

Time-Frequency Phase Retrieval for Audio — The Effect of Transform Parameters

Andrés Marafioti, Nicki Holighaus, and Piotr Majdak

Abstract—In audio processing applications, phase retrieval (PR) is often performed from the magnitude of short-time Fourier transform (STFT) coefficients. Although PR performance has been observed to depend on the considered STFT parameters and audio data, the extent of this dependence has not been systematically evaluated yet. To address this, we studied the performance of three PR algorithms for various types of audio content and various STFT parameters such as redundancy, time-frequency ratio, and the type of window. The quality of PR was studied in terms of objective difference grade and signal-to-noise ratio of the STFT magnitude, to provide auditory- and signal-based quality assessments. Our results show that PR quality improved with increasing redundancy, with a strong relevance of the time-frequency ratio. The effect of the audio content was smaller but still observable. The effect of the window was only significant for one of the PR algorithms. Interestingly, for a good PR quality, each of the three algorithms required a different set of parameters, demonstrating the relevance of individual parameter sets for a fair comparison across PR algorithms. Based on these results, we developed guidelines for optimizing STFT parameters for a given application.

I. INTRODUCTION

Phase is a crucial component of audio signals and affects how humans perceive sounds [1] and speech [2], [3]. When processing audio, a signal is often represented in the complex-valued short-time Fourier transform (STFT) domain [4]–[6], although many audio applications focus on processing the STFT magnitude [7]–[10]. In order to synthesize the targeted time-domain signal, they estimate the STFT phase from the processed STFT magnitudes by performing phase retrieval (PR) [11], [12]. The necessity of PR also arises in the generation of a signal described only by STFT magnitude [13], [14]. PR algorithms have been used successfully in the field of audio [15]–[17], including specific applications such as audio inpainting [18], [19], but many applications exist beyond the audio domain, e.g., in X-ray crystallography [20], [21] and imaging [22], [23].

The input to PR algorithms is usually given by phaseless transform coefficients with respect to some dictionary. The classic problem of Fourier-based PR [24] has been extended to deal with various time-frequency (TF) representations, most

notably the STFT, the best understood and most widely used TF representation in the field of audio processing. However, different STFT parameters may yield different PR results. From the mathematical perspective, PR is a difficult inverse problem and various conditions ensuring its feasibility have been derived, e.g., [25]–[31], yielding conditions on the window, transform parameters, or limitations in the processed input. These conditions, while theoretically correct, can be impractical in actual applications. To find practical conditions, we evaluated the performance of various PR algorithms under systematic variation of STFT parameters and audio type.

Phase¹ retrieval for audio signals reached its first milestone in 1984, when the Griffin-Lim algorithm (GLA) was introduced [32]. It is iterative and computationally intensive in each iteration and unsuitable for real-time applications. Further improvements with respect to quality [33]–[35] and computation time [36], [37] yielded algorithms such as the fast Griffin-Lim algorithm (FGLA) [38] and real-time iterative spectrogram inversion [39]. Nowadays, GLA is a widely used iterative PR algorithm, e.g., [40], [41].

Alternative, non-iterative algorithms such as single-pass spectrogram inversion (SPSI) [42], phase unwrapping [43], and phase gradient heap-integration (PGHI) [44], have been proposed. SPSI assumes a sinusoidal model with linear phase progression and phase locking [45] to the closest spectral peak. It is fast and directly suitable for real-time usage, but it relies on the assumption that the signal consists of slowly varying sinusoidal components. PGHI is efficient and equally suitable for real-time processing. Even though PGHI does not rely on any signal assumptions, it is based on the phase-magnitude relations [46], a property of the continuous STFT, which when used in a discrete realm may introduce inaccuracies governed by the parameters of the discrete STFT [47].

All these algorithms are applied to magnitude STFT coefficients which can be obtained with different parameters. Unfortunately, besides general introductions to phase-aware processing [15], [48], there are only a few hints towards which are ‘good’ STFT parameters for PR algorithms. For example, a large number of frequency channels seems to be beneficial in music applications [14]. For a fixed number of frequency channels, the window function and redundancy seems to affect the PR quality [49], with larger redundancy improving the PR quality [50], [51]. Still, a systematic investigation of the transform parameters affecting the PR quality of audio seems to be missing.

¹From now on, we use *phase* when referring to the STFT phase.

Manuscript received on November 2020; Revised in April 2021.

The authors are with the Acoustics Research Institute, Austrian Academy of Sciences, Wohllebengasse 12–14, 1040 Vienna, Austria.

Accompanying web page: <https://github.com/andimarafioti/phaseRetrievalEvaluation>.

We thank Nathanaël Perraudin for the fruitful discussions we had over the years about the importance of the quality of phase retrieval in more complex systems. This work has been supported by Austrian Science Fund (FWF) project MERLIN (Modern methods for the restoration of lost information in digital signals; I 3067-N30).

In this article, we first revisit relevant properties of the discrete STFT [52]–[54] in the context of PR. We then systematically evaluate three PR algorithms: PGHI, FGLA, and SPSI. We consider five redundancies of the STFT between 2 and 32, many window sizes ranging from 32 to 61440 samples, and four types of windows: Gaussian, Blackman, Hann, and Bartlett. In addition, we evaluate the PR performance in a simple setting with processed magnitude spectrograms. We also consider the performance of PR for various types of audio signals. Finally, we describe guidelines for obtaining good STFT parameters for a given PR algorithm. The code used to perform our experiments is available at <https://github.com/andimarafioti/phaseRetrievalEvaluation>.

II. THE DISCRETE SHORT-TIME FOURIER TRANSFORM

We consider finite signals $s \in \mathbb{C}^L$ and indices in the signal domain are to be understood modulo L . The STFT of s , with the *analysis window* $g \in \mathbb{R}^L$, time step $a \in \mathbb{N}$ and $M \in \mathbb{N}$ frequency channels is given by

$$\begin{aligned} S_g(s)[m, n] &= \sum_{l=0}^{L-1} s[l]g[l - na]e^{-2\pi iml/M} \\ &= |S_g(s)[m, n]| e^{i\phi_g(s)[m, n]}, \end{aligned} \quad (1)$$

for $n \in \{0, \dots, L/a - 1\}$ and $m \in \{0, \dots, M - 1\}$. If s and g are real-valued, the STFT is conjugate symmetric in m and it is sufficient to store the first $M_{\mathbb{R}} = \lfloor (M/2) + 1 \rfloor$ channels. Note that $\phi_g(s)$ refers to the TF phase, which we refer to simply by *phase* throughout this document. Accordingly, TF PR is concerned with estimating the phase, or equivalently $S_g(s)$, from the *magnitude* $|S_g(s)|$.

A. Properties of the STFT

Depending on the choice of transform parameters a , M , and the window g , the discrete STFT encodes time and frequency information with different properties. The *full* STFT, i.e., with $a = 1$ and $M = L$, is a slowly varying function, owing to significant overlap between both the time range covered by adjacent time positions and the frequency range covered by adjacent frequency channels. When increasing the time step a over 1, the *time resolution* of the STFT decreases. Similarly, when decreasing the number of channels M below L , the *frequency resolution* of the STFT decreases. Jointly, time and frequency resolution can be likened to the pixel resolution in digital imaging. This joint resolution is characterized by the *redundancy* (D) of the STFT, $D = M/a$. Figure 1 shows examples of STFT magnitudes calculated with the same window g , for various redundancies D . Especially at redundancy $D = 2$, it can be seen that some characteristic features of the STFT magnitude are obscured.

Further, the window g and its Fourier transform \widehat{g} control the inherent TF uncertainty [54] of the STFT, independently of a and M . Namely, every window function g has a certain shape in time and \widehat{g} in frequency, which determine how spectro-temporal signal components are *smear*ed in the STFT magnitude. The shape of a window is usually characterized by its width. In the

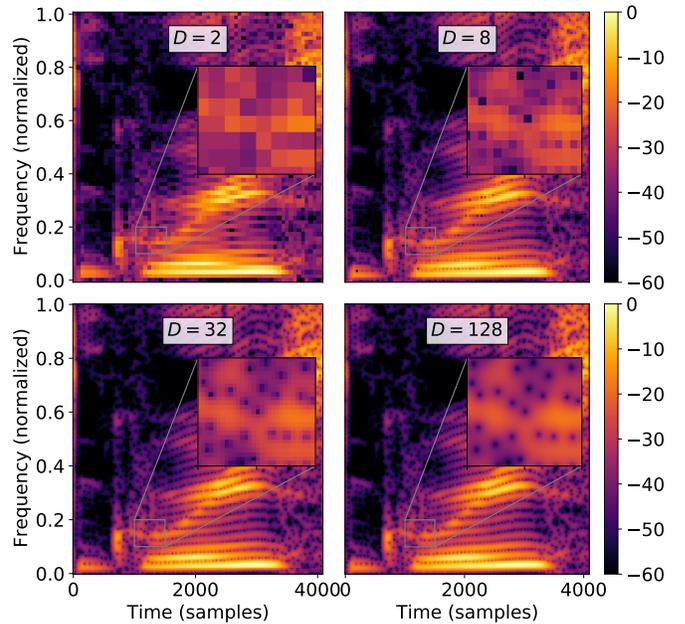


Fig. 1: Exemplary spectrograms calculated for various redundancies D . Calculations were done with the Gaussian window and TF ratio of $\lambda = 8$.

classic uncertainty principle, a window’s time and frequency width are defined as the standard deviation of g and of \widehat{g} , respectively. This notion is reasonable for any smooth, roughly bell-shaped window.

The classic example of a bell-shaped window is the Gaussian window. The Gaussian window minimizes the product of time and frequency width and its Fourier transform is a Gaussian as well. The *discrete, periodized* Gaussian, simply referred to as *Gaussian* in this study, is defined as:

$$g_\lambda[l] := \sum_{k=-\infty}^{\infty} e^{-\frac{\pi(l-kL)^2}{\xi_s \lambda}}, \quad \lambda, \xi_s \in \mathbb{R}^+, \quad (2)$$

where ξ_s is the assumed sampling rate (in Hz). It inherits from its continuous counterpart the property that its DFT \widehat{g}_λ is again a Gaussian. The parameter λ defines jointly the width of g_λ and \widehat{g}_λ , as illustrated in Figure 2 by the inverse relation between the width of g_λ (measured in samples) is λ times as large as the width of \widehat{g}_λ (measured in Hz). This is why λ can be referred to as the *TF ratio* of a Gaussian window. The effect of λ on the STFT is shown in Figure 3, for different STFTs at the same redundancy D .

The Gaussian window is available in public libraries such as Scipy [55] and LTFAT [56]. However, it is not the most commonly used window function. Therefore, libraries specializing in a particular field often do not implement it, e.g., PyTorch [57] for machine learning. Instead, it is more common to compute the STFT using windows with short support, such as the Hann, Hamming or rectangular windows. For those windows g there is no exact equivalent to the TF ratio λ . Instead, one can determine their equivalent λ through comparison to the

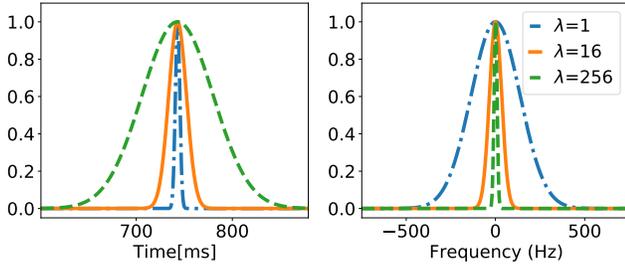


Fig. 2: Examples of Gaussian windows in the time domain (left) and magnitude of their Fourier transform (right) for various TF ratios λ . The length of the windows was the same, i.e. $L = 65536$ samples, in all examples.

Gaussian window. Precisely, given a window g , we find λ as $\operatorname{argmin}_{\lambda} \|g - g_{\lambda}\|$, assuming that g is peak-normalized, i.e., $\max |g[l]| = 1$. Alternatively, g can be fit to a given λ by adjusting the length of g to minimize the norm distance to g_{λ} .

In conclusion, the overall numerical properties of the discrete STFT depend on the joint choice of the parameters g_{λ} , governing its uncertainty, and a and M , controlling its resolution. The most favorable properties can be achieved when the uncertainty is matched to the resolution, [58], [59]. With the definitions in Eqs. (1) and (2), uncertainty and resolution are matched if and only if

$$\lambda = aM/\xi_s. \quad (3)$$

In this case, the STFT samples lie on a grid on which the ratio between time- and frequency-steps coincides with λ , leading to an optimally uniform covering of time-frequency space. In all our experiments, λ , a and M are linked in this fashion.

B. Inverse STFT for signal synthesis

For any *synthesis window* $\tilde{g} \in \mathbb{R}^L$, the inverse STFT of $S \in \mathbb{C}^{M \times N}$ with respect \tilde{g} is given by

$$\tilde{s}[l] = \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} S[m, n] \tilde{g}[l - na] e^{2\pi i m l / M}, \quad (4)$$

for $l \in \{0, \dots, L-1\}$. If a \tilde{g} exists, such that $\tilde{s} = s$ for all $s \in \mathbb{C}^L$, and with $S = S_g(s)$, then the STFT S_g is invertible, i.e., it forms a frame in the sense of [6], [60], [61] and \tilde{g} is a *dual window* for g . Generally, in order to obtain an invertible STFT, a redundancy equal to or larger than one, $D = M/a \geq 1$ is required².

For redundancies $D > 1$, the STFT is overcomplete (or redundant), and S_g maps into a strict subspace of $\mathbb{C}^{M \times N}$. In other words, not every matrix $S \in \mathbb{C}^{M \times N}$ represents an STFT. We call S *consistent* if there is a signal s , such that $S = S_g(s)$ and *inconsistent* otherwise. *Implicitly*, the inverse STFT operation applied to S projects onto the image of S_g before synthesis, as visualized in Fig. 4. In practice, this means that the inverse STFT, applied to inconsistent coefficients S

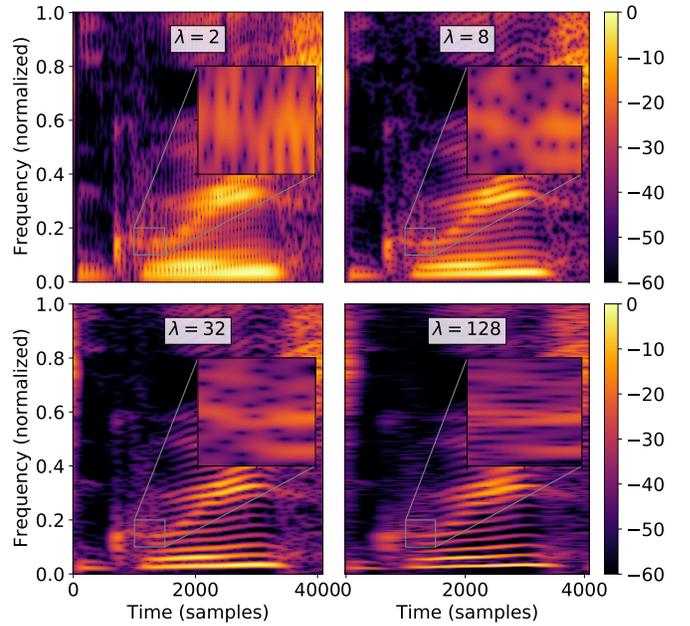


Fig. 3: Exemplary spectrograms calculated for various TF ratios λ of the Gaussian window. The same redundancy of $D = 128$ was used in all examples.

produces a signal \tilde{s} with $S_g(\tilde{s}) \neq S$. Applied to consistent coefficients S , we instead obtain $S = S_g(S_g^{-1}(S))$. Thus, in the setting of PR, synthesis from a given spectrogram $|S_g(s)|$ with a mismatched phase estimate ϕ will often lead to a poor reconstruction. In some cases, we may also use (*in*)consistent to describe magnitude coefficients $M \in (\mathbb{R}_0^+)^{M \times N}$ that do not correspond to the STFT of *any* signal, i.e., $M \neq |S_g(s)|$ for all $s \in \mathbb{C}^L$.

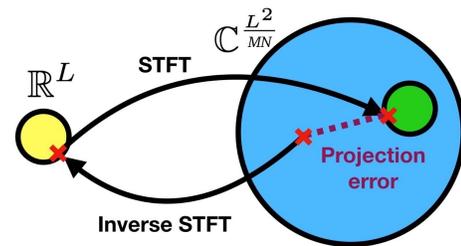


Fig. 4: Blue circle: Set of all possible TF coefficients. Yellow circle: set of time-domain signals. Green circle: set of consistent STFT coefficients. An inverse STFT done on a point from the blue set yields a point from the yellow set. An STFT from this point yields a point from the green circle, introducing a projection error. An inverse STFT done on a point from the green circle yields a point from the yellow set, which after a subsequent STFT is remapped to the original point in the green set without any projection errors.

²In contrast to common practice, the number of channels M may be smaller than the number of nonzero samples in g .

III. GENERAL METHODS

A. Datasets

The phase of different types of audio signals may have completely different characteristics, thus, we considered three types of signals for the evaluation attempting to cover a wide range of audio characteristics. First, we included speech signals to the evaluation, because it is a widely used class of signals, it consists not only of harmonic components, but also transients and stochastic segments such as fricatives. Second, we considered piano music synthesized from MIDI because it represents a class of simple polyphonic sounds resembling a linear combination of sine waves, without any ambient recording noise. Third, we considered actual music recordings, which included the natural variations from the musicians and ambient noise from the recording setup.

1) Speech. For speech, we used LJ Speech [62], which is a public-domain English speech dataset consisting of 13,100 short audio clips of a single speaker reading passages from seven non-fiction books. All clips have a sampling rate of 22050 Hz and vary in length from 1 to 10 seconds and have a total length of approximately 24 hours.

2) Midi synthesized music. We used the Lakh MIDI dataset [63], a collection of 176,581 unique simple piano MIDI files, to synthesize piano audio signals. This MIDI set was created with the goal of facilitating large-scale music information retrieval, both symbolic (using the MIDI files alone) and audio content-based (using information extracted from the MIDI files as annotations for the matched audio files). The audio files were synthesized from MIDI data using `pretty_midi` [64], specifically its `fluidsynth` API. We generated just one instrument with a sampling rate of 22050 Hz and set it to the piano program 1.

3) Music. We segmented the ‘small’ dataset of the free music archive (FMA, [65]) by genre and used the genre ‘electronic’. This was done to reduce the variability in the music structure in our evaluations. FMA is an open and easily accessible dataset, usually used for evaluating tasks in musical information retrieval. The subset we used is comprised 1,000 30-s segments of songs sampled at 44.1 kHz.

The audio material was processed at a sampling rate of 22050 Hz after resampling to that rate if necessary. Our experiments use the Large Time-Frequency Analysis Toolbox (LTFAT, [56]). Hence, since LTFAT’s STFT requires a, M to be divisors of the length of the signal L , we chose the first $L = 122880$ samples of a randomly selected 128 signals for each subset, resulting in approximately 5.6 seconds per signal. For the experiments with varying number of channels M , after reconstruction a portion of signal of length M was removed at the beginning and the end of the signals to avoid issues introduced by the circularity of the considered STFT implementation.

B. PR algorithms

We evaluated three PR algorithms implemented in the Phase Retrieval Toolbox (PHASERET, [66]): Phase-gradient heap integration (PGHI) [44], fast Griffin-Lim (FGLA) [38], and

single-pass spectrogram inversion (SPSI) [42]. These algorithms were chosen to cover a wide range of PR strategies and based on the availability of tested and reliable implementations in the toolbox, enabling a fair comparison. PHASERET relies on LTFAT for STFT computation and other basic functionalities.

1) PGHI is a non-iterative method. PGHI implies no assumptions on the signal. Instead, it is based on the phase-magnitude relations of an STFT computed using a Gaussian window [46], namely, the relation between the partial phase derivatives of the continuous STFT with a Gaussian window and the partial derivatives of the logarithmic STFT magnitudes [44], [52]. In PGHI, this relation is approximated for the discrete STFT as:

$$\begin{aligned} \partial_n \phi_g[m, n] &\approx \frac{aM}{\lambda} \partial_m \log(|S_g|)[m, n], \\ \partial_m \phi_g[m, n] &\approx -\frac{\lambda}{aM} \partial_n \log(|S_g|)[m, n] - 2\pi na/M. \end{aligned} \quad (5)$$

Here, ∂_n, ∂_m denote numerical differentiation with respect to n and m , respectively. This step may be realized, e.g., by a finite difference scheme. The dependence on properties of the Gaussian STFT suggests a dependence on the window function, which was already observed in [49]. From the phase-magnitude relation, the phase (ϕ_g) is reconstructed in an adaptive integration scheme. The reliance on numerical differentiation and integration suggests that the results of PGHI also depend on the STFT parameters a, M, λ and, in particular, its redundancy D . In our experiments, PGHI was initialized with no prior knowledge of the original phase.

2) FGLA is an iterative algorithm relying on alternating projections and it is based on the Griffin-Lim algorithm (GLA) [32], which is itself an extension of the seminal Gerchberg-Saxton algorithm [67]. Specifically, given a target STFT magnitude combined with an initial phase estimate (in our experiments, the phase was uniformly set to zero), the algorithm performs first a projection onto the space of consistent STFTs. Since the latter is a strict subspace of $\mathbb{C}^{M \times N}$ whenever the STFT is redundant, this step is expected to yield a magnitude different from the target. Therefore, the second step keeps only the new phase, an imposes the target magnitude. Both steps of GLA are repeated until convergence or until a certain number of iterations. A final inverse STFT is then applied to synthesize a time-domain signal. Thus, GLA does not rely on a signal model, but only on the redundancy of the STFT. In our experiments we employ fast Griffin-Lim (FGLA) [38], which adds an acceleration term, governed by a hyperparameter α to the GLA update, and empirically yields better results at the same number of iterations, often significantly. For our experiment we use the default value $\alpha = 0.99$ proposed by the reference implementation in LTFAT [56]. Furthermore, we set the number of iterations to 100, which provides a decent trade-off between quality and computation time. The convergence curve after 100 iterations is already rather flat such that a large number of iterations is required to achieve significant improvements.

3) SPSI is another non-iterative method. In contrast to PGHI, SPSI does not rely on mathematical properties of the STFT. And in contrast to FGLA, it is fast and directly suitable for real-time usage. SPSI implicitly assumes a sinusoidal signal model and thus fails for transient and broadband components in the signal. At every time step, SPSI locates peaks in the TF coefficients obtained and predicts the phase by assuming a linear phase progression at the rate of the closest peak frequency. Hence, the rate at which the TF coefficients vary over time is expected to be a limiting factor of PR by SPSI. The phase prediction, being similar to the integration scheme of PGHI, depends on the time step parameter a and, to a lesser degree, on the number of frequency channels M . In our experiments, SPSI was initialized without any information regarding the original phase.

C. Evaluation measures

The results were evaluated numerically by means of several measures. First, we considered the signal-to-noise ratio calculated on the magnitude spectrogram, (SNR_{MS}). This quantity is also sometimes referred to as *spectral convergence*. Second, to consider human-like performance, we computed the objective difference grade (**ODG**) based on perceptually motivated models.

1) Spectrogram signal-to-noise ratio (SNR_{MS}) is the logarithmic ratio between the energy of the spectrogram $|S|$ of the original signal s and the energy of the spectrogram difference ($|S_r| - |S|$), where S_r is the STFT of the reconstructed time-domain signal s_r :

$$\text{SNR}_{\text{MS}}(S, S_r) = 10 \log_{10} \frac{\|S\|^2}{\||S_r| - |S|\|^2}. \quad (6)$$

To compute SNR_{MS} , we used the STFT as in Eq. (1) with $M = 2048$, $a = 128$ (thus, $D_{\text{SNR}} = 16$), and the Gaussian window g_λ with $\lambda_{\text{SNR}} = aM/\xi_s \approx 11.886$. In Section IV-A, we show that SNR_{MS} only exhibits minor dependence on this parameter choice.

2) Objective difference grade (**ODG**) is the overall quality measure introduced in PEAQ [68], [69] and designed to mimic perceptual quality ratings made by a human listener. PEAQ is a full-reference algorithm, i.e., it performs a direct comparison between a modified signal and a target signal³. It relies on an auditory model obtained by processing STFT coefficients and ranges from 0 to -4 with the interpretation shown in Tab. I. We used the implementation from [70].

Internally, **ODG** computes the STFT of the analysed signal. Thus, it might prefer a particular set of STFT parameters. To consider this effect in our evaluation, we initially calculated ODGs based on PEMO-Q [71] as well, which uses a Gammatone filterbank. Due to this difference in analysis dictionary, it is unlikely that PEMO-Q exhibits preference for certain STFT parameters. We used the implementation from [72] and refer to this measure as ODG_{PEMO} .

³In our case, the original signal.

ODG	Impairment
0	Imperceptible
-1	Perceptible, but not annoying
-2	Slightly annoying
-3	Annoying
-4	Very annoying

TABLE I: Interpretation of **ODG**.

IV. EXPERIMENTS

A. Sensitivity of the evaluation measures

For PR performance, we expected a significant effect of the TF ratio and redundancy across the tested PR algorithms. In order to distinguish between actual effects of the PR algorithm and effects induced by the evaluation measure, we first determined the sensitivity of the evaluation measures **ODG**, ODG_{PEMO} , and SNR_{MS} to changes of these parameters.

SNR_{MS} reflects the average amount of phase distortion over signal duration, thus, we hypothesized that it is largely insensitive to changes in the TF ratio λ . On the other hand, adjacent TF coefficients are correlated (with the correlation increasing with D) and any uncorrelated distortion imposed on the coefficients partially cancels in the synthesis process. Thus, we expected the PR quality to improve with increasing redundancy, and SNR_{MS} to reflect this. However, SNR_{MS} itself is based on an STFT, the parametrization of which, determined by λ_{SNR} and D_{SNR} , might affect the results. To account for this effect, for each evaluated condition, we calculated SNR_{MS} for all combinations of $\lambda_{\text{SNR}} \in \{10^{-3}, 10^4\}$ and $D_{\text{SNR}} \in \{2, 8, 32\}$ and evaluated the statistics by means of the average and standard deviation.

The manifestation of phase distortion in synthesized time-domain signals \tilde{s} depends on the width of the synthesis window \tilde{g} , which further depends on λ . Therefore, we expected an effect of the TF ratio on **ODG**. To account for possible presence effects caused by the STFT parametrization in PEAQ, we additionally calculated ODG_{PEMO} for the same set of conditions.

For the evaluation, we first computed STFTs for various TF ratios $\lambda \in \{10^{-3}, 10^4\}$ and three redundancies $D \in \{2, 8, 32\}$. Note that a λ, D combination uniquely determines both a and M . Then, we added Gaussian white noise to the phase of these STFTs. We tested three standard deviations $\sigma \in \{1, 0.5, 0.1\}$ and phase values were wrapped onto the range $\pm\pi$ after applying the distortion. Further note that the condition with the highest distortion level corresponds to reconstructing spectrograms with a nearly random phase. Finally, we calculated the inverse STFT and applied the three measures to the result. This setup provides direct evidence for the extent of sensitivity of **ODG**, ODG_{PEMO} , and SNR_{MS} to the TF ratio λ and redundancy D .

Fig. 5 shows the results. For the SNR_{MS} , the statistics across λ_{SNR} and D_{SNR} show negligible standard deviation as compared to the effect of changes of λ and D . This indicates that the parametrization of SNR_{MS} had nearly no effect on the sensitivity and reliability of this measure when evaluating

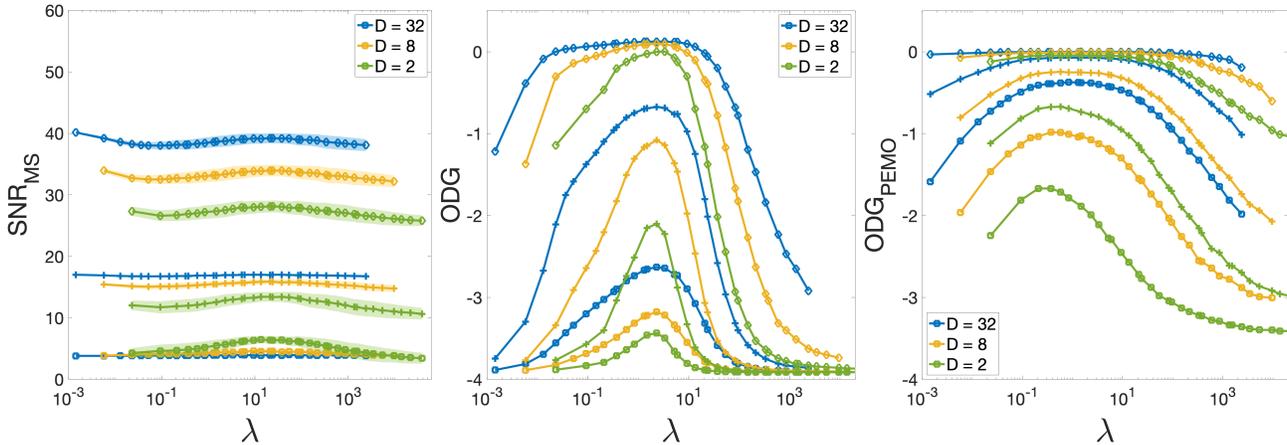


Fig. 5: SNR_{MS} , ODG , and ODG_{PEMO} resulting from the inverse STFT of speech signals with distorted phases. Symbols: severity of distortion with $\sigma = 0.1$ (diamonds), $\sigma = 0.5$ (plus), $\sigma = 1.0$ (squares).

phase effects for various λ and D . The average SNR_{MS} correspond roughly to the values calculated for $D_{\text{SNR}} = 16$ and $\lambda_{\text{SNR}} = 10$. Thus, in the following, we use parameters as indicated in Sec. III-C to calculate SNR_{MS} in the following experiments.

For the highest distortion level ($\sigma = 1$), the SNR_{MS} was below 8 dB and ODG was worse than ‘annoying’ in most cases, indicating that this amount of distortion substantially destroyed the original signals in every considered combination of λ and D . These results reflect the output one would expect of our measures when reconstructing spectrograms with a random phase. On the other side, ODG_{PEMO} was between ‘not annoying’ and ‘annoying’ for the majority of conditions, implying that even for nearly random phases, the results would have been acceptable when relying on this measure alone. This indicates that ODG_{PEMO} is most probably not suitable as a measure to analyze the effects of phase distortion.

For the moderate ($\sigma = 0.5$) and low ($\sigma = 0.1$) distortion levels, a pattern emerges: at fixed TF width ratio, larger D yielded better performance in terms of larger SNR_{MS} and better ODG . With most results being better than ‘not annoying’, ODG_{PEMO} showed little sensitivity to these amounts of noise, except at extreme λ and low redundancy D . We take this as further evidence to the poor sensitivity of this measure to phase distortion. SNR_{MS} showed little effect of TF ratio, with a small peak at λ of approximately 10. In contrast, ODG seems to be more sensitive to the TF ratio, following a bell shape with a clear peak at the same single-digit λ for all D s. This peak seems to be wide for low levels of phase distortions and to become sharper for increasing distortion level.

In summary, we observed the following: 1) the small differences, when computing SNR_{MS} parametrized to various λ_{SNR} and D_{SNR} combinations, indicate that the particular choice of its own parametrization is not essential. 2) SNR_{MS} , calculated with the default parametrization, is sensitive to phase distortions, but for a fixed amount of distortion, it is only mildly sensitive to the TF ratio. Moreover, at low to moderate

distortion, SNR_{MS} increases with D . SNR_{MS} below 10 dB was in line with what we expect for reconstructions with a random phase and can be used as a rule of thumb when evaluating PR algorithms. 2) ODG is sensitive to both TF ratio and redundancy, however, it showed ceiling effects, i.e., it saturated at low levels of distortions and high redundancies. The sensitivity to the TF ratio seems to depend on the level of distortion, with single-digit λ being a good choice at most redundancies. 3) ODG_{PEMO} was barely sensitive to phase distortions and did not use the full range even for nearly random phases. This provides strong indication that it is not a suitable measure for evaluating phase effects. 4) All measures showed consistently better results with increasing redundancy, demonstrating the increased robustness of the inverse STFT to phase distortions with increasing redundancy. 5) By combining SNR_{MS} and ODG we can avoid our evaluation to be hampered by the low sensitivity of SNR_{MS} to λ and the ceiling effects of ODG for high D , obtaining an adequate scheme for evaluating PR algorithms in the following experiments.

B. Effect of STFT parameters on PR

In this experiment, we studied the effect of the choice of STFT parameters on PR in terms of SNR_{MS} and ODG . Evaluation was performed for Gaussian windows, while varying the redundancy D , and TF ratio λ . The experiment aims to not only assess the effect of D and λ , but also to demonstrate performance differences between the algorithms.

To achieve this, we created spectrograms of the speech dataset considering redundancies $D \in \{2, 4, 8, 16, 32\}$ and a large range of TF ratios $\lambda \in [10^{-3}, 10^4]$, applied all considered PR algorithms (Sec. III-B) on those spectrograms, and calculated SNR_{MS} and ODG of the reconstructed signals. The results are presented in Fig. 6.

For all three algorithms, λ had a clear effect on both SNR_{MS} and ODG . We can clearly see that the change of ODG in λ is more pronounced than in Exp. A, indicating actual impact of λ on PR quality. In contrast to Exp. A,

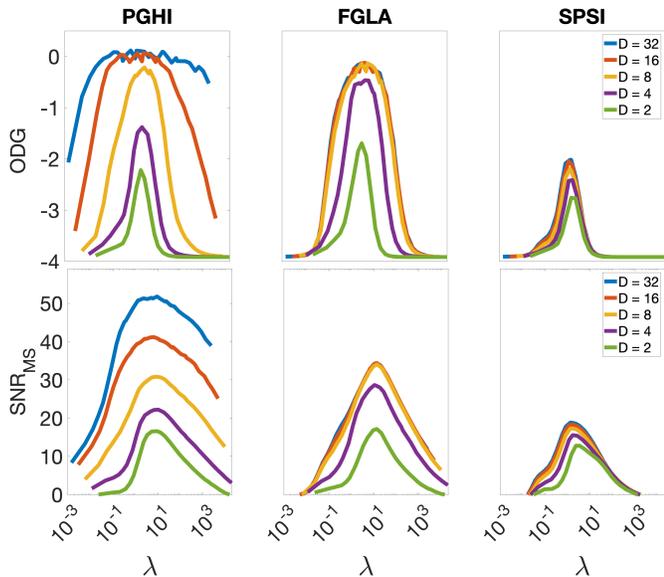


Fig. 6: PR performance in terms of ODG and SNR_{MS} obtained with three PR algorithms: PGHI, SPSI, and FGLA. Calculations done with the Gaussian window, and various redundancies and TF ratios.

SNR_{MS} also shows significant variation in λ , providing further evidence that this is the case. For λ below 0.1 and above 100, the level of phase distortions introduced by the PR corresponded to that of high level of noise ($\sigma = 1.0$). Although the general trend is similar across all algorithms, the extent to which STFT parameters affect PR quality depended on the chosen algorithm.

For PGHI, both measures showed peaks at λ , and those peaks were shared for every D . For PGHI peaks were less pronounced than for FGLA, indicating that PGHI is less sensitive to a particular choice of λ . For ODG , the peak was at $\lambda = 2.32$, corresponding to $M = 320$ for $D = 2$. For SNR_{MS} , the peak was at $\lambda = 5.94$, corresponding to $M = 512$ for $D = 2$. The performance increased with the redundancy, showing ceiling effects in ODG for redundancy of 16 or larger. For redundancies of $D \geq 16$, SNR_{MS} showed little distortions, comparable to our low distortion level with Gaussian white noise. From these results, we conclude that for speech signals PGHI works best at $D \geq 16$ with $\lambda \approx 3$. For lower redundancies, the performance degraded and the choice of λ became even more important.

For FGLA, both measures followed a thin bell shape with peaks at λ . This peak was the same at every redundancy D . For ODG , the peak was at $\lambda = 3.34$, corresponding to $M = 384$ for $D = 2$. For SNR_{MS} , the peak was at $\lambda = 13.37$, corresponding to $M = 768$ for $D = 2$. For both measures, the performance increased with the redundancy for D of up to 8, showing no improvements beyond that redundancy. This is in contrast with Sec. IV-A, where performance improved with increasing redundancy. We conclude that for FGLA, the choice of λ is crucial, with $\lambda \approx 8$ providing good results. For redundancies $D \leq 4$, FGLA performed better than PGHI, but, there is no gain in increasing D beyond 8.

For SPSI, even the best performance corresponded to large or moderate level of distortion when compared to that obtained for Gaussian white noise from Sec. IV-A. Performance increased slightly with the redundancy, however still remained at a low level. The performance depended on λ , showing a peak $\lambda = 1.4$ for every D , corresponding to $M = 256$ at $D = 2$. The low general performance of SPSI is probably an effect of the underlying signal assumption of slowly varying sinusoidal components, which does not hold for speech signals.

In all three algorithms, the performance increased with the redundancy. This is not surprising because with increasing redundancy, phase and magnitude are more dependent and deducing one from the other becomes easier. This is related to the fact that phase can be perfectly calculated from the magnitude of a continuous STFT (up a fixed scalar factor and disregarding numerical precision) [46] and, by increasing redundancy, the discrete STFT approximates the continuous setting. Algorithms utilizing this principle particularly benefit from increased redundancy. This explains the performance increase with redundancy provided by PGHI (which is based on that principle) and the limited performance gain provided by FGLA and SPSI (which rely on other principles).

The time-frequency ratio λ affected the performance of all tested algorithms. While the redundancy extends the range of reasonable choices for λ , generally good performance can be obtained for λ between 0.2 and 20, whereas performance was mostly poor for λ below 0.1 and above 100. In order to explain why particular time-frequency ratios are beneficial for PR, we need to look into the interaction between the audio signal and the window duration resulting from a particular λ . For example, at this sampling rate, $\lambda = 5$ implies a window duration of approximately 35 ms and provides a good trade-off between temporal and spectral resolution, see Fig. 3. Substantially shorter and longer windows, i.e., resulting from substantially smaller and larger time-frequency ratios, respectively, create more spectral and temporal smearing in the magnitudes, respectively.

C. Effect of STFT parameters on inconsistent spectrograms

In the previous experiment we considered PR from unmodified magnitude spectrograms. This allowed us to investigate the PR task in isolation, without having to consider inconsistency, e.g., introduced in processing. However, in practical applications, PR is mostly applied to modified or synthetic spectrograms, which are rarely consistent. Reconstruction from inconsistent spectrograms introduces errors, recall Fig. 4. When combined with PR, these errors cannot be uniquely attributed to either inconsistency or PR artifacts. Hence, we investigated the PR effect on inconsistent spectrograms in a simple setting: Approximating a time-invariant filter with nonnegative frequency response by weighting spectrogram channels according to the (sampled) frequency response and subsequent PR of the phaseless weighted spectrogram. The target signal is obtained by applying the frequency response directly to the DFT of the full input signal. As a reference, we use reconstruction from the weighted complex STFT, i.e.,

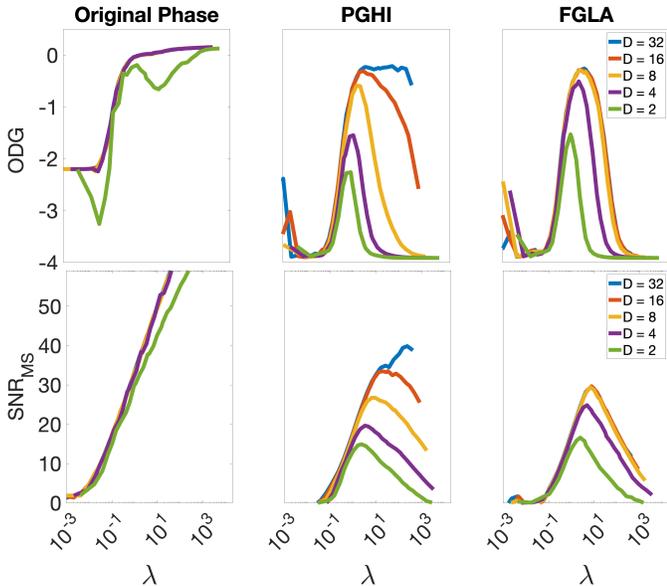


Fig. 7: Effect of the synthesis of inconsistent spectrograms in terms of **ODG** and **SNR_{MS}** obtained with original phases (left columns) and reconstructed phases by PGHI (center columns) and FGLA (right column). Calculations done with the Gaussian window and various redundancies and TF ratios.

we use the original input phase. Given the poor performance of SPSI in Exp. B, we did not consider it for this experiment. The frequency response of the considered filter is

$$\max(\{\min(\{0.1 + \cos(2\pi \cdot 5\xi), 1\}), 0.1\}), \quad (7)$$

with frequency index $\xi \in \{0, \dots, L-1\}$. The filter was chosen only for the purpose of illustration. The filter had the same length as the signal and possessed 15 equidistant peaks where the output's energy was unmodified, and 14 valleys where it was reduced by a factor of 10. The sampled filter resolved all valleys and peaks from the original filter for $M \geq 96$. In this setting, the larger the M , the milder the inconsistencies in the processed STFTs.

The results are shown in Fig. 7. Using the original phase, performance increased monotonically with λ . This corroborates our initial hypothesis that the larger the M , the less inconsistencies are introduced. The performance obtained by the three PR algorithms also increased with λ until a tipping point where their performance decreased similarly to the decrease found in the previous experiment.

We conclude that the inconsistencies in spectrograms interact with the effect of STFT parameters on PR. This is clear when comparing to Exp. B, where the optimal performance was found for a similar λ at every redundancy. In contrast, in this experiment, where the phase is reconstructed from inconsistent spectrograms, the optimal λ increased by a factor of ten when doubling the redundancy.

D. Effect of the signal content

In this experiment, we investigate the relationship between PR performance and the signal content. Based on the explanation for the optimal PR parameters from IV-B, we expected the signal content to affect the optimal PR parameters. To examine this, we performed a reduced version of Exp. B on three classes of audio signals and analyzed their effect on the PR performance. Given that λ determines the TF resolution trade-off and uncertainty, we expected to find optimal λ depending on the signal content.

PR was performed by all three algorithms on each of the three considered datasets. For PGHI, we used three redundancies $D \in \{2, 8, 32\}$, for FGLA two redundancies $D \in \{2, 8\}$, and for SPSI only $D = 8$, considering the reduced significance of redundancy for FGLA and SPSI in previous experiments. Figure 8 shows the results.

The PR performance depended on the signal class, however, it not as strong as the dependency on the STFT parameters. For example, for PGHI and redundancy of $D = 32$, we found an average difference in **SNR_{MS}** of approximately 10 dB between speech and music, with a largest difference of 20 dB for $\lambda = 10$. Such a difference is smaller than that when switching from $D = 32$ to $D = 8$ or varying λ by one order of magnitude. The other PR algorithms showed even smaller effect of the signal class.

Still, the signal class had a general effect on the optimal λ , i.e., compared to speech, the optimal λ increased by factor three on average when switching to MIDI, with the optimal λ for music being between those found for speech and MIDI. This effect can be (again) explained by looking at the interaction between the audio signal and the window duration linked with the time-frequency ratio. Music contains generally lower frequencies, i.e., down to 20 Hz, while speech has the lowest fundamental frequency around 80 Hz. Compared to the MIDI-generated piano sounds, our electronic music had a larger bandwidth. The MIDI-generated piano sounds, on the other hand, contained on average more energy in the low frequency region. Thus, our MIDI required longer window durations, i.e., larger time-frequency ratios, in order to encode the phase information in the magnitude at the same level of accuracy as for speech.

The general conclusion from this experiment is that 1) PR performance depends only a little on the signal class as compared to the dependency on the STFT parameters, and 2) the optimal λ depends on the frequency content to be processed, with lower for speech and higher by up to a factor of three when applied to low frequency signals such as music.

E. Effect of the window function

From the three PR algorithms, only PGHI places explicit assumptions on the STFT parameters, specifically the window being a Gaussian. Therefore, we expect a particular influence of the window function for PGHI. FGLA and SPSI, on the other hand, make little assumptions on the transform, so we expect no large effect of the window function.

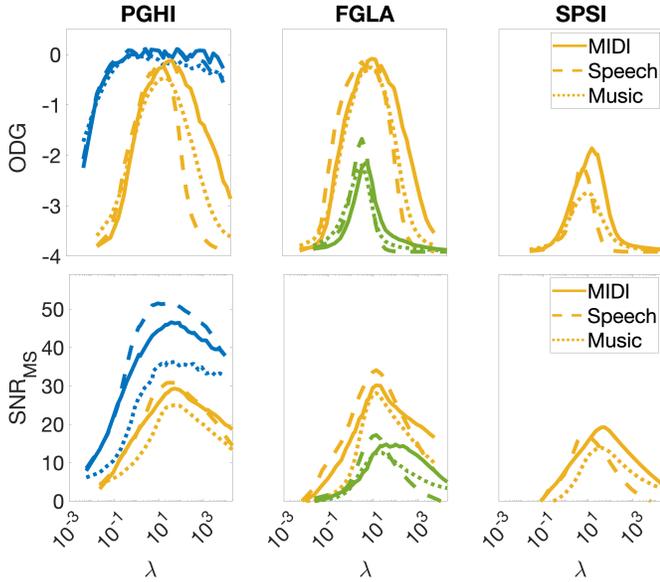


Fig. 8: PR performance as an effect of the signal class. MIDI represented in solid lines, speech in dashed lines, and music in dotted lines. Color indicates redundancy: Blue ($D = 32$), yellow ($D = 8$), and green ($D = 2$). All other aspects are as in Fig. 6.

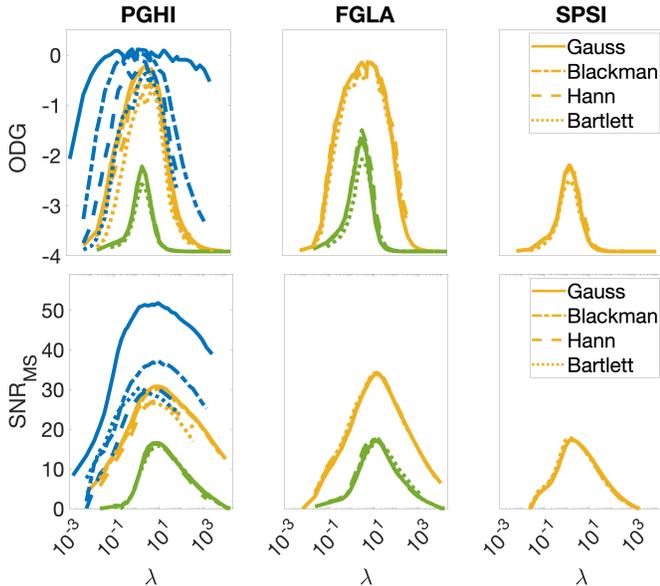


Fig. 9: PR performance as an effect of the window. Gauss window represented in solid lines, Blackman in dashed lines, Hann in dotted lines, and Bartlett in dash-dot line. All other aspects are as in Fig. 8

To verify these hypotheses, this experiment repeats the Exp. B, with the difference that we used either the Gaussian, the Blackman, the Hann, or the Bartlett window in the STFT computations. To match the windows to λ , we determine g closest to the Gaussian g_λ , as discussed in Sec. II-A. Following this, we completed the procedure from Exp. B for a comparable range of TF ratios. We evaluated this experiment for all redundancies, but only show results for the same redundancies as in Exp. D in order for the results to be easier to interpret. The results are presented in Fig. 9.

Only PGHI showed a significant sensitivity to the window. In particular, at a high redundancy such as 32, the difference between every window is significant, even larger than the difference for signal content. At this redundancy, the Gaussian window clearly outperforms every other window, with the Bartlett window performing even worse than the Gaussian window at redundancy 8. For both PGHI and FGLA, the effect of the window was not significant and was well below the effect for signal content. From this experiment we conclude that the choice of window does not significantly affect PR algorithms which do not rely on particular structures of the STFT.

F. Effect of the convergence of FGLA

A major drawback of iterative PR algorithms is the necessity to perform multiple time-consuming iterations. In the previous experiments, we were looking for the optimal TF ratio λ and used 100 iterations of FGLA in all comparisons. However, there might be an interaction between the TF ratio λ and the performance per iteration, yielding a different optimum range λ at different number of iterations.

To this end, we investigated the interaction between the STFT parameters and the convergence properties of FGLA on the speech dataset. The evaluation considered the Gaussian window, the redundancy at which FGLA performed best, $D = 8$, and a wide range of TF ratios, $\lambda \in \{10^{-3}, 10^4\}$. The results were collected after 5, 30, 100, and 300 iterations and are presented in Fig. 10.

After only five iterations, the range of TF ratios yielding good PR performance can be identified, both in terms of ODG and, to a lesser extent, of SNR_{MS} . After 30 iterations, this range is clear for both measures. While SNR_{MS} improved with the increasing number of iteration, ODG showed ceiling effects after 100 iterations for a wider range of TF ratios. Both measures agreed that PR performance at 100 iterations was very good, even though SNR_{MS} continued to improve afterwards, at the cost of higher computation time.

In the next step, we looked into the time-performance trade-off for a good TF ratio. To this end, we fixed the TF ratio at $\lambda = 3.34$ and performed PR with FGLA for redundancies $D \in \{2, 4, 8, 16, 32\}$. We set a minimum number of 120 iterations at redundancy 32, such that we always perform more than the default 100 iterations. Per halving redundancy, we doubled the number of iterations, to have similar computation times per redundancy. Figure 11 shows the PR performance plotted

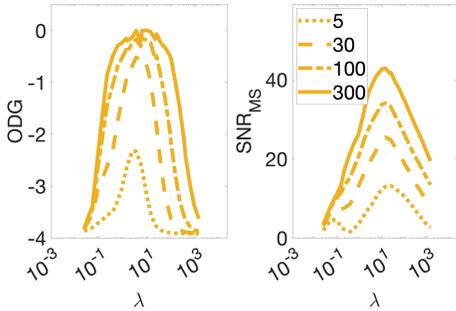


Fig. 10: FGLA's performance after various numbers of iterations for $D = 8$ and various TF ratios λ .

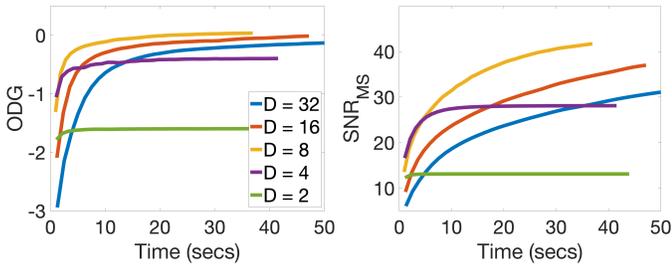


Fig. 11: FGLA's performance as a function of the iteration number for five redundancies and $\lambda = 3.34$.

against the computational time consumed in our workstation⁴. The redundancy $D = 8$ resulted in the best time-performance trade-off, with the exception of the first 5 seconds at $D = 4$, where SNR_{MS} showed slightly better results.

In conclusion, this experiment reveals that the optimal range of λ for the iterative PR algorithm FGLA can be obtained after as little as 5 iterations, greatly reducing the computation time required to find the optimal λ for a new dataset. We also learned that redundancy 8 does not only perform the best in terms of quality, but it also maximizes the performance/computation time trade-off.

G. Optimal parameters for future applications

Our results from the experiments allow us to compose a procedure for finding the optimal set of STFT parameters (λ, D) in future applications. This procedure retrieves the range of optimal parameters based on user's input such as the PR algorithm and the audio content. The work flow is as follows:

- 1) Select the phase-retrieval method to evaluate, the audio signals, and the range of λ and D .
- 2) Select error measures to use for evaluation. We recommend a combination of measures such as SNR_{MS} and ODG .
- 3) Compute the STFT magnitudes with a given (λ, D) , apply the PR method, and compute the average error across all signals.

⁴Our workstation is a Windows 10 PC equipped with an Intel i5 7400 processor and 16 GB of RAM. For these experiments we used the MEX backend for LTFAT and the PHASERET toolbox.

- 4) Increase D and repeat step 3. If the method performs above the given threshold, repeat this step. If it does not, continue with the previous D .
- 5) Repeat step 3 with smaller and larger λ until the performance starts decreasing. In this way, find the range of λ for the given D that perform inside of your given threshold.

An implementation of an algorithm following this guidelines is freely available⁵. As a proof of concept, we applied our algorithm to generate parameter sets for some representative use cases and show the obtained parameters in Table II.

Algorithm	Audio	Best for	D	λ	M
PGHI	Speech	Quality	32	0.14-53.5	320-6144
PGHI	Speech	Speed	16	0.65-11.89	480-2048
FGLA (300)	Speech	Quality	8	2.32-37.15	640-2560
FGLA (50)	Speech	Speed	8	2.32-13.38	640-1536
SPSI	Speech	Quality	8	0.83-1.49	384-512
SPSI	Speech	Speed	4	1.16-1.67	320-384
PGHI	Music	Quality	32	0.21-53.5	384-6144
PGHI	Music	Speed	16	1.16-107.0	640-6144
FGLA (300)	Music	Quality	8	9.28-334.37	1280-7680
FGLA (50)	Music	Speed	8	9.28-95.11	1280-4096
SPSI	Music	Quality	8	3.34-20.9	768-1920
SPSI	Music	Speed	4	6.68-18.57	768-1280

TABLE II: Optimal parameter sets for various use cases. The number in parenthesis after FGLA refers to the number of applied iterations. Audio content as in Sec. III-A. M was derived from λ and the redundancy D . The threshold was set to 15 dB for SNR_{MS} and to 0.3 for ODG .

V. CONCLUSIONS

We systematically studied the effects of STFT parameters and audio content on the quality of PR algorithms. The goal was to demonstrate the effects, develop guidelines for optimizing STFT parameters, and explain why they improve the PR performance. To this end, we considered three classes of algorithms reconstructing phase from STFT magnitude: iterative (represented by FGLA) and non-iterative with and without a signal model (represented by SPSI and PGHI, respectively). Our results show that the PR performance depends on the algorithm, the redundancy D , the TF ratio λ , the signal content, and even the window type. Further, we provide guidelines to find the best combinations of the parameters for a specific application.

As for the algorithms, PGHI showed the best and SPSI the worst performance in terms of SNR and ODG. The performance increased with D , with PGHI showing improvements beyond D of 8, and FGLA and SPSI having their performance limited even for $D > 8$. We explain this clear advantage of PGHI by its direct exploitation of the theoretical phase-magnitude relation present in the continuous STFT. Still, FGLA has the advantage of providing better results at lower redundancies ($D \leq 8$), however, at the cost of significantly higher computation times. To find the optimal STFT parameters for FGLA, as little as five iterations sufficiently represent the relative performance of the fully converged method. In general, PGHI with $D = 32$

⁵<https://github.com/andimarafioti/phaseRetrievalEvaluation>

seems to be a good choice, achieving a high SNR and an ODG corresponding to the category ‘imperceptible’. While such high redundancies are currently rarely considered in practice, their advantages on the performance of phase retrieval may push forward their popularity in future systems.

In all three algorithms, the optimal TF ratio was in the range of 0.2 to 20. The particular choice seems to depend on the type of application, though. For example, in our experiment of PR algorithms applied to processed spectrograms, larger λ provided an improvement. Also, the audio content requires special consideration as we found higher optimal λ required for signals having more energy in the lower frequencies. This can be explained by the need of longer windows for low-frequency content, an issue well-known in the parametrization of STFT for audio applications, which, as our results show, also hold in the problem of phase reconstruction.

The window type had an effect on the PR performance. While for FGLA and SPSI, that effect was tiny compared to that caused by other parameters, for PGHI, the Gaussian window provided up to 20 dB more SNR compared to other windows. This can be explained by the Gaussian STFT adhering most closely to the PGHI model, which relies on the phase-magnitude relations.

While our results demonstrate how to choose the optimal parameters, many applications receive suboptimal TF representations for PR. Thus, future work may consider the development of a system transforming a TF representation computed with a suboptimal set of parameters into TF representations better suited for PR.

REFERENCES

- [1] M.-V. Laitinen, S. Disch, and V. Pulkki, “Sensitivity of human hearing to changes in phase spectrum,” *Journal of the Audio Engineering Society*, vol. 61, no. 11, pp. 860–877, 2013.
- [2] L. Liu, J. He, and G. Palm, “Effects of phase on the perception of intervocalic stop consonants,” *Speech Communication*, vol. 22, no. 4, pp. 403–417, 1997.
- [3] K. K. Paliwal and L. D. Alsteris, “On the usefulness of STFT phase spectrum in human listening tests,” *Speech Communication*, vol. 45, no. 2, pp. 153–170, 2005.
- [4] A. V. Oppenheim and J. S. Lim, “The importance of phase in signals,” *Proceedings of the IEEE*, vol. 69, no. 5, pp. 529–541, 1981.
- [5] J. Allen, “Short term spectral analysis, synthesis, and modification by discrete Fourier transform,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 3, pp. 235–238, 1977.
- [6] J. Wexler and S. Raz, “Discrete Gabor expansions,” *Signal Processing*, vol. 21, no. 3, pp. 207 – 220, 1990.
- [7] E. Vincent, T. Virtanen, and S. Gannot, *Audio source separation and speech enhancement*. John Wiley & Sons, 2018.
- [8] S. Chowdhury, A. V. Portabella, V. Haunschmid, and G. Widmer, “Towards explainable music emotion recognition: The route via mid-level features,” in *Proc. of the 20th International Society for Music Information Retrieval Conference*, 2019, pp. 237–243.
- [9] L. Pepino and L. Bender, “Separación de fuentes musicales mediante redes neuronales convolucionales con múltiples decodificadores,” in *Jornadas de Audio, Acústica y Sonido*. UNTREF, 2018.
- [10] S. Ghose and J. J. Prevost, “Enabling an IoT system of systems through auto sound synthesis in silent video with DNN,” in *2020 IEEE 15th International Conference of System of Systems Engineering (SoSE)*, 2020, pp. 563–568.
- [11] P. Magron, K. Drossos, S. Mimilakis, and T. Virtanen, “Reducing interference with phase recovery in DNN-based monaural singing voice separation,” in *Proc. of INTERSPEECH*, 2018.
- [12] B. Liu, A. Cao, and H. Kim, “Unified signal compression using generative adversarial networks,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 3177–3181.
- [13] A. Marafioti, P. Majdak, N. Holighaus, and N. Perraudin, “GACELA – A generative adversarial context encoder for long audio inpainting,” *IEEE Journal of Selected Topics in Signal Processing, Special issue on reconstruction of audio from incomplete or highly degraded observations*, p. to appear, 2020.
- [14] J. Engel, K. K. Agrawal, S. Chen, I. Gulrajani, C. Donahue, and A. Roberts, “GANSynth: Adversarial neural audio synthesis,” in *Proc. of ICLR*, 2019.
- [15] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, “Phase processing for single-channel speech enhancement: History and recent advances,” *IEEE signal processing Magazine*, vol. 32, no. 2, pp. 55–66, 2015.
- [16] P. Mowlaee, J. Kulmer, J. Stahl, and F. Mayer, *Single Channel Phase-Aware Signal Processing in Speech Communication: Theory and Practice*. Wiley-IEEE Press, 2017.
- [17] P. Magron, R. Badeau, and B. David, “Phase recovery in NMF for audio source separation: An insightful benchmark,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 81–85.
- [18] A. Marafioti, N. Perraudin, N. Holighaus, and P. Majdak, “A context encoder for audio inpainting,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2362–2372, 2019.
- [19] A. Marafioti, N. Holighaus, P. Majdak, and N. Perraudin, “Audio inpainting of music by means of neural networks,” in *Audio Engineering Society Convention 146*, Mar 2019.
- [20] R. W. Harrison, “Phase problem in crystallography,” *Journal of the Optical Society of America A*, vol. 10, no. 5, pp. 1046–1055, 1993.
- [21] J. Miao, T. Ishikawa, Q. Shen, and T. Earnest, “Extending X-ray crystallography to allow the imaging of noncrystalline materials, cells, and single protein complexes,” *Annual Review of Physical Chemistry*, vol. 59, pp. 387–410, 2008.
- [22] F. Fogel, I. Waldspurger, and A. d’Aspremont, “Phase retrieval for imaging problems,” *Mathematical programming computation*, vol. 8, no. 3, pp. 311–335, 2016.
- [23] Y. Shechtman, Y. C. Eldar, O. Cohen, H. N. Chapman, J. Miao, and M. Segev, “Phase retrieval with application to optical imaging: a contemporary overview,” *IEEE signal processing magazine*, vol. 32, no. 3, pp. 87–109, 2015.
- [24] T. Bendory, R. Beinert, and Y. C. Eldar, *Fourier Phase Retrieval: Uniqueness and Algorithms*. Cham: Springer International Publishing, 2017, pp. 55–91.
- [25] R. Balan, P. Casazza, and D. Edidin, “On signal reconstruction without phase,” *Applied and Computational Harmonic Analysis*, vol. 20, no. 3, pp. 345–356, 2006.
- [26] S. Nawab, T. Quatieri, and J. Lim, “Signal reconstruction from short-time Fourier transform magnitude,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 31, no. 4, pp. 986–998, 1983.
- [27] I. Bojarovska and A. Flinth, “Phase retrieval from Gabor measurements,” *Journal of Fourier Analysis and Applications*, vol. 22, no. 3, pp. 542–567, 2016.
- [28] K. Jaganathan, Y. C. Eldar, and B. Hassibi, “STFT phase retrieval: Uniqueness guarantees and recovery algorithms,” *IEEE Journal of selected topics in signal processing*, vol. 10, no. 4, pp. 770–781, 2016.
- [29] L. Li, C. Cheng, D. Han, Q. Sun, and G. Shi, “Phase retrieval from multiple-window short-time Fourier measurements,” *IEEE Signal Processing Letters*, vol. 24, no. 4, pp. 372–376, 2017.
- [30] R. Alaifari and M. Wellershoff, “Ill-conditionedness of discrete Gabor phase retrieval and a possible remedy,” in *2019 13th International conference on Sampling Theory and Applications (SampTA)*, 2019, pp. 1–4.
- [31] I. Waldspurger, “Phase retrieval with random gaussian sensing vectors by alternating projections,” *IEEE Transactions on Information Theory*, vol. 64, no. 5, pp. 3301–3312, 2018.
- [32] D. Griffin and J. Lim, “Signal estimation from modified short-time Fourier transform,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.
- [33] Y. Masuyama, K. Yatabe, Y. Koizumi, Y. Oikawa, and N. Harada, “Deep Griffin–Lim iteration: Trainable iterative phase reconstruction using neural network,” *IEEE Journal of Selected Topics in Signal Processing*, pp. 1–1, 2020.
- [34] Y. Masuyama, K. Yatabe, Y. Koizumi, Y. Oikawa, and N. Harada, “Deep Griffin–Lim iteration,” in *Proc. of ICASSP*. IEEE, 2019, pp. 61–65.

- [35] Y. Masuyama, K. Yatabe, Y. Koizumi, Y. Oikawa, and N. Harada, "Phase reconstruction based on recurrent phase unwrapping with deep neural networks," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 826–830.
- [36] Y. Masuyama, K. Yatabe, and Y. Oikawa, "Griffin–Lim like phase recovery via alternating direction method of multipliers," *IEEE Signal Processing Letters*, vol. 26, no. 1, pp. 184–188, 2019.
- [37] J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama, "Fast signal reconstruction from magnitude STFT spectrogram based on spectrogram consistency," in *Proc. Int. Conf. Digital Audio Effects*, vol. 10, 2010.
- [38] N. Perraudin, P. Balazs, and P. L. Søndergaard, "A fast Griffin-Lim algorithm," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*. IEEE, 2013, pp. 1–4.
- [39] X. Zhu, G. T. Beauregard, and L. Wyse, "Real-time iterative spectrum inversion with look-ahead," in *2006 IEEE International Conference on Multimedia and Expo*, 2006, pp. 229–232.
- [40] S. I. Mimitakis, K. Drossos, J. F. Santos, G. Schuller, T. Virtanen, and Y. Bengio, "Monaural singing voice separation with skip-filtering connections and recurrent inference of time-frequency mask," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 721–725.
- [41] S. Vasquez and M. Lewis, "MelNet: A generative model for audio in the frequency domain," in *Proc. of ICLR*, 2020.
- [42] G. T. Beauregard, M. Harish, and L. Wyse, "Single pass spectrogram inversion," in *2015 IEEE international conference on digital signal processing (DSP)*, 2015, pp. 427–431.
- [43] P. Magron, R. Badeau, and B. David, "Phase reconstruction of spectrograms with linear unwrapping: application to audio signal restoration," in *2015 23rd European Signal Processing Conference (EUSIPCO)*, 2015, pp. 1–5.
- [44] Z. Průša, P. Balazs, and P. L. Søndergaard, "A noniterative method for reconstruction of phase from STFT magnitude," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 25, no. 5, May 2017.
- [45] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Transactions on Speech and Audio processing*, vol. 7, no. 3, pp. 323–332, 1999.
- [46] M. Portnoff, "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, 1979, pp. 186–189.
- [47] A. Marafioti, N. Perraudin, N. Holighaus, and P. Majdak, "Adversarial generation of time-frequency features with application in audio synthesis," in *Proc. of the 36th ICML*, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. Long Beach, California, USA: PMLR, 09–15 Jun 2019, pp. 4352–4362.
- [48] P. Mowlae, R. Saeidi, and Y. Stylianou, "Advances in phase-aware signal processing in speech communication," *Speech Communication*, vol. 81, pp. 1–29, 2016.
- [49] Z. Průša and P. L. Søndergaard, "Real-time spectrogram inversion using phase gradient heap integration," in *Proc. Int. Conf. Digital Audio Effects (DAFx-16)*, Sep 2016, pp. 17–21.
- [50] N. Holighaus, G. Koliander, Z. Průša, and L. D. Abreu, "Characterization of analytic wavelet transforms and a new phaseless reconstruction algorithm," *IEEE Transactions on Signal Processing*, vol. 67, no. 15, pp. 3894–3908, 2019.
- [51] N. Holighaus, G. Koliander, L. D. Abreu, and Z. Průša, "Non-iterative phaseless reconstruction from wavelet transform magnitude," in *Proceedings of the 22nd International Conference on Digital Audio Effects, Birmingham, UK*, 2019, pp. 2–6.
- [52] M. Portnoff, "Implementation of the digital phase vocoder using the fast Fourier transform," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 24, no. 3, pp. 243–248, 1976.
- [53] F. Auger, É. Chassande-Mottin, and P. Flandrin, "On phase-magnitude relationships in the short-time Fourier transform," *IEEE Signal Process. Lett.*, vol. 19, no. 5, pp. 267–270, 2012.
- [54] K. Gröchenig, *Foundations of Time-Frequency Analysis*, ser. Appl. Numer. Harmon. Anal. Birkhäuser, 2001.
- [55] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. Jarrod Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. Carey, Í. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.
- [56] Z. Průša, P. L. Søndergaard, N. Holighaus, C. Wiesmeyer, and P. Balazs, "The large time-frequency analysis toolbox 2.0," in *Sound, Music, and Motion*, ser. LNCS. Springer International Publishing, 2014, pp. 419–442.
- [57] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035.
- [58] T. Strohmer and S. Beaver, "Optimal OFDM system design for time-frequency dispersive channels," *IEEE Trans. Comm.*, vol. 51, no. 7, pp. 1111–1122, July 2003.
- [59] M. Faulhuber and S. Steinerberger, "Optimal Gabor frame bounds for separable lattices and estimates for Jacobi theta functions," *Journal of Mathematical Analysis and Applications*, vol. 445, no. 1, pp. 407–422, 2017.
- [60] T. Strohmer, "Numerical algorithms for discrete Gabor expansions," in *Gabor Analysis and Algorithms: Theory and Applications*, ser. Applied and Numerical Harmonic Analysis, H. G. Feichtinger and T. Strohmer, Eds. Birkhäuser Boston, 1998, pp. 267–294.
- [61] A. Janssen, "From continuous to discrete Weyl-Heisenberg frames through sampling," *Journal of Fourier Analysis and Applications*, vol. 3, no. 5, pp. 583–596, 1997.
- [62] K. Ito and L. Johnson, "The LJ speech dataset," <https://keithito.com/LJ-Speech-Dataset/>, 2017.
- [63] C. Raffel, "Learning-based methods for comparing sequences, with applications to audio-to-MIDI alignment and matching," Ph.D. dissertation, 2016.
- [64] C. Raffel and D. P. W. Ellis, "Intuitive analysis, creation and manipulation of MIDI data with pretty_midi," in *Proc. of the 15th ISMIR, Late Breaking and Demo Papers*, 2014.
- [65] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "Fma: A dataset for music analysis," in *18th International Society for Music Information Retrieval Conference*, 2017.
- [66] Z. Průša, "The phase retrieval toolbox," in *AES International Conference On Semantic Audio*, Erlangen, Germany, June 2017.
- [67] R. W. Gerchberg and W. O. Saxton, "A practical algorithm for the determination of phase from image and diffraction plane pictures," *Optik*, vol. 35, pp. 237–246, 1972.
- [68] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, and C. Colomes, "PEAQ-The ITU standard for objective measurement of perceived audio quality," *J. Aud. Eng. Soc.*, vol. 48, no. 1/2, pp. 3–29, 2000.
- [69] ITU-R Recommendation, "1387: Method for objective measurements of perceived audio quality," *International Telecommunication Union, Geneva, Switzerland*, 2001.
- [70] P. Kabal *et al.*, "An examination and interpretation of ITU-R BS. 1387: Perceptual evaluation of audio quality," *TSP Lab Technical Report, Dept. Electrical & Computer Engineering, McGill University*, pp. 1–89, 2002.
- [71] R. Huber and B. Kollmeier, "Pemo-q—a new method for objective audio quality assessment using a model of auditory perception," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 6, pp. 1902–1911, 2006.
- [72] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, 2011.