is estimated and combined with the partial frame $y_{m-1}(n)w(n-mS)$ and the RTISI process continues with the successive frames until the reconstruction of the whole signal. The frame m is estimated from only the previous frames in the RTISI algorithm. In this process, shown in Fig. 1, the future frames are not considered in the signal estimation. In contrast, in the RTISI-LA algorithm, the reconstruction of the current frame m is influenced only by the k future frames. After the estimation of the frame m, it is kept uncommitted until the generation of the frame $(m + k)$ [62].

The optimal parameter setting for iterative algorithms, in particular RTISI, was studied in [1]. It has been shown that using an iteration number $N_{it} = 7$, an overlap rate of 1/2, a Hamming-type window, a frame length equal to 16 ms for a signal sampled at 44.1 KHz or 48 kHz, and equal to 32 ms for a sampling rate of 16 kHz, gives the highest signal-to-error ratio, ($\text{SER}_{\text{dB}} \geq 40$ dB), and the best mean opinion score (MOS ≥ 4) as well. This proves that a fine tuning of the algorithm parameters is also necessary to improve the quality of the reconstructed signals [1, 33].

2.2.2 Non iterative methods

Though iterative algorithms give satisfactory results under certain conditions, the delay due to the iterative implementations is not compatible with real-time requirements. Therefore, some non-iterative signal reconstruction algorithms have recently been proposed.

Single Pass Spectrogram Inversion Algorithm (SPSI), Beauregard et al. proposed a good estimation of the phase spectrum in a single iteration. The SPSI algorithm [7] employs peak picking followed by quadratic interpolation of the magnitude spectrum in order to identify the instantaneous frequency. This technique can also provide an excellent estimation

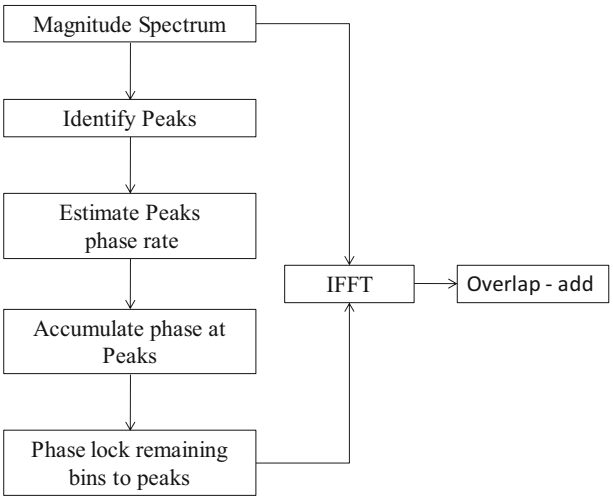


Fig. 2 Single Pass Spectrogram Inversion (SPSI) algorithm [7]

of the initial phase that improves the performance of the iterative spectrogram inversion algorithms. For each frame, SPSI algorithm proceeds as follows (cf. Fig. 2):

(a) The peaks of the magnitude spectrum $|X(mS, \omega_j)|$, $|X(mS, \omega_{j-1})|$ and $|X(mS, \omega_{j+1})|$ are estimated (S is the synthesis step size, m is the frame index and $w_j = 2\pi j/N$ is the angular frequency of the peak and N is the discrete Fourier transform's number of points).

(b) Using a quadratic interpolation, the true values of the bins of the magnitude spectrum are identified using (5)-(8).

$$p = \frac{\alpha - \gamma}{2(\alpha - 2\beta + \gamma)}, \tag{5}$$

where

$$\alpha = |X(mS, \omega_{j-1})|, \tag{6}$$

$$\beta = |X(mS, \omega_j)|, \tag{7}$$

and

$$\gamma = |X(mS, \omega_{j+1})|. \tag{8}$$

The value of p is in the range $[-0.5, 0.5]$. It represents the deviation between the estimated and the true peaks as a proportion of the bin size.

(c) The true frequency of the peak is calculated by $w_j = 2\pi(j + p)/N$ where j indicates the position of the peak bin. The value of the adjusted phase corresponding to this peak is then calculated and the accumulated phase for the bin j is given by $\phi_{m,j} = \phi_{m-1,j} + S_{\omega,j}$.

(d) The phase values of the remaining bins are calculated according to the sign of p.

(e) The signal is reconstructed by combining the phase obtained so forth with the original magnitude [7].

Real-Time spectrogram inversion using Phase Gradient Heap Integration algorithm (RTPGHI) The goal of this algorithm is to estimate the phase spectrum from the magnitude

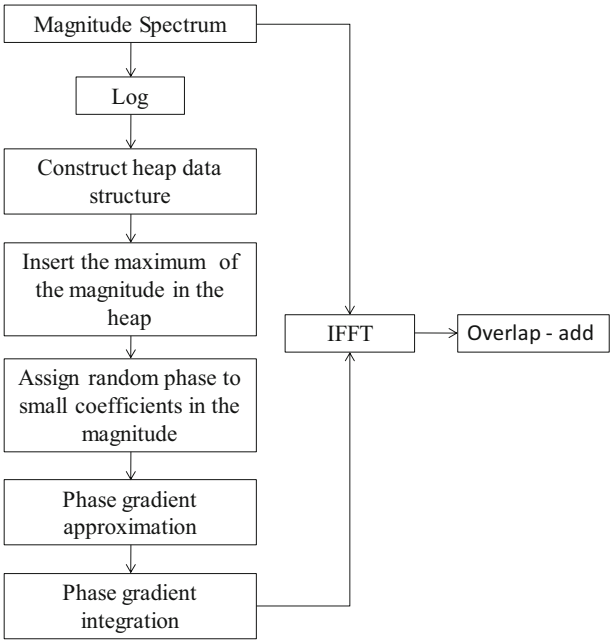


Fig. 3 Real-time phase gradient heap integration algorithm (RTPGHI) [47]

spectrum in order to reconstruct the signal. The reconstruction process is performed in two steps [47], as follows (cf. Fig. 3):

(a) Obtaining the phase gradient via numerical differentiation. Therefore, the discrete STFT phase gradient $\nabla\phi = (\widetilde{\phi}_\omega(m,n), \widetilde{\phi}_t(m,n))$ can be approximated using a centered difference scheme (cf. (9) and (10)):

$$\widetilde{\phi}_\omega(m,n) = -\frac{\delta}{2aM}(A_{\log}(m,n+1) - A_{\log}(m,n-1)), \tag{9}$$

$$\widetilde{\phi}_t(m,n) = \frac{aM}{2\delta}(A_{\log}(m+1,n) - A_{\log}(m-1,n)) + \frac{2\pi am}{M}, \tag{10}$$

where $A_{\log}(m,n) = \log(A(m,n))$, $A(m,n)$ is the discrete STFT magnitude, δ is the time frequency ratio of the window, a is the time step (in number of samples) and M is the number of frequency channels.

(b) Applying a numerical line integration: The gradient integration is done using the cumulative sum. The integration paths are chosen according to the coefficient magnitude.

The algorithm starts off by marking coefficients from the frame n as unknown and continues by inserting coefficients from the frame $(n-1)$ into the heap making them all potential initial points. The integration itself starts by removing the biggest coefficient from the heap and it is used to spread the phase to the only neighbour in the frame n. The algorithm continues until no coefficients with unknown phase are left. The reconstructed phase $\hat{\phi}(m,n)$ is combined with the magnitude such that the STFT $\hat{S}(m,n) = A(m,n)e^{\hat{\phi}(m,n)}$ and the signal $\hat{s}$ is reconstructed using a dual window such us Gabor window and applying the

overlap-and-add (OLA) procedure [47].

2.3 Other methods

In addition to STFT-based methods, signal reconstruction and in particular phase retrieval can be carried out using other types of time-frequency transforms, mainly wavelet transform (WT). Actually, the magnitude spectrum of the WT, called also scalogram, can be inverted in a similar way to the STFT spectrogram to yield an estimated phase, and then a reconstructed signal, through inverse WT. WT can be continuous, discrete, discrete-time or stationary discrete time, with some particularities in each case [18].

In the pioneering work of [25], signal reconstruction is carried out using the CWT magnitude coefficients. To achieve that, least square error estimate (LSEE) is minimized in the same way as for the Griffin-and-Lim algorithm [16]. Lopes & White [31] proposed an algorithm for signal reconstruction using GWT (Generalized WT). In a similar way to [25], the time-domain signal is obtained by applying modified inverse GWT, where at each iteration the scalogram is updated, and its magnitude is replaced by that of the original one. Recently, [40] proposed a faster version of the CWT-based algorithm of [25]. In this method, the phase estimate problem is formulated using a consistency condition, whereas the iterative algorithm is derived using the same process as in [25]. The consistency condition means that the output of the adjacent channels within the overlapping spectral sub-bands have to be consistent, so the waveform reconstructed as $\hat{s} = \mathrm{W}\mathrm{W}^{+}s$ (where W is the CWT matrix) is as close as possible to the original signal s. To achieve that, the error term $O^{(i)} = \hat{s}^{(i)} - \mathrm{W}\mathrm{W}^{+}\hat{s}^{(i)}$ is minimized at each iteration (i). In [40], this improved method is reported to be 100 times faster than the original one [25].

Further reading about DWT and their interpretation can be found in [17]. Also, novel theoretic results regarding the *well-posedness* of phase retrieval as an inverse problem have recently been presented in [58].

3 Proposed approach

The approach that we propose in this paper is based on some fundamental results about the existence of an explicit relationship between STFT magnitude and phase spectra. Such a relationship has been recently proved in the work of [50], based on earlier works of [45] and [5]. Our main contribution consists in leveraging this relationship to build an efficient algorithm for audio signal reconstruction through phase retrieval from the original STFT magnitude spectrum.

3.1 Explicit relationship between STFT magnitude and phase

The relationship between STFT magnitude and phase spectra has been investigated since the beginning of using STFT as an analysis tool. However, no explicit generalized relationships had been proved until recently. Besides, exact estimation of one spectrum, without a perfect knowledge of the other one, is difficult in most cases, as shown hereafter.

In [45], STFT is described as the output of a filter-bank spectrum analyzer, as in (11):

$$X(t,\omega) = \int_{-\infty}^{+\infty} w(t-u)x(u)e^{-ju\omega}du, \tag{11}$$

where $X(t,\omega) = A(t,\omega)e^{j\phi(t,\omega)}$; $A(t,\omega)$ and $\phi(t,\omega)$ are the STFT magnitude and phase, respectively, and w is the analysis window. If w is a Gaussian window then it can be written as in (12):

$$w(t,\omega) = e^{-\frac{t^2}{2\sigma^2}}, \tag{12}$$

where σ is a duration standard deviation. The first order derivatives of STFT magnitude and phase are related by the system of (13) and (14):

$$\frac{\partial \log(A(t,\omega))}{\partial t} + \frac{1}{\sigma^2}\frac{\partial(\phi(t,\omega))}{\partial \omega} = -\frac{T}{\sigma^2}, \tag{13}$$

$$\frac{\partial \phi(t,\omega)}{\partial t} - \frac{1}{\sigma^2}\frac{\partial \log(A(t,\omega)}{\partial \omega} = \omega. \tag{14}$$

In [50], (13) and (14) were reformulated using the definitions of group delay (GD) and instantaneous frequency (IF), given by (15) and (16), respectively:

$$GD(t,\omega) = -\frac{\partial \phi(t,w)}{\partial \omega}, \tag{15}$$

$$IF(t,\omega) = \frac{1}{2\pi}\frac{\partial \phi(t,w)}{\partial t}, \tag{16}$$

so that

$$GD(t,\omega) = \sigma^2\frac{\partial A(t,\omega)}{\partial t} + \frac{T}{2}, \tag{17}$$

and

$$IF(t,\omega) = \frac{1}{2\pi\sigma^2}\frac{\partial A(t,\omega)}{\partial \omega} + \frac{\omega}{2\pi}. \tag{18}$$

This means that group delay and instantaneous frequency, which are by definition phase-related quantities, are also explicitly linked to magnitude [50]. It implies that if the STFT magnitude $A(t,\omega)$ is known, then the STFT phase $\phi(t,\omega)$ can be determined using either the GD-related equation, i.e. (17), or the IF-related one, i.e. (18), thus giving (19) and (20) respectively:

$$\phi(t,w) = -\sigma^2\int \frac{\partial \log(A(t,\omega))}{\partial t}d\omega - \frac{T\omega}{2} + C_1, \tag{19}$$

and

$$\phi(t,w) = \frac{1}{\sigma^2}\int \frac{\partial \log(A(t,\omega))}{\partial \omega}dt + \omega t + C_2. \tag{20}$$

where C_1 and C_2 are constants depending on the initial conditions.

Since our goal is to estimate the phase spectrum $\phi(t,\omega)$ from the magnitude spectrum $A(t,\omega)$ for signal reconstruction using inverse STFT, then (19) and (20) will be transformed into discrete-time domain, where $n = t.f_s$ is the discrete-time index, $k = \omega N/(2\pi f_s)$ is the frequency bin and $N = T.f_s$ is the length of the analysis window. Hence, the discrete group delay (DGD) and the discrete instantaneous frequency (DIF) are calculated by (21) and (22) as in [50]:

$$DGD(n,k) = \frac{\sigma^2 f_s}{2}\log\frac{A_{n+1,k}+\delta}{A_{n-1,k}+\delta} + \frac{N}{2f_s}, \tag{21}$$

and

$$DIF(n,k) = \frac{N}{8\pi^2\sigma^2 f_s}\log\frac{A_{n,k+1}+\delta}{A_{n,k-1}+\delta} + \frac{kf_s}{N}, \tag{22}$$

where δ is a small positive constant used for numerical stability [50]. A recursive solution of (21) and (22) is given by (23) and (24):

$$\hat{\phi}_{n,k} = \hat{\phi}_{n,k-1} - \frac{\pi\sigma^2 f_s^2}{2N} \log \frac{A_{n+1,k}A_{n+1,k-1} + \delta}{A_{n-1,k}A_{n-1,k-1} + \delta} - \pi, \tag{23}$$

$$\hat{\phi}_{n,k} = \hat{\phi}_{n-1,k} + \frac{N}{8\pi^2\sigma^2 f_s} \log \frac{A_{n,k+1}A_{n-1,k+1} + \delta}{A_{n,k-1}A_{n-1,k-1} + \delta} + \frac{2\pi k}{N}. \tag{24}$$

where the initial values in case of real signals are $\hat{\phi}(n, 0) = 0$ or π and $\hat{\phi}(0, k) = 0$ [50].

3.2 Proposed algorithm

The relationships between STFT magnitude and phase spectra were already investigated in [5, 45] and [50] in slightly different forms and using different techniques. Actually, in the aforementioned references, the authors have used different STFT phase conventions so that the explicit equations relating phase and magnitude spectra differ. In this paper, the phase spectrum is estimated from the original magnitude spectrum following the results of [50], in particular using (24). The proposed algorithm aims to retrieve phase from the original STFT magnitude spectrum of each frame and then to reconstruct the frame by inverse STFT (cf. Fig. 4). For the first frame, the reconstruction is performed by the inverse Fourier transform using the original magnitude whereas the phase is initialized to zero (cf. (25)):

$$\begin{cases} A_{n,0} = A_{k=0}^{original}, \\ \hat{\phi}_{n,0} = 0. \end{cases} \tag{25}$$

Assuming that the phase spectrum until the frame $(n-1)$ is reconstructed, the frame $s(n-1, :)$ is shifted, with an appropriate overlap rate. Initially the new samples introduced by the overlap are set to zero. At the first iteration, the estimated phase $\hat{\phi}_{n,k}$ is calculated using (24), then, at each iteration, the phase estimate $\hat{\phi}_{n,k}^{(i)}$ is updated using the same technique used in RTISI algorithm (cf. Fig. 4).

4 Objective evaluation

In this section, we start by presenting the objective evaluation results, in terms of standard error measures used in speech reconstruction problems. This type of evaluation is carried out in two steps: First the performance of the proposed algorithm is assessed individually, and secondly in comparison to state-of-the-art algorithms used for benchmarking.

4.1 Test audio data

For evaluation purposes, we used different types of audio signals, like monophonic signals, i.e. speech, with male and female voices, and polyphonic signals, like classical music. Thus, 3 test sample sets were randomly selected from standard speech and music corpora, namely TIMIT [15], PTDB [42] and SQAM [53]. Information about type of sound, sampling rate, gender of speakers, number and mean duration of samples for each database is detailed in Table 1.

The choice of speech databases was based on three main reasons: i) TIMIT covers a wide range of male and female voices uttering the same speech, which allows distinguishing quality in the subjective tests; ii) PTDB is basically designed to assess pitch-tracking

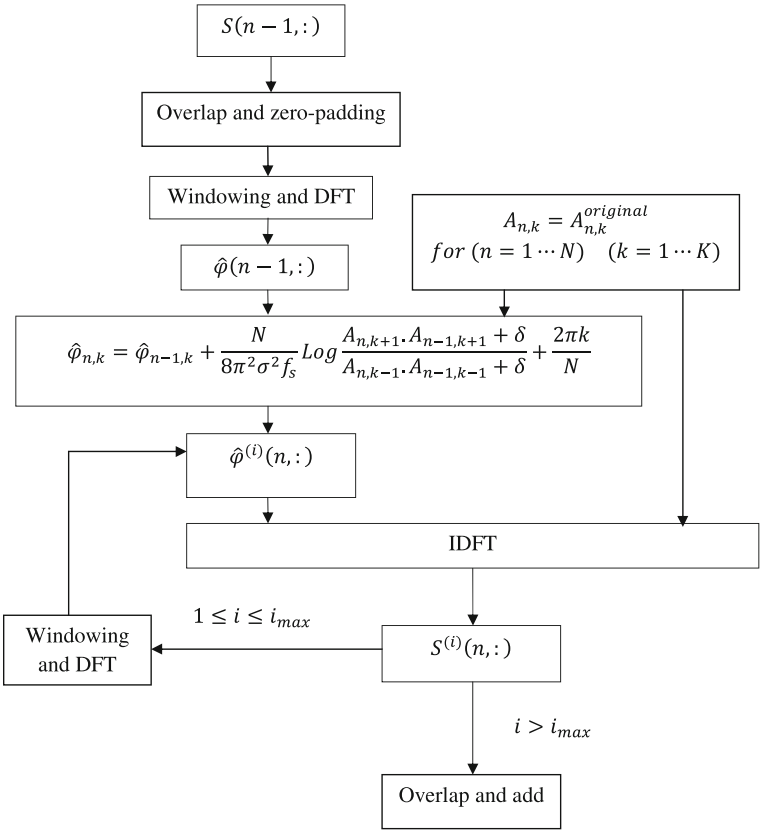


Fig. 4 The proposed signal reconstruction algorithm using phase retrieval: At each frequency channel $k = 1, \ldots, K$, The STFT phase spectrum $\hat{\phi}_{n,k}$ is estimated from the original STFT magnitude spectrum $A_{n,k}$ using (24), and then iteratively updated

algorithms, which makes it potentially suitable for evaluating phase estimation algorithms; iii) Both databases have already been utilized for benchmarking in a wide range of speech reconstruction, enhancement and synthesis algorithms.

Table 1 Test database characteristics

Sound	Type	Database	Sampling rate	Voice	# of samples	Mean duration
Speech	Monophonic	TIMIT [15]	16 KHz	Male	10	3.56 s
				Female	10	3.26 s
		PTDB [42]	48 KHz	Male	10	7.81 s
				Female	10	7.84 s
Music	polyphonic	SQAM [53]	44.1 KHz		10	10.00 s

Regarding the music test signals, similar reasons can be invoked for the choice of SQAM database [53]. Actually this database has been designed for the particular purpose of subjective audio signal processing assessment [53].

4.2 Objective evaluation measures

Generally in signal reconstruction problems, the objective evaluation consists in comparing the original signal to the reconstructed one using standard measures, mainly signal-to-error ratio ($\mathrm{SER_{dB}}$). Since phase retrieval is basically a signal reconstruction problem, we opted to use error measures used in similar problems, mainly speech enhancement, in order to enrich the evaluation process, as $\mathrm{SER_{dB}}$ might not cover all aspects of objective evaluation, such as spectral distortion. Therefore, we also used time-domain and cepstrum-related error measures, classically used for speech enhancement quality assessment, as described in [24].

4.2.1 Time-Frequency domain error measures

Signal-to-error ratio ($\mathrm{SER_{dB}}$): It has been used as a standard error measure for phase retrieval-based signal reconstruction, since the very beginning, as advanced by Griffin and Lim [16]. It is based on STFT and is similar to the frequency-domain signal-to-noise ratio ($\mathrm{SNR_{dB}}$) where noise is replaced by the error between the reconstructed and the original signal. $\mathrm{SER_{dB}}$ provides a good measure of the fidelity of the reconstructed signal, since any distortion of phase or magnitude would be accounted for noise. The signal-to-error ratio ($\mathrm{SER_{dB}}$) is defined as in [16]:

$$\mathrm{SER_{dB}} = 10\log_{10}\left(\frac{\sum_{m=-\infty}^{+\infty}\frac{1}{2\pi}\int_{\omega=-\pi}^{\pi}|X(mS,\omega)|^2 d\omega}{\sum_{m=-\infty}^{+\infty}\frac{1}{2\pi}\int_{\omega=-\pi}^{\pi}(|X(mS,\omega)|-|\hat{X}(mS,\omega)|)^2 d\omega}\right). \tag{26}$$

4.2.2 Time-domain error measures

Segmental SNR ($\mathrm{segSNR_{dB}}$) SNR is a classical performance measure in audio reconstruction problems, such as speech enhancement [24], either in time or in frequency domains. A particular measure of time-domain SNR is called *segmental SNR* and is computed in a frame-wise manner, as in [21]:

$$\mathrm{segSNR_{dB}} = \frac{10}{M}\times\sum_{m=0}^{M-1}\log_{10}\frac{\sum_{i=Nm}^{N(m+1)-1}x^2(i)}{\sum_{i=Nm}^{N(m+1)-1}(x(i)-\hat{x}(i))^2}, \tag{27}$$

where M is the number of temporal frames, N is the frame length, x and $\hat{x}$ are the target and the reconstructed signals, respectively. It should be emphasized that unlike speech enhancement, where the target signal is the clean one, in our case the target signal is the original one, i.e. the signal from which the magnitude spectrum is extracted, regardless the presence of noise.

4.2.3 Cepstrum-related distortion measures

In [24, 30], three types of cepstrum domain error measures are proposed to assess the objective quality of enhanced speech. Since speech enhancement is basically a signal reconstruction problem, where the goal is to approximate the original clean signal, we believe that such an evaluation could be useful for the problem of audio signal reconstruction using

phase retrieval. These measures are computed from the cepstrum using LPC (linear predictive coding). It should be noted that in [24], it is recommended to calculate the LPC vector at the 14^{th} order.

Log-likelihood ratio (LLR) It measures the scalar product of both reconstructed and original signals, here presented as vectors, in the logarithmic domain, cf. (28). It is applied frame by frame and then averaged over all frames [24]. Unlike $\mathrm{SER_{dB}}$ and $\mathrm{segSNR_{dB}}$, the smaller is LLR, the better is the reconstruction quality:

$$\mathrm{LLR}(\vec{\hat{x}}, \vec{x}) = \log \frac{\vec{\hat{x}} R \vec{\hat{x}}^T}{\vec{x} R \vec{x}^T}, \tag{28}$$

where $\hat{x}$ and x are the LPC vector of the original signal frame and of the reconstructed one, respectively, and R is the autocorrelation matrix of the original signal.

Itakura-Saito distortion measure (IS) This is also a measure of the cepstrum distortion. It differs from LLR in using the cepstrum gain as a weighting factor, and as a logarithmic offset, cf. (29). In [24], it is mentioned that IS values should be limited to the range of [0,100] to minimize the number of outliers:

$$\mathrm{IS}(\vec{\hat{x}}, \vec{x}) = \frac{\sigma^2}{\hat{\sigma}^2}\left(\frac{\vec{\hat{x}} R \vec{\hat{x}}^T}{\vec{x} R \vec{x}^T}\right) + \log\left(\frac{\sigma^2}{\hat{\sigma}^2}\right) - 1, \tag{29}$$

where σ^2 and $\hat{\sigma}^2$ are the LPC gain of the original signal and of the reconstructed signal one, respectively.

Cepstrum distance This distance is calculated between the cepstrum coefficients of the reconstructed and the original signals:

$$\mathrm{CEP}(\hat{c}, c) = \frac{10}{\log 10}\sqrt{2\sum_{m=1}^{p}(\hat{c}(m) - c(m))^2}, \tag{30}$$

where c and $\hat{c}$ are the cepstrum coefficients of the original signal and the reconstructed one, respectively, and p is the LPC order (e.g. $p = 14$). The cepstrum coefficients are obtained from LPC coefficients $\{a_m\}_{m=1,\ldots,p}$ such that:

$$c(m) = a_m + \sum_{k=1}^{m-1}\frac{k}{m}c(k)a_{m-k}, \quad 1 \le m \le p. \tag{31}$$

4.3 Performance of the proposed algorithm

To evaluate the performance of the proposed algorithm, $\mathrm{SER_{dB}}$ was measured on the aforementioned test sets, using different values of the parameters, i.e. the Gaussian window σ parameter, the overlap rate, the analysis window length and the number of iterations. Tables 2 and 3 show that for all types of signals, i.e. monophonic (speech) and polyphonic (music), and for all test databases, i.e. TIMIT, PTDB and SQAM, the maximum $\mathrm{SER_{dB}}$ is asymptomatically reached from 5 iterations. Also, experiments have shown that the following parameter settings: the duration standard deviation of the Gaussian window in (24) $\sigma = 2.5$, a Hamming window of 32 ms and an overlap rate of 50%, give a satisfactory result, i.e. $\mathrm{SER_{dB}} \ge 40$ dB.

Table 2 Error measures by type of sound (M: Male, F: Female)

Error	Sound	Voice	# of iterations				
Measure			1	5	10	15	20
SER_{dB}	Speech	M	24.0	40.2	40.3	39.2	38.3
		F	25.0	41.4	40.3	39.0	38.3
	Music		24.3	40.6	39.4	37.8	36.6
$segSNR_{dB}$	Speech	M	0.53	0.19	0.41	0.51	0.59
		F	0.50	0.20	0.38	0.47	0.52
	Music		0.49	0.27	0.39	0.56	0.57
LLR	Speech	M	6.0e-3	2.8e-3	2.3e-3	2.4e-3	2.4e-3
		F	5.0e-3	2.5e-3	2.1e-3	2.3e-3	2.5e-3
	Music		0.62	0.63	0.64	0.64	0.64
IS	Speech	M	0.04	0.02	0.02	0.03	0.03
		F	0.04	0.02	0.02	0.03	0.03
	Music		1.81	1.77	1.78	1.78	1.79
CEP	Speech	M	3.5e-3	1.5e-3	1.2e-3	1.4e-3	1.5e-3
		F	2.7e-3	1.3e-3	1.4e-3	1.5e-3	1.7e-3
	Music		0.14	0.15	0.15	0.15	0.15

Also, Table 2 shows that the proposed algorithms reconstructs speech waveforms slightly better than music. This may be due to the monophonic aspect of speech, whereas music is firstly a multi-pitch signal, and secondly may have more variability in frequency than

Table 3 Error measures by source database (T: TIMIT [15], P: PTDB [42], S: SQAM [53])

Error	Sound	Database	# of iterations				
Measure			1	5	10	15	20
SER_{dB}	Speech	T	22.1	41.5	42.5	42.4	42.0
		P	26.8	40.1	38.1	35.9	34.6
	Music	S	24.3	40.6	39.4	37.8	36.6
$segSNR_{dB}$	Speech	T	0.52	-0.03	0.11	0.18	0.24
		P	0.51	0.42	0.68	0.81	0.87
	Music	S	0.49	0.27	0.39	0.57	0.57
LLR	Speech	T	5.1e-3	3.5e-3	3.7e-3	4.1e-3	4.4e-3
		P	5.9e-3	1.8e-3	0.8e-3	0.6e-3	0.6e-3
	Music	S	0.62	0.63	0.64	0.64	0.64
IS	Speech	T	0.05	0.02	0.02	0.02	0.03
		P	0.04	0.02	0.03	0.03	0.03
	Music	S	1.81	1.77	1.78	1.78	1.79
CEP	Speech	T	3.8e-3	1.5e-3	1.3e-3	1.6e-3	1.8e-3
		P	2.4e-3	1.3e-3	1.3e-3	1.4e-3	1.4e-3
	Music	S	0.14	0.15	0.15	0.15	0.15

speech. Regarding databases, TIMIT voices perform better than PTDB ones (cf. Table 3). This can be also explained by the particular diversity of pitch in PTDB voices.

Finally, it is worth noting that in both Tables 2 and 3, the cepstrum-related errors, i.e. LLR, IS and CEP, are much higher for music. This may be due to the fact that the cepstrum is rather fitted to speech perception than to other audio signals such as music.

4.4 Comparison to state-of-the-art algorithms

The performance of the proposed algorithm is compared to state-of-the-art algorithms, whether iterative such as LGLA [29] and RTISI [8], or non-iterative like SPSI [7] and RTPGHI [47] with respect to: (a) the type of sound and the test database, (b) the number of iterations, and (c) the computational cost.

For so doing, we used the implementation of these algorithms provided in the Phase Retrieval Toolbox [46]. It should be mentioned that all tested algorithms were run using the same parameter settings, i.e. a Hamming window of 32 ms-duration and an overlap rate equal to 50%.

4.4.1 Comparison by type of sound and by database

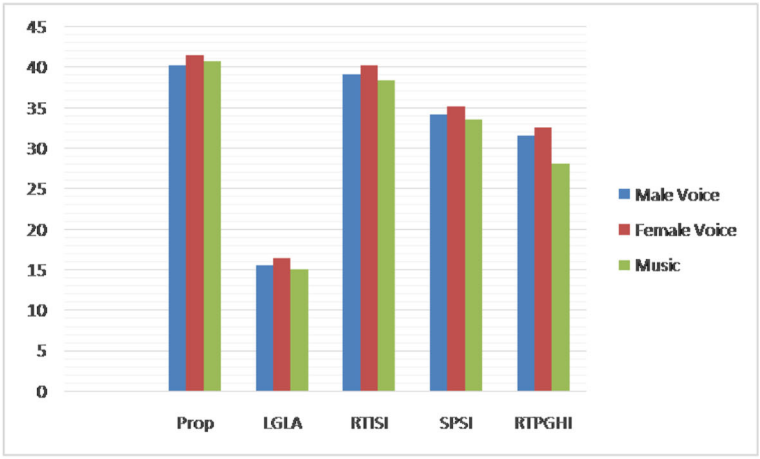
Similar results come up from Fig. 5, where the proposed method is compared to the state-of-the-art algorithms for each type of sound and for each database, respectively. Figure 5a shows that the proposed method outperforms the state-of-the-art algorithms for all types of sound. In addition, speech signals in general and male voices in particular are better reconstructed using the proposed method, whereas music signal reconstruction is slightly less accurate. This matches with the hypothesis that the phase spectrum of polyphonic music is more difficult to reconstruct, since this type of signals has essentially a multi-pitch.

In a similar way, Fig. 5a shows that the proposed method outperforms the other ones, both on speech and music databases. It can also be noticed that for most algorithms, including the proposed one, speech databases signals, i.e. TIMIT [15] and PTDB [42], are better reconstructed than music database signals, i.e. SQAM [53]. This confirms that the more pitch interval is extended, the less phase reconstruction is accurate.

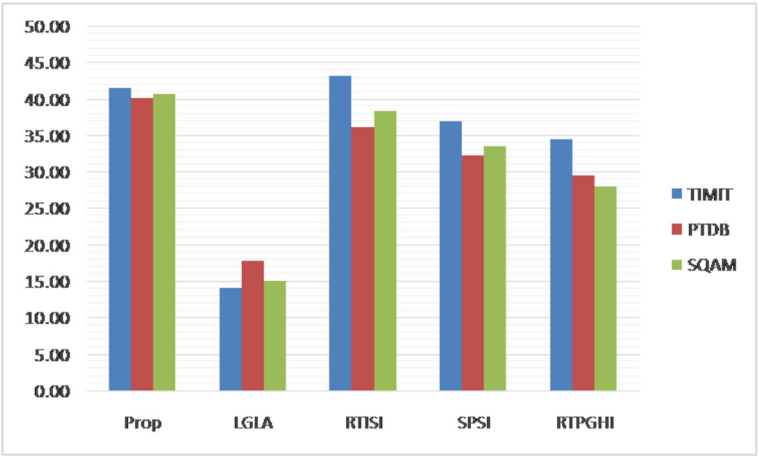
4.4.2 Comparison by number of iterations

Some examples of signal and phase reconstruction using the benchmarking algorithms are given in Fig. 6. First, signal reconstruction (cf. Fig. 6i) shows that when the same number of iterations is used for all iterative algorithms, the proposed algorithm (Prop.) reconstructs the signal better than the other ones, i.e. LGLA [29] and RTISI [8]. Besides, when (Prop.) runs at only one iteration, its performance is not far away from non-iterative algorithms, i.e. SPSI [7] and RTPGHI [47]. The same behaviour can be noticed when looking to the unwrapped phase reconstructed by the different algorithms (cf. Fig. 6ii).

Effect of the number of iterations on SER error measures In terms of $\mathrm{SER}_{\mathrm{dB}}$, Table 4 shows that the proposed algorithm (Prop.) is able to outperform the iterative methods, ie. RTISI [8] and LGLA [29], for the same number of iterations, on all of the three databases, i.e. for different types of signals and using different sampling rates. In comparison to non-iterative methods, i.e. SPSI [7] and RTPGHI [47], the proposed algorithm is doing better when it is executed at 5 iterations (cf. Table 4). However, it is outperformed when it runs at only one iteration, for both monophonic and polyphonic signals (cf. Fig. 7).



(a) by type of sound

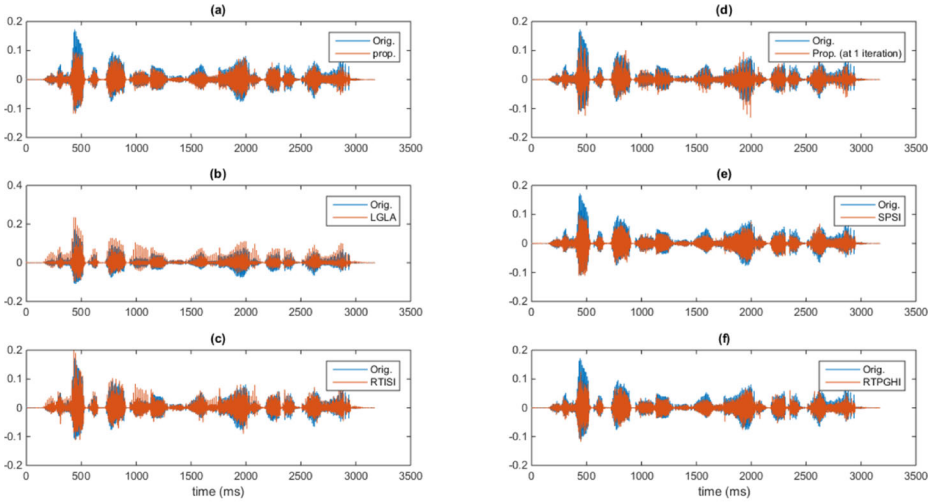


(b) by database

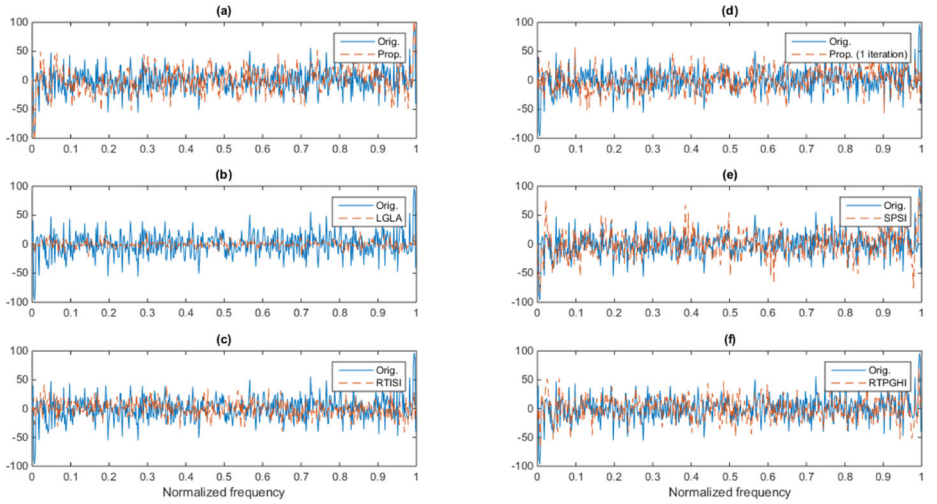
Fig. 5 Comparison of SER_{dB} error by database (5 iterations for (Prop.), LGLA [29] and RTISI [8])

This may be explained by the original iterative scheme which the proposed algorithm has been implemented with. Actually, the first iteration serves only for a first estimation of the phase spectrum from the original magnitude spectrum using (24), whereas this estimation is refined through the iterative process. Tracking the performance of iterative algorithms through the number of iterations, it could be easily noticed that the proposed algorithm is outperforming iterative algorithms for a small number of iterations ($N_{iter} < 10$). Actually, RTISI [8] needs at least 15 iterations to reach the same performance for speech (cf. Fig. 7a), and 20 iterations for music (cf. Fig. 7b).

Finally, we notice that SER_{dB} error reaches its peak at 5 iterations and then slows down for all iterative algorithms, including (Prop.) (cf. Fig. 7). In fact, it was expected to see SER_{dB} rather increasing than decreasing when the number of iterations increases, however the opposite phenomenon has been noticed. To understand the reasons of such a behaviour, the cepstrum-related error measures are also analyzed.



(i) Original vs. reconstructed signal



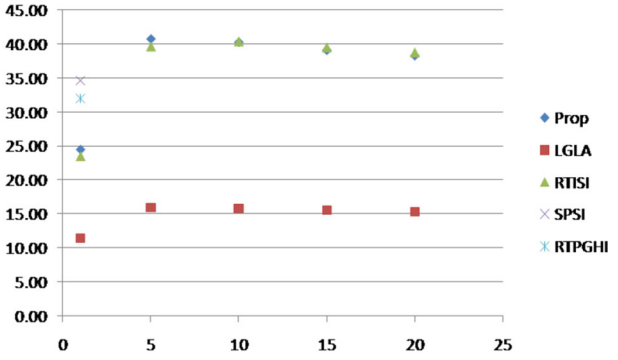
(ii) Original vs. reconstructed unwrapped phase

Fig. 6 Reconstruction examples: (i) original vs. reconstructed speech waveforms and (ii) their respective unwrapped phase with: (a) the proposed algorithm (Prop.) executed at 5 iterations, a Gaussian window with $\sigma = 2.5ms$ for (24), and overlap rate of 1/2, (b) LGLA algorithm executed at 5 iterations, (c) RTISI algorithm, executed at 5 iterations and overlap rate of 50%, (d) the proposed algorithm executed at only 1 iteration, (e) Non-iterative SPSI algorithm, (f) Non-iterative RTPGHI algorithm

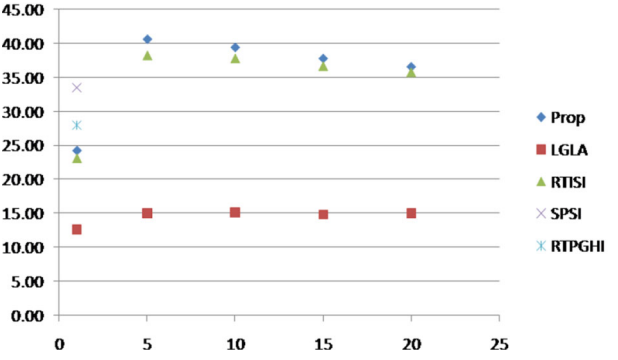
Effect of the number of iterations on cepstrum-related error measures Figure 8 shows the evolution of cepstrum-related errors vs. the number of iterations. Cepstrum-related errors, i.e. LLR, IS and CEP, are also checked for the each number of iterations for all iterative algorithms. Figure 8 shows clearly that LLR error is slowly increasing with the number of iterations. Actually, the purpose of the iterative processing is to decrease the phase estimation error, but apparently, too many iterations seem to increase the spectral distortion, which is expressed by the cepstrum distance measures. This might account for an

Table 4 Comparison of error measures by type of sound and by algorithm using 5 iterations for iterative algorithms (Prop, LGLA [29] and RTISI [8]) and one pass for non-iterative algorithms (SPSI [7] and RTPGHI [47])

Sound	Voice	Algorithm	SER_{dB}	$segSNR_{dB}$	LLR	IS	CEP
Speech	Male	(Prop.)	40.19	0.19	2.8e-3	0.02	1.5e-3
		LGLA	15.45	1.74	6.3e-3	0.11	5.2e-3
		RTISI	39.00	-2.72	0.5e-3	0.01	1.8e-3
		SPSI	34.11	0.68	0.1e-3	0.01	1.0e-3
		RTPGHI	31.47	0.33	0.8e-3	0.01	1.4e-3
	Female	(Prop.)	41.38	0.20	2.5e-3	0.02	1.3e-3
		LGLA	16.41	0.48	10.6e-3	0.10	6.2e-3
		RTISI	40.19	-0.71	1.1e-3	0.01	2.4e-3
		SPSI	35.07	0.82	2.3e-3	0.01	1.8e-3
		RTPGHI	32.50	0.63	2.0e-3	0.01	2.3e-3
Music		(Prop.)	40.62	0.27	0.63	1.77	0.15
		LGLA	14.97	2.47	0.90	2.68	0.21
		RTISI	38.25	0,01	0.79	1.78	0.21
		SPSI	33.48	1.57	0.74	2.22	0.22
		RTPGHI	28.00	1.28	0.74	2.35	0.20

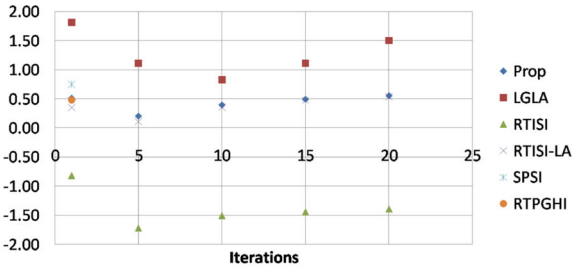


(a) Monophonic signals

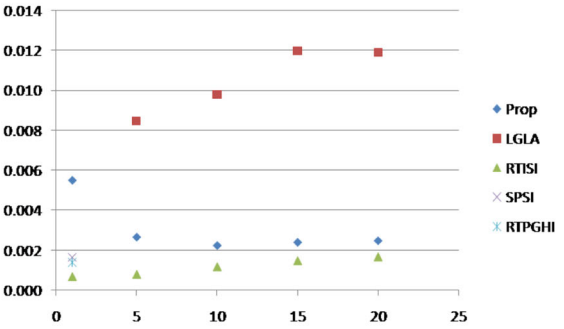


(b) Polyphonic signals

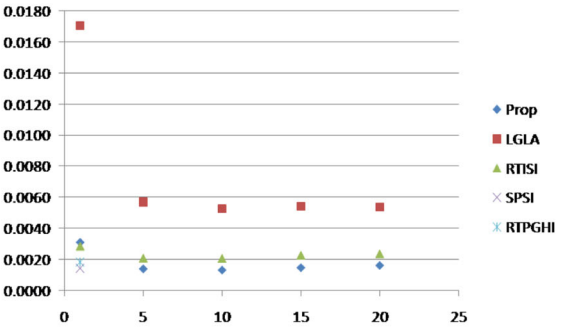
Fig. 7 Comparison of SER_{dB} error signals by number of iterations for each algorithm: from 1 to 20 iterations for (Prop.), LGLA [29] and RTISI [8], 1 iteration for SPSI [7] and RTPGHI [47]



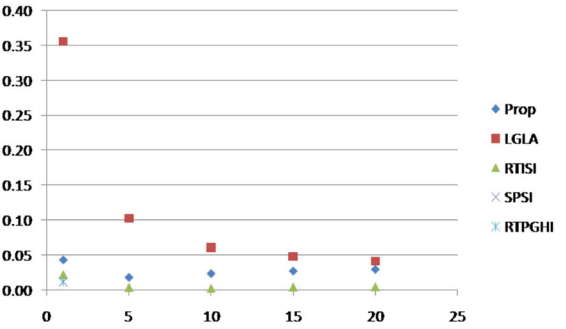
(a) $\text{segSNR}_{\text{dB}}$



(b) LLR



(c) CEP



(d) IS

Fig. 8 Comparison of $\text{segSNR}_{\text{dB}}$ error and cepstrum-related errors (LLR, CEP and IS) vs. number of iterations by algorithm for monophonic signals only

important finding, since the global belief in phase retrieval was that increasing iterations should improve reconstruction quality [3]. However, these experiences show that there is a tradeoff between the reconstruction error and the cepstrum distortion.

4.4.3 Comparison by computational cost

In order to compare the proposed algorithm to the state of the art in terms of computational cost, the time required to reconstruct one second of audio has been calculated at the same conditions, i.e sampling rate, STFT window length and overlap for all algorithms (Table 5). Besides, for iterative algorithms, the number of iterations was varied in the same way as in the previous experiences. However, it should be noted that the reported results apply only for an STFT with a window length of 32 ms and overlap rate of 50%. Results could have been different for other STFT settings, but we opted to keep the same conditions that has been set for the evaluation mentioned above (cf. Tables 2, 3, 4).

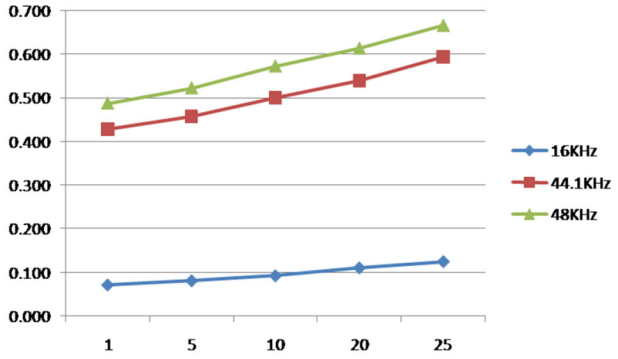
Moreover, Table 5 shows the following facts:

- Non-iterative algorithms are clearly faster, not only because they proceed in a single pass, but also for the absence of STFT in their processing. Actually, even though STFT is a powerful spectrogram inversion tool, it requires more computational complexity, typically $\mathcal{O}(N \log N)$, which makes it difficult running such a method in real-time mode.
- Amongst the iterative STFT-based algorithms, (Prop.) is doing much better than the classical Griffin-and-Lim method [29], and equally to RTISI [8], especially for a sampling rate of 16KHz.
- By crossing results of Table 5 and the different error measures in Table 4, it looks obvious that computational speed should not be the only criterion to evaluate signal

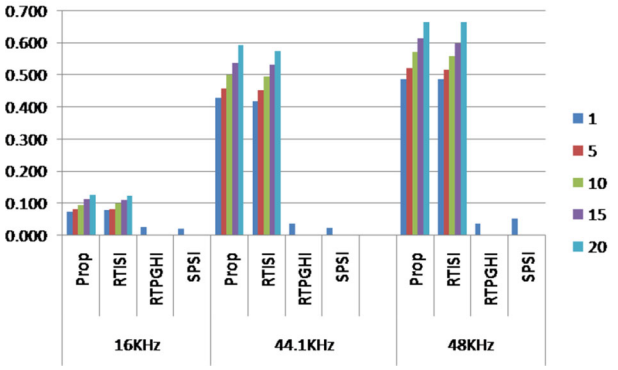
Table 5 Mean computation time (sec) vs. sampling rate (KHz) by number of iterations for 1 sec of reconstructed audio for each algorithm, using an STFT with 32-msec STFT window with 50% overlap for iterative algorithms (Prop., LGLA [29] and RTISI [62]) and one pass for non-iterative algorithms (SPSI [7] and RTPGHI [47])

Algorithm	Sampling	# of iterations				
	rate(KHz)	1	5	10	15	20
Prop.	16	**0.071**	**0.081**	**0.093**	**0.111**	**0.125**
	44.1	**0.428**	**0.458**	**0.501**	**0.539**	**0.594**
	48	**0.488**	**0.523**	**0.573**	**0.614**	**0.666**
LGLA [29]	16	0.239	1.092	2.150	3.208	4.262
	44.1	0.748	3.605	7.349	10.853	14.458
	48	0.691	3.309	6.533	9.910	13.125
RTISI [8]	16	0.077	0.079	0.100	0.110	0.123
	44.1	0.417	0.451	0.494	0.533	0.574
	48	0.486	0.517	0.560	0.599	0.665
SPSI [7]	16	0.018				
	44.1	0.022				
	48	0.050				
RTPGHI [47]	16	0.023				
	44.1	0.034				
	48	0.036				

Bold entries indicate the performance of the proposed method (Prop)



(a) For the proposed algorithm



(b) For benchmarking algorithms

Fig. 9 Mean reconstruction time (sec) vs. number of iterations (from 1 to 20) for 1sec of audio at different sampling rates for all benchmarking algorithms (Prop. and RTISI [8], 1 iteration for SPSI [7] and RTPGHI [47]) (LGLA [29] is not mentioned for its higher reconstruction time, cf. Table 5)

reconstruction methods. Actually, even though non-iterative algorithms are faster, they fail to provide lower error measures.

– Another positive fact in favor of iterative algorithms, notwithstanding their slower execution, is revealed by the slow increase of the computation time vs. the number of iterations. In particular, for (Prop.) and RTISI, the evolution of the computational time vs. the number of iterations seems to follow a logarithmic scale.

However, we believe that a further investigation of the effect of increasing the number of iterations and the sampling rate on the computational cost should be undertaken, since this problem can be approached not only in a temporal standpoint, i.e. computation time, but also for spatial considerations, i.e. required memory (Fig. 9).

5 Subjective evaluation

According the ITU recommendations [26], the quality of audio signal should be evaluated not only objectively, using standard quantitative measures, but also subjectively through qualitative assessment by human listeners.

5.1 Subjective evaluation protocol

Subjective evaluation for audio signal quality is mostly based on ratings by human auditors. The principle consists in listening to the sound and rating the quality in terms of intelligibility and naturalness, or of degradation, according to a scale ranging from 0 to 5 for example. There are two important steps to conduct such a test: collecting responses from subjects and achieving a statistical analysis to obtain the final assessment [6]. To estimate the performance of an audio system, three subjective measures are commonly used: the mean opinion score (MOS) [26, 57], the degradation mean opinion score (DMOS) [13, 54], and the preference test [14].

10 persons participated in the subjective evaluation, equally divided into male and female. The participants were selected as not specialized in acoustics, phonetics or any related field, to avoid any interference. Also, they were not aware of the purpose of the review, to avoid any biased decision. They are all between 20 and 50 year-old with healthy conditions: None of them presents any hearing or any other speech-related impairment, such as dyslexia or dysarthria. All of them all used a headphone with controllable volume. The listening test was blind, as the participants did not know whether signals were original or reconstructed, nor how they were generated. Also, before starting the test, every participants had to listen to some signals as a training.

5.2 MOS test results

Mean opinion score is the most popular method in subjective assessment of audio signal quality. MOS score of the each test sample is the mean of the scores collected from listeners [26]. In MOS test, listeners are free to assign their score to audio signal quality. Nevertheless, this may present a serious disadvantage because the variation of individual listener's scales can result in a bias in the listeners judgments [57].

To carry out this evaluation, 10 signals were selected from the test sample set and reconstructed in four versions, each corresponding to an algorithm, namely: a) simple reconstruction by inverse STFT using the original STFT magnitude and phase spectra (Rec.), b) the proposed algorithm (Prop.), c) RTISI algorithm [8], d) SPSI algorithm [7]. Actually, only RTISI and SPSI algorithms were selected from the benchmarking methods based on their performance in the objective evaluation, since they were the best performing ones amongst the iterative and the non-iterative methods, respectively. Also we opted to use a signal reconstructed with original magnitude and phase (Rec.) instead of the original one, to avoid the bias than can be caused by the reconstruction process.

Then the listeners were asked to rate these signals using the MOS score grades (cf. Table 6) by answering the following question: *How do you judge the overall quality of the audio signal that you have just heard?*

The results obtained from this evaluation are illustrated in Fig. 10b. It shows clearly that the proposed algorithm outperforms RTISI and SPSI algorithms in terms of subjective quality. Even better, the score of the proposed algorithms is not too far away from that of the signal reconstructed using original magnitude and phase (Rec.).

5.3 DMOS Test results

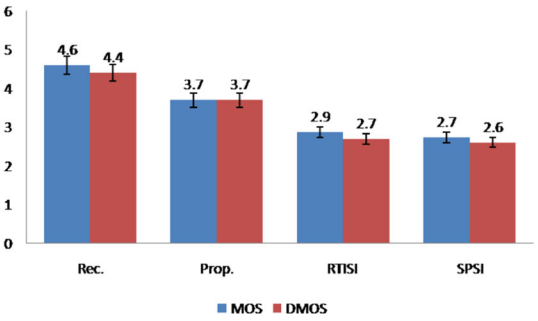
DMOS is a measure of annoyance or degradation level by comparing the reconstructed audio signal to the original one. Unlike MOS test, the original signal is provided and mentioned, then DMOS gives greater sensitivity than MOS in evaluating speech degradation

Table 6 Scales of speech quality (MOS) and speech degradation (DMOS)

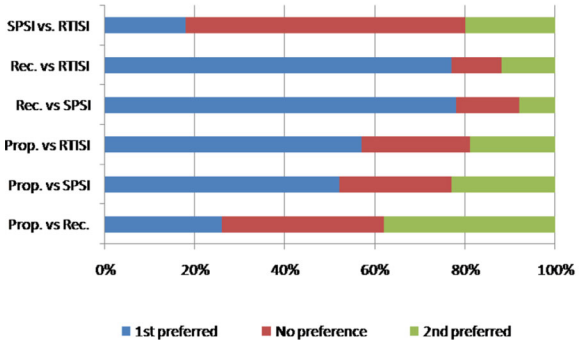
Score	Speech quality	Speech degradation
5	Excellent	Inaudible
4	Good	Audible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Very poor	Very annoying

[60]. Table 6 presents DMOS grades and their corresponding meaning. This evaluation was carried out by presenting a reconstructed signal (by each algorithm) and its original version at each time. The auditors were asked to answer the following question: *How do you judge the degradation of the reconstructed version of the signal in comparison to the original one?*

The results obtained from this evaluation are illustrated in Fig. 10b. They are very similar to those obtained by the MOS test, as the proposed algorithm is ranked second after (Rec.), i.e. the signal reconstructed using original magnitude and phase.



(a) MOS and DMOS test results with 95% confidence interval



(b) Preference test

Fig. 10 Subjective evaluation statistics for the methods selected in terms of objective evaluation scores (Rec: signal reconstructed with original STFT phase and magnitude spectra, (Prop.): The proposed algorithm, RTISI [8]: state-of-the-art iterative algorithm, SPSI [47]: state-of-the-art non-iterative algorithm)

5.4 Preference test results

The preference test is useful to rate which algorithm is preferred to the other ones, in a one-to-one confrontation. This evaluation was carried out by presenting a pair of signals, each reconstructed by a different algorithm at each time. Then, the auditors were asked to indicate which version of the reconstructed signal they prefer. The results obtained from this evaluation are illustrated in Fig. 10b. The preference test results show that: i) The version reconstructed by the proposed algorithm (Prop.) is much preferred to those of RTISI and SPSI; ii) The reconstructed signal with original magnitude and phase (Rec.) is much preferred to all algorithms, but with a lesser degree to the version given by the proposed algorithm, iii) SPSI and RTISI are perceived nearly the same, iv) Looking to the score of preference between the proposed version (Prop.) and that reconstructed with original data (Rec.), only less than 40% find that (Rec.) is clearly better, hence more more than 60% find that they are at least equal, including 25% who find (Prop.) better than (Rec.). This proves that the proposed algorithm succeeds to provide a satisfactory phase estimation.

6 Conclusion

This paper studied a signal reconstruction method using phase estimation from the STFT magnitude spectrum. As presented in the works of [5, 45, 50], phase retrieval can be efficiently justified by establishing theoretic relationships between STFT magnitude and phase spectra. Based on these works, an algorithm was developed and evaluated in this paper. The main novelty of this algorithm consists in updating the phase estimate using an explicit relationship between phase and magnitude to estimate the phase from the original STFT magnitude.

To assess the performance of the proposed method, quantitative and qualitative tests were undertaken, first for the proposed algorithm on its own, and secondly in comparison to state-of-the-art methods. Both types of evaluation were carried out using different audio databases, covering monophonic signals (speech), with male and female voices, and polyphonic signals (music). The objective evaluation was achieved through calculating different types of error measures, whether in the time domain, i.e. segmental SNR ($\text{segSNR}_{\text{dB}}$), or in time-frequency domain, i.e. the signal-to-error ratio (SER_{dB}). Besides, cepstrum-related measures, such as log-likelihood ratio (LLR), Itakura-Saito distortion (IS) and cepstrum distance (CEP), were computed to quantify the cepstrum distortion. The tests show that the proposed algorithm is able to outperform similar state-of-the-art iterative algorithms, such as LGLA [29] and RTISI [8] for the same number of iterations. Notwithstanding it is outperformed by non-iterative like SPSI [7] and RTPGHI [47] when it is executed at only one iteration, it provides better results from a small number of iterations ($N_{\text{iter}} \leq 5$).

Subjective tests (MOS, DMOS and preference test) were conducted to assess the perception of the reconstructed signals by human listeners. These tests have confirmed the results obtained by the objective evaluation, since only the signal version reconstructed with original magnitude and phase spectra (Rec.) was preferred to the version reconstructed by the proposed algorithm (Prop.). An objective interpretation could be the effect of using an explicit relationship between STFT magnitude and phase, in order to estimate the phase in the proposed algorithm, whereas RTISI algorithm tries to retrieve the phase using a convergence criterion, and SPSI is based on an approximation of the phase from the magnitude peak frequency bins.

As an outlook, future work would focus on using this proposed algorithm in audio signal applications such as noise reduction, speech enhancement, source separation, etc. Also, compressive sensing is potentially a relevant application. Another question which requires to be addressed with a higher interest consists in transforming the proposed method into a non-iterative process without loosing its performance.

References

1. Abdelmalek R, Mnasri Z, Benzarti F (2018) Determining the optimal conditions for signal reconstruction based on stft magnitude. Int J Speech Technol 21(3):619–632
2. Abdelmalek R, Mnasri Z, Benzarti F (2018) Signal reconstruction based on the relationship between stft magnitude and phase spectra. In: International conference on the sciences of electronics, technologies of information and telecommunications, Springer, pp 24–36
3. Alsteris LD, Paliwal KK (2007) Iterative reconstruction of speech from short-time fourier transform phase and magnitude spectra. Comput Speech Lang 21(1):174–186
4. Alsteris LD, Paliwal KK (2007) Short-time phase spectrum in speech processing: a review and some experimental results. Digital Signal Process 17(3):578–616
5. Auger F, Chassande-Mottin É, Flandrin P (2012) On phase-magnitude relationships in the short-time fourier transform. IEEE Signal Process Lett 19(5):267–270
6. Barnwell IIITP, Clements M., Quackenbush S. (1988) Objective measures for speech quality testing
7. Beauregard GT, Harish M, Wyse L (2015) Single pass spectrogram inversion. In: 2015 IEEE International conference on digital signal processing (DSP), IEEE, pp 427–431
8. Beauregard GT, Zhu X, Wyse L (2005) An efficient algorithm for real-time spectrogram inversion. In: Proceedings of the 8th international conference on digital audio effects, pp 116–118
9. Bendory T, Eldar YC, Boumal N (2017) Non-convex phase retrieval from stft measurements. IEEE Trans Inf Theory 64(1):467–484
10. Davies ME, Plumbley MD (2007) Context-dependent beat tracking of musical audio. IEEE Transactions on Audio, Speech, and Language Processing 15(3):1009–1020
11. De Leon PL, Hernaez I, Saratxaga I, Pucher M, Yamagishi J (2011) Detection of synthetic speech for the problem of imposture. In: 2011 IEEE International conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 4844–4847
12. Degottex G, Erro D (2014) A measure of phase randomness for the harmonic model in speech synthesis. In: Fifteenth annual conference of the international speech communication association
13. Dimolitsas S, Corcoran FL, Ravishankar C (1995) Dependence of opinion scores on listening sets used in degradation category rating assessments. IEEE Trans Speech Audio Process 3(5):421–424
14. Emiya V, Vincent E, Harlander N, Hohmann V (2011) Subjective and objective quality assessment of audio source separation. IEEE Trans Aud Speech Lang Process 19(7):2046–2057
15. Garofolo JS (1993) Timit acoustic phonetic continuous speech corpus. Linguistic Data Consortium, 1993
16. Griffin D, Lim J (1984) Signal estimation from modified short-time fourier transform. IEEE Trans Acoust Speech Sign Process 32(2):236–243
17. Guido RC (2017) Effectively interpreting discrete wavelet transformed signals [lecture notes]. IEEE Signal Proc Mag 34(3):89–100
18. Guido RC, Pedroso F, Furlan A, Contreras RC, Caobianco LG, Neto JS (2020) Cwt× dwt× dtwt× sdtwt: Clarifying terminologies and roles of different types of wavelet transforms. Int J Wavelets Multiresolution and Info Process 18(06):2030001
19. Gunawan D, Sen D (2010) Iterative phase estimation for the synthesis of separated sources from single-channel mixtures. IEEE Signal Process Lett 17(5):421–424
20. Guo Y, Wang T, Li J, Wang A, Wang W (2019) Multiple input single output phase retrieval. Circ Syst Sign Process 38(8):3818–3840
21. Hansen JH, Pellom BL (1998) An effective quality evaluation protocol for speech enhancement algorithms. In: Fifth international conference on spoken language processing
22. Hayes M, Lim J, Oppenheim A (1980) Signal reconstruction from phase or magnitude. IEEE Trans Acoust Speech Sign Process 28(6):672–680
23. Holzapfel A, Stylianou Y (2008) Beat tracking using group delay based onset detection. In: ISMIR-International conference on music information retrieval, ISMIR, pp 653–658

24. Hu Y, Loizou PC (2007) Evaluation of objective quality measures for speech enhancement. IEEE Trans Audio Speech Lang Process 16(1):229–238
25. Irino T, Kawahara H (1993) Signal reconstruction from modified auditory wavelet transform. IEEE Trans Sign Process 41(12):3549–3554
26. ITU-T RP (1996) 861:" objective quality measurement of telephone-band (300-3400 hz) speech code
27. Iwen M, Viswanathan A, Wang Y (2017) Robust sparse phase retrieval made easy. Appl Comput Harmon Anal 42(1):135–142
28. Laroche J, Dolson M (1997) Phase-vocoder: About this phasiness business. In: Proceedings of 1997 workshop on applications of signal processing to audio and acoustics, IEEE, pp 4–pp
29. Le Roux J, Kameoka H, Ono N, Sagayama S (2010) Fast signal reconstruction from magnitude stft spectrogram based on spectrogram consistency. In: Proc Int Conf Digital audio effects, vol 10
30. Loizou PC (2013) Speech enhancement: Theory and practice. CRC press
31. Lopes D, White P (2000) Signal reconstruction from the magnitude or phase of a generalised wavelet transform. In: 2000 10Th european signal processing conference, IEEE, pp 1–4
32. Magron P, Virtanen T (2020) Online spectrogram inversion for low-latency audio source separation. IEEE Sign Process Lett 27:306–310
33. Malek RA, Mnasri Z, Benzarti F (2018) Optimal conditions for signal reconstruction based on stft magnitude spectrum. In: 2018 15Th international multi-conference on systems, signals & devices (SSD), IEEE, pp 1084–1090
34. Mayer F, Mowlaee P (2015) Improved phase reconstruction in single-channel speech separation. In: Sixteenth annual conference of the international speech communication association
35. Mayer F, Williamson DS, Mowlaee P, Wang D (2017) Impact of phase estimation on single-channel speech separation based on time-frequency masking. J Acoust Soc Am 141(6):4668–4679
36. Moravec ML, Romberg JK, Baraniuk RG (2007) Compressive phase retrieval. In: Wavelets XII, vol 6701, International Society for Optics and Photonics, pp 670120
37. Mowlaee P, Kulmer J (2015) Harmonic phase estimation in single-channel speech enhancement using phase decomposition and snr information. IEEE/ACM Trans Aud Speech Lang Process 23(9):1521–1532
38. Mowlaee P, Saeidi R, Stylianou Y (2014) Phase importance in speech processing applications. In: Fifteenth annual conference of the international speech communication association
39. Mowlaee P, Stahl J, Kulmer J (2017) Iterative joint map single-channel speech enhancement given non-uniform phase prior. Speech Comm 86:85–96
40. Nakamura T, Kameoka H (2014) Fast signal reconstruction from magnitude spectrogram of continuous wavelet transform based on spectrogram consistency. In: DAFX, pp 129–135
41. Ohlsson H, Yang A, Dong R, Sastry S (2012) Cprl–an extension of compressive sensing to the phase retrieval problem. In: Advances in neural information processing systems, pp 1367–1375
42. Pirker G, Wohlmayr M, Petrik S, Pernkopf F (2011) A pitch tracking corpus with evaluation on multipitch tracking scenario. In: Twelfth annual conference of the international speech communication association
43. Pobloth H, Kleijn WB (1999) On phase perception in speech. In: 1999 IEEE International conference on acoustics, speech, and signal processing. Proceedings. ICASSP99 (cat. no. 99CH36258), vol 1, IEEE, pp 29–32
44. Portnoff M (1976) Implementation of the digital phase vocoder using the fast fourier transform. IEEE Trans Acoust Speech Sign Process 24(3):243–248
45. Portnoff M (1979) Magnitude-phase relationships for short-time fourier transforms based on gaussian analysis windows. In: ICASSP'79. IEEE International conference on acoustics, speech, and signal processing, vol 4, IEEE, pp 186–189
46. Pruša Z (2017) The phase retrieval toolbox. In: AES International conference on semantic audio, Erlangen, Germany
47. Pruša Z, Søndergaard PL (2016) Real-time spectrogram inversion using phase gradient heap integration. In: Proc Int Conf Digital audio effects (DAFx-16), pp 17–21
48. Sanchez J, Saratxaga I, Hernaez I, Navas E, Erro D, Raitio T (2015) Toward a universal synthetic speech spoofing detection using phase information. IEEE Trans Info Forensics Secur 10(4):810–820
49. Saratxaga I, Erro D, Hernáez I, Sainz I, Navas E (2009) Use of harmonic phase information for polarity detection in speech signals. In: Tenth annual conference of the international speech communication association
50. Shimauchi S, Kudo S, Koizumi Y, Furuya K (2017) On relationships between amplitude and phase of short-time fourier transform. In: 2017 IEEE International conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 676–680
51. Smaragdis P, Raj B, Shashanka M (2011) Missing data imputation for time-frequency representations of audio signals. J Signal Process Syst 65(3):361–370

52. Takaki S, Kameoka H, Yamagishi J (2017) Direct modeling of frequency spectra and waveform generation based on phase recovery for dnn-based speech synthesis. In: INTERSPEECH, pp 1128–1132
53. Tech E (2008) 3253: Sound quality assessment material recordings for subjective tests. EBU Geneva
54. Thorpe L, Shelton B (1993) Subjective test methodology: Mos vs. dmos in evaluation of speech coding algorithms. In: Proceedings., IEEE workshop on speech coding for telecommunications, IEEE, pp 73–74
55. Van Hove P, Hayes M, Lim J, Oppenheim A (1983) Signal reconstruction from signed fourier transform magnitude. Trans Acous Speech Sign Process 31(5):1286–1293
56. Virtanen T (2007) Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. Trans Aud Speech Lang Process 15(3):1066–1074
57. Voiers WD (1976) Methods of predicting user acceptance of voice communication systems. Tech. rep. DYNASTAT INC AUSTIN TX
58. Waldspurger I (2017) Phase retrieval for wavelet transforms. IEEE Trans Inf Theory 63(5):2993–3009
59. Wang D, Lim J (1982) The unimportance of phase in speech enhancement. Trans Acous Speech Sign Process 30(4):679–681
60. Yang W (1999) Enhanced modified bark spectral distortion (EMBSD): An objective speech quality measure based on audible distortion and cognitive model temple university
61. Yegnanarayana B, Saikia D, Krishnan T (1984) Significance of group delay functions in signal reconstruction from spectral magnitude or phase. Trans Acous Speech Sign Process 32(3):610–623
62. Zhu X, Beauregard GT, Wyse L (2006) Real-time iterative spectrum inversion with look-ahead. In: 2006 IEEE International conference on multimedia and expo, IEEE, pp 229–232