Timo Gerkmann, Martin Krawczyk-Becker, and Jonathan Le Roux

Phase Processing for Single-Channel Speech Enhancement

History and recent advances

With the advancement of technology, both assisted listening devices and speech communication devices are becoming more portable and also more frequently used. As a consequence, users of devices such as hearing aids, cochlear implants, and mobile telephones, expect their devices to work robustly anywhere and at any time. This holds in particular for challenging noisy environments like a cafeteria, a restaurant, a subway, a factory, or in traffic. One way to making assisted listening devices robust to noise is to apply speech enhancement algorithms. To improve the corrupted speech, spatial diversity can be exploited by a constructive combination of microphone signals (so-called beamforming), and by exploiting the different spectrotemporal properties of speech and noise. Here, we focus on single-channel speech enhancement algorithms which rely on spectrotemporal properties. On the one hand, these algorithms can be employed when the miniaturization of devices only allows for using a single microphone. On the other hand, when multiple microphones are available, single-channel algorithms can be employed as a postprocessor at the output of a beamformer. To exploit the short-term stationary properties of natural sounds, many of these approaches process the signal in a time-frequency representation, most frequently the short-time discrete Fourier transform (STFT) domain. In this domain, the coefficients of the signal are complex-valued, and can therefore be represented by their absolute value (referred to in the literature both as STFT magnitude and STFT amplitude) and their phase. While the modeling and processing of the STFT magnitude has been the center of interest in the past three decades, phase has been largely ignored.

In this article, we review the role of phase processing for speech enhancement in the context of assisted listening and speech communication devices. We explain why most of the research conducted in this field used to focus on estimating spectral magnitudes in the STFT domain, and why recently phase processing is attracting increasing interest in the speech

Digital Object Identifier 10.1109/MSP.2014.2369251
Date of publication: 12 February 2015

enhancement community. Furthermore, we review both early and recent methods for phase processing in speech enhancement. We aim to show that phase processing is an exciting field of research with the potential to make assisted listening and speech communication devices more robust in acoustically challenging environments.

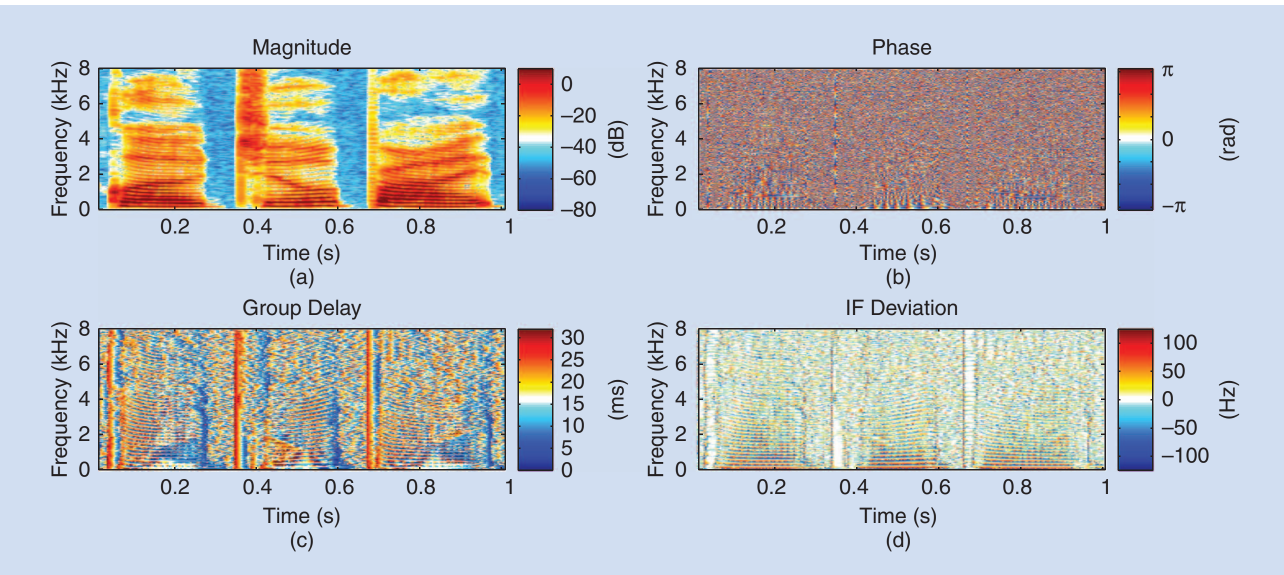
WITH THE ADVANCEMENT OF TECHNOLOGY, BOTH ASSISTED LISTENING DEVICES AND SPEECH COMMUNICATION DEVICES ARE BECOMING MORE PORTABLE AND ALSO MORE FREQUENTLY USED.

INTRODUCTION

Let us first consider the common speech enhancement setup consisting of STFT analysis, spectral modification, and subsequent inverse STFT (iSTFT) resynthesis. The analyzed digital signal $x(n)$, with time index n, is chopped into L segments with a length of N samples, overlapping by $N-R$ samples, where R denotes the segment shift. Each segment ℓ is multiplied with the appropriately shifted analysis window $w_a(n-\ell R)$ and transformed into the frequency domain by applying the discrete Fourier transform (DFT), yielding the complex-valued STFT coefficients $X_{k,\ell} \in \mathbb{C}$ for every segment ℓ and frequency band k. To compactly describe this procedure, we define the STFT operator: $X = \text{STFT}(\boldsymbol{x})$. Here, $\boldsymbol{x}$ is a vector containing the complete time-domain signal $x(n)$ and X is an $N \times L$ matrix of all $X_{k,\ell}$, which we will refer to as the *spectrogram*. Since we are interested in real-valued acoustic signals, we consider only complex symmetric spectrograms $X \in \mathcal{S} \subset \mathbb{C}^{N \times L}$, where $\mathcal{S}$ denotes the subset of spectrograms for which $X_{N-k,\ell} = \overline{X}_{k,\ell}$ for all ℓ and k, with $\overline{X}$ being the complex conjugate of X.

After some processing, such as magnitude improvement, is applied on the STFT coefficients, a modified spectrogram $\widetilde{X}$ is obtained. From $\widetilde{X}$ a time-domain signal can be resynthesized through an iSTFT operation, denoted by $\widetilde{\boldsymbol{x}} = \text{iSTFT}(\widetilde{X})$. For this, the inverse DFT of the STFT coefficients is computed and each segment is multiplied by a synthesis window $w_s(n-\ell R)$; the windowed segments are then overlapped and added to obtain the modified time-domain signal. A final renormalization step is performed to ensure that, if no processing is applied to the spectral coefficients, there is perfect reconstruction of the input signal, i.e., $\text{iSTFT}(\text{STFT}(\boldsymbol{x})) = \boldsymbol{x}$. The renormalization term, equal to $\sum_{q=-\infty}^{+\infty} w_{\text{a}}(n+qR)\, w_{\text{s}}(n+qR)$, is R-periodic and can be included in the synthesis window. A common choice for both $w_a(n)$ and $w_s(n)$ is the square-root Hann window, which for overlaps such that $N/R \in \mathbb{N}$ (e.g., 50%, 75%, etc.) only requires normalization by a scalar. If the spectrogram is modified, using the same window for synthesis as for analysis can be shown to lead to a resynthesized signal whose spectrogram is closest to $\widetilde{X}$ in the least-squares sense [1]. This fact will turn out to be important for the iterative phase estimation approaches discussed later.

Until recently, in STFT-based speech enhancement, the focus was on modifying only the magnitude of the STFT components, because it was generally considered that most of the insight about the structure of the signal could be obtained from the magnitude, while little information could be obtained from the phase component. This would seem to be substantiated by Figure 1 when considering only (a) and (b), where the STFT magnitude (a) and STFT phase (b) of a clean speech excerpt are depicted. In contrast to the magnitude spectrogram, the phase spectrogram appears to show only little temporal and spectral regularities. There are nonetheless distinct structures inherent to the spectral phase, but they are hidden to a great extent because the phase is



[FIG1] (a) Magnitude spectrogram, (b) phase spectrogram, (c) group delay, and (d) IF deviation of the utterance "glowed jewel-bright" using a segment length of 32 ms and a shift of 4 ms.

wrapped to its principle value, i.e., $-\pi \leq \phi^X_{k,\ell} = \angle X_{k,\ell} \leq \pi$. To reveal these structures, alternative representations have been proposed, which consider phase relations between neighboring time-frequency points instead of absolute phases. Two examples of such representations are depicted in Figure 1(c) and (d). In (c), the negative derivative of the phase along frequency, known as the *group delay*, is shown. It has been shown to be a useful tool for speech enhancement, e.g., by Yegnanarayana and Murthy [2]. Besides the group delay, the derivative of the phase along time, i.e., the instantaneous frequency (IF), also unveils structures in the spectral phase. For an improved visualization, in (d), we do not show the IF, but rather its deviation from the respective center frequency in Hz, which reduces wrapping along frequency [3], [4]. It is interesting to remark that the temporal as well as the spectral derivatives of the phase both reveal structures similar to those in the magnitude spectrogram in (a). Please note that both phase transformations are invertible and thus carry the same information as the phase itself. No additional prior knowledge has been injected.

The observed structures in the spectral phase can well be explained by employing models of the underlying signal, e.g., by sinusoidal models in the case of voiced speech [5]. Besides the structures in the phase that are caused by signal characteristics, neighboring time-frequency points also show dependencies due to the STFT analysis: first, because of the finite length of the segments, neighboring frequency bands are not independent; second, successive segments overlap and hence share partly the same signal information. This introduces particular spectrotemporal relations between STFT coefficients within and across frames of the spectrogram, regardless of the signal. If the spectrogram is modified, these relations are not guaranteed to be maintained and the modified spectrogram $\widetilde{X}$ may not correspond to the STFT of any time-domain signal anymore. As a consequence, the resynthesized signal may have a spectrogram $\mathcal{G}(\widetilde{X})$, where

$$\mathcal{G}(\widetilde{X}) := \mathrm{STFT}(\mathrm{iSTFT}(\widetilde{X})), \qquad (1)$$

which is different from the desired spectrogram $\widetilde{X}$, as illustrated in Figure 2. Such spectrograms are called *inconsistent*, while *consistent* spectrograms verify $\mathcal{G}(X) = X$ and can be obtained from a time-domain signal.

Since the majority of speech enhancement approaches only modify the magnitude, the mismatch between the enhanced magnitude and the degraded phase will most likely lead to an inconsistent spectrogram. This implies that even if the estimated magnitudes $|\widetilde{X}|$ are optimal with respect to some objective function, the magnitude spectrogram of the synthesized time-domain signal is not, as $|\mathcal{G}(\widetilde{X})| \neq |\widetilde{X}|$ (where $|\cdot|$ denotes the element-wise absolute value). To maintain consistency, and thus also optimality, the STFT phase has to be taken into account as well.

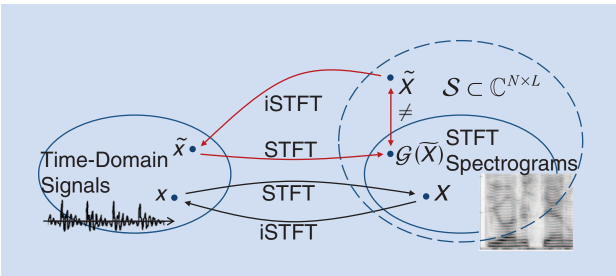
As a final illustration emphasizing the power of phase, it is interesting to remark that, from a particular magnitude spectrogram, it is possible to reconstruct virtually any time-domain signal with a carefully crafted phase. For instance, one can derive a magnitude spectrogram from that of a speech signal such that it yields either a speech signal similar to the original or a piece of rock music, depending on the choice of the phase. The point here is to exploit the inconsistency between magnitude and phase to make contributions of neighboring frames cancel each other just enough to reconstruct the energy profile of the target sound. Reconstruction is thus done up to a scaling factor, and quality is good albeit limited by dynamic range issues. An audio demonstration is available in http://www.jonathanleroux.org/research/LeRoux2011ASJ03_sound_transfer.html.

SPEECH ENHANCEMENT IN THE STFT DOMAIN

Speech enhancement is a field of research with a long-standing history. In this section, we will wrap up the different fields of research that have led to remarkable progress over the years. For a more detailed treatment and references to the original publications, see [6].

In the STFT domain, noisy spectral coefficients can, for instance, be improved using spectral subtraction or using minimum mean squared error (MMSE) estimators of the clean speech spectral coefficients [6, Ch. 4]. Examples of the latter are the Wiener filter as an estimator of the complex speech coefficients and the short-time spectral amplitude estimator [7]. These MMSE estimators are driven by estimates of the speech and noise power spectral densities (PSDs). The noise PSDs can be estimated in speech pauses as signaled by a voice activity detector, by searching for spectral minima in each subband, or based on the speech presence probability [6, Ch. 6]. With the noise PSD at hand, the speech PSD can be estimated by subtracting the noise PSD from the periodogram of the noisy signal. This has been shown to be the maximum likelihood (ML) optimal estimator of the clean speech PSD when considering isolated and independent time-frequency points and complex Gaussian distributed speech and noise coefficients [6, Sec. 4.2]. To reduce outliers, the ML speech PSD estimate is often smoothed, for instance, using the decision-directed approach [7] or more advanced smoothing techniques [6, Ch. 7].

Over the years, many improvements have been proposed resulting in a considerable progress thanks to better statistical models of speech and noise [6, Ch. 3], improved estimation of speech and noise PSDs [6, Ch. 6 and 7], combination with speech presence probability estimators [6, Ch. 5], and integration of perceptual models [6, Sec. 2.3.3]. Recent years have seen an explosion of interest in data-driven methods, with model-based approaches



[FIG2] **An illustration of the notion of consistency.**

such as nonnegative matrix factorization, hidden Markov models, and discriminative approaches such as deep neural networks. However, mainstream approaches have tended to ignore the phase, mainly due to the difficulty of modeling it and the lack of clarity about its importance, as discussed next.

RISE, DECLINE, AND RENAISSANCE OF PHASE PROCESSING FOR SPEECH ENHANCEMENT

The first proposals for noise reduction in the STFT domain arose in the late 1970s. While the spectral subtraction approaches only modified the spectral magnitudes, the role of the STFT phase was also actively researched at the time. In particular, several authors investigated conditions under which a signal is uniquely specified by only its phase or only its magnitude and proposed iterative algorithms for signal reconstruction from either one or the other (e.g., [1], [8], and references therein). For minimum or maximum phase systems, log-magnitude and phase are related through the Hilbert transform, meaning that only the spectral phase (or only the spectral magnitude) is required to reconstruct the entire signal. But the constraint of purely minimum or maximum phase is too restrictive for real audio signals, and Quatieri [8] showed that more constraints are needed for mixed-phase signals. For instance, imposing a causality or a finite-length constraint on the signal and specifying a few samples of the phase or the signal itself is in some cases sufficient to uniquely characterize the entire phase function from only the magnitude. Quatieri [8] also showed how to exploit such constraints to estimate a signal from its spectral magnitude: assuming some time-domain samples are known, and starting with an initial phase estimate and the known spectral magnitude, the signal is transformed to the time domain, where the given set of known samples is used to replace the corresponding time-domain samples. Then the time-domain signal is transformed back to the frequency domain, where the resulting magnitude is replaced by the known magnitude. This procedure is repeated for a certain number of iterations. In the case of the STFT domain, the correlation between overlapping short-time analysis segments can be exploited to derive similar iterative algorithms that do not require time-domain samples to be known. A popular example of such methods is that of Griffin and Lim (GL) [1], which we describe in more detail later along with more recent approaches. While algorithms such as GL can also be employed with magnitudes that are estimated rather than measured from an actual signal, the quality of the synthesized speech and the estimated phase strongly depends on the accuracy of the estimated speech spectral magnitudes and artifacts such as echo, smearing, and modulations may occur [9].

To explore the relevance of phase estimation for speech enhancement, Wang and Lim [10] performed listening experiments where the magnitude of a noisy speech signal at a certain signal-to-noise ratio (SNR) was combined with the phase of the same speech signal but distorted by noise at a different SNR. Listeners were asked to compare this artificial test stimulus to a noisy reference speech signal and to set the SNR of the reference such that the perceived quality was the same for the reference and the test stimulus. The result of this experiment was that the SNR gain obtained by mixing noisy magnitudes with a less distorted phase resulted in typical SNR improvements of 1 dB or less. Hence, Wang and Lim concluded that improving phase was not critical in speech enhancement [10]. Similarly, Vary [11] showed that only for local SNRs below 6 dB a certain roughness could be perceived if the noisy phase was kept unchanged. Finally, Ephraim and Malah [7] investigated the role of phase improvement from a statistical perspective: they showed that, under a zero-mean circular Gaussian speech and noise model and assuming that time-frequency points are mutually independent given the speech and noise PSDs, the MMSE estimate of the complex exponential of the speech phase has an argument equal to the noisy phase. Also, for more general models for the speech magnitudes with the same circularity assumption, it has been shown that the noisy phase is the ML optimal estimator of the clean speech phase, e.g., [12]. Note, however, that the independence assumption does not hold in general, and especially not for overlapping STFT frames, where part of the relationship is actually deterministic.

As a consequence of these observations, subsequent research in speech enhancement focused mainly on improving magnitude estimation, while phase estimation received far less attention for the next two decades. Even methods that considered phase, either by use of complex domain models, or by integrating out phase in log-magnitude-based models in a sophisticated way [13], ultimately used the noisy phase because of similar circularity assumptions.

However, as the performance of magnitude-only methods can only go so far without considering phase, and with the increase in computational power of assisted listening and speech communication devices, all options for improvements are back on the table. Therefore, researchers started reinvestigating the role of the STFT phase for speech intelligibility and quality [14], [15]. For instance, Kazama et al. [14] investigated the influence of the STFT segment length on the role of phase for speech intelligibility for a segment overlap of 50%. They found that, while for signal segments between 4 ms and 64 ms the STFT magnitude spectrum is more important than the phase spectrum, for segments shorter than 2 ms and segments longer than 128 ms, the phase spectrum is more important. These results are consistent with Wang and Lim's earlier conclusions [10]. To focus on practical applications, Paliwal et al. [15] investigated signal segments of 32 ms length, but in contrast to Wang and Lim [10] and Kazama et al. [14], they used a segment overlap of 7/8th instead of 1/2 in the STFT analysis, and they also zero-padded the time segments before computing the Fourier transform. With this increased redundancy in the STFT, the performance of existing magnitude-based speech enhancement can be significantly improved [15] if combined with enhanced phases. For instance, Paliwal et. al [15, case 4] report an improvement of 0.2 points of the mean opinion score (MOS) predicted by the instrumental "perceptual evaluation of speech quality" (PESQ) measure for white Gaussian noise at an SNR of 0 dB when combining an MMSE estimate of the clean speech magnitude with the oracle clean speech phase in a perfectly reconstructing STFT framework.

Paliwal et al.'s research confirmed the importance of developing and improving phase processing algorithms. This has recently been the focus of research by multiple groups. We now survey the main directions that have been investigated so far: better and faster phase estimation from magnitude, modeling of the signal phase, group delay and transient processing, and joint estimation of phase and magnitude.

ITERATIVE ALGORITHMS FOR PHASE ESTIMATION

Among the first proposals for phase estimation are iterative approaches, which aim at estimating a time-domain signal whose STFT magnitude is as close as possible to a target one [1], [8]. Indeed, if the STFT magnitude of two signals are close, the signals will in general be perceptually close as well. Thus, finding a signal whose STFT magnitude is close to a target one is considered a valid goal when looking to obtain a signal that "sounds" like that target magnitude. This motivated intense research on algorithms to estimate signals (or equivalently a corresponding phase) given target magnitudes, with applications such as speech enhancement or timescale modification. In the case of speech enhancement, the magnitude is typically obtained through one of the many magnitude estimation algorithms mentioned earlier, while some estimate of the phase, such as that of the noisy mixture, may further be exploited for initialization or as side information.

FINDING A SIGNAL WHOSE STFT MAGNITUDE IS CLOSE TO A TARGET ONE IS CONSIDERED A VALID GOAL WHEN LOOKING TO OBTAIN A SIGNAL THAT "SOUNDS" LIKE THAT TARGET MAGNITUDE.

The most well known and fundamental of these approaches is that of Griffin and Lim [1], which consists in applying STFT synthesis and analysis iteratively while retaining information about the updated phases and replacing the updated magnitudes by the given ones. This exploits correlations between neighboring STFT frames to lead to an estimate of the spectral phases and the time-domain signal.

Given a target magnitude spectrogram A, Griffin and Lim formulated the problem as that of estimating a real-valued time-domain signal x such that the magnitude of its STFT X is closest to A in the least-squares sense, i.e., estimating a signal x which minimizes the squared distance

$$d(x,A) = \sum_{k,\ell} \| |X_{k,\ell}| - A_{k,\ell} |^2. \tag{2}$$

They proposed an iterative procedure which can be proven to minimize, at least locally, this distance. Starting from an initial signal estimate $x^{(0)}$ such as random noise, iterate the following computations: compute the STFT $X^{(i)}$ of the signal estimate $x^{(i)}$ at step i; compute the phase estimate $\phi^{(i)}$ as the phase of $X^{(i)}$, $\phi^{(i)} = \angle X^{(i)}$; compute the signal estimate $x^{(i+1)}$ at step $i+1$ as the iSTFT of $Ae^{j\phi^{(i)}}$. Using the operator $\mathcal{G}$ defined in (1), this can be reformulated as

$$\phi^{(i+1)} = \angle \mathcal{G}(Ae^{j\phi^{(i)}}). \tag{3}$$

This procedure can be proven to be nonincreasing as well for a measure of inconsistency of the spectrogram $Ae^{j\phi^{(i)}}$ defined directly in the time-frequency domain:

$$\mathcal{I}(\phi) = \| \mathcal{G}(Ae^{j\phi}) - Ae^{j\phi} \|_2^2. \tag{4}$$

Indeed, one can easily show that $d(x^{(i+1)}, A) \le \mathcal{I}(\phi^{(i)}) \le d(x^{(i)}, A)$. Interestingly, if only parts of the phase are updated according to (3), the nondecreasing property still holds for $\mathcal{I}(\phi)$, but whether it still does for $d(x, A)$ has not been established.

Due to the extreme simplicity of its implementation and to its perceptually relatively good results, GL was used as the standard benchmark and a starting point for multiple extensions in the three decades that have followed, even after better and only marginally more involved algorithms had been devised. Most of the algorithms that have been developed since attempted to fix GL's issues, of which there are several: first, convergence typically requires many iterations; second, GL does not provide a good initial estimate, starting from random phases with no considerations for cross-frame dependencies; third, the updates rely on computing STFTs, which are computationally costly even when implemented using fast Fourier transforms (FFTs); fourth, the updates are typically performed on whole frames, without emphasis on local regularities; and finally, the original version of GL processes signals in batch mode.

On this last point, it is interesting to note that Griffin and Lim did actually hint at how to modify their algorithm to use it for online applications. They described briefly in [1] and with more details in [16] how to sequentially update the phase using "cascaded processors" that each take care of one iteration; their particular proposal however still incurs an algorithmic delay of I times the window length if performing I iterations. In [16], Griffin also presented several methods that he referred to as "sequential estimation methods": these only incur a single frame delay and could thus be used for online application, the best performing one being reported as on par with batch GL.

While one can already see in Griffin's account [16] several elements to modify GL into an algorithm that can lead to high quality reconstruction in a real-time setting, such as sliding-block analysis across the signal and the use of windows that compensate for partially reconstructed frames, these ideas seem to have gone largely unnoticed and it is not until much later that they were incorporated into more refined methods. Beauregard, Zhu, and Wyse proposed consecutively two algorithms for real-time signal reconstruction from STFT magnitude, the real-time iterative spectrogram inversion (RTISI) algorithm and RTISI with look ahead (RTISI-LA) [17]. RTISI aims at improving the original batch GL in two respects: allowing for online implementation, and generating better initial phase estimates. The algorithm considers the frames sequentially in order, and at frame ℓ, it only uses information from the current frame's magnitude and the previous overlapping frames. The initial phase estimate $\phi_\ell^{(0)}$ for frame ℓ is obtained as the phase of the partial reconstruction from the previous frames, windowed by an analysis window, which already ensures some consistency between the phases of the current and previous frames. An iterative procedure similar to GL is then applied, limited to the current frame's phase: at each iteration, frame ℓ's contribution to the signal is obtained by the inverse DFT of the phase $\phi_\ell^{(i)}$ combined with the target magnitude; frame ℓ's contribution is then combined by overlap-add to the contribution of the previous frames, leading to a signal estimate for frame ℓ; the phase $\phi_\ell^{(i+1)}$ is estimated as the phase of this signal estimate to which the analysis window is applied.

RTISI does lead to better results than GL for the first few iterations, but it quickly reaches a plateau and is ultimately significantly outperformed by GL. This is mainly due to the fact that RTISI does not consider information from future frames at all, even though the contribution of these future frames will later on be added to that of the past and current frames, effectively altering the estimation performed earlier. Its authors thus proposed an extension to RTISI including an M frame look-ahead, RTISI-LA. Instead of considering only the current frame as active, RTISI-LA performs GL-type updates on the phases in a block of multiple frames. The contribution of future frames outside the block is discarded during the updates, because the absence of a reliable phase estimate for them is regarded as likely to make their contribution more of a disturbance than a useful clue. This creates an asymmetry, which Zhu et al. [17] proposed to partially compensate by using asymmetric analysis windows with a reverse effect. Although the procedure relies on heuristic considerations, the authors show that it leads to much better performance than GL for a given number of iterations per block.

While RTISI and RTISI-LA were successful in overcoming GL's issues regarding online processing and poor initialization, they did not tackle the problems of heavy reliance on costly FFT computations and lack of care for local regularities in the time-frequency domain. Solving these problems was difficult in the context of classical approaches relying on enforcing constraints both in the time-frequency domain (to impose a given magnitude) and the time domain (to ensure that magnitude and phase are consistent), because they inherently had to go back and forth between the two domains, processing whole frames at a time. A solution was proposed by Le Roux et al. [18], whose key idea was to bypass the time domain altogether and reformulate the problem within the time-frequency domain. The standard operation of classical iterative approaches, i.e., computing the STFT of the signal obtained by iSTFT from a given spectrogram, can indeed be considered as a linear operator in the time-frequency domain. Le Roux et al. noticed that the result of that operation at each time-frequency bin can be well approximated by a local weighted sum (LWS) with complex coefficients on a small neighborhood of that bin in the original spectrogram. While the very small number of terms in the sum does not suffice to reduce the complexity of the operation compared to using FFTs, the locality of the sum opens the door to selectively updating certain time-frequency bins, as well as to immediately propagating the updated value for a bin in the computations of its neighbors' updates. Taking advantage of the sparseness of natural sound signals, Le Roux et al. showed in particular that focusing first on updating only the bins with high energy not only reduced greatly the complexity of each iteration, but also could lead to better initializations, the high energy regions serving as anchors for lower energy ones. While the LWS algorithm was originally proposed as an extension to GL for batch-mode computations, the authors later showed that it could be effectively used in online mode as well in combination with RTISI-LA [19]. Interestingly, a different prioritization of the updates based on energy, at the frame level instead of the bin level, was also successfully used by Gnann and Spiertz to improve RTISI-LA [20].

Recently, several authors investigated signal reconstruction from magnitudes with specific task-related side information. Those developed in the context of source separation are of particular interest to this article. Gunawan and Sen [21] proposed the multiple input spectrogram inversion (MISI) algorithm to reconstruct multiple signals from their magnitude spectrograms and their mixture signal. The phase of the mixture signal acts as very powerful side information, which can be exploited by imposing that the reconstructed complex spectrograms add up to the mixture complex spectrogram when estimating their phases, leading to much better reconstruction quality than in situations where the mixture signal is not available. Sturmel and Daudet's partitioned phase retrieval (PPR) method [9] also handles the reconstruction of multiple sources. Their proposal was to reconstruct the phase of the magnitude spectrogram obtained by Wiener filtering by applying a GL-like algorithm, which keeps the mixture phase in high SNR regions as a good estimate for the corresponding source and only updates the phase in low- to mid-SNR regions. Both methods, however, only modify the phase of the sources, and thus implicitly assume that the input magnitude spectrograms are close to the true source spectrograms, which is not realistic in general in the context of blind or semiblind source separation. Sturmel and Daudet proposed to extend MISI to allow for modifications of both the magnitude and phase, leading to the informed source separation using iterative reconstruction (ISSIR) method [22], and showed that it is efficient in the context of informed source separation where a quantized version of the oracle magnitude spectrograms is available. Methods to jointly estimate phase and magnitude for blind source separation and speech enhancement will be presented later.

SINUSOIDAL MODEL-BASED PHASE ESTIMATION

In contrast to the iterative approaches presented in the previous section, sinusoidal model-based phase estimation [4] does not require estimates of the clean speech spectral magnitudes. Instead, the clean spectral phase is estimated using only an estimate of the fundamental frequency, which can be obtained from the degraded signal. However, since usage of the sinusoidal model is reasonable only for voiced sounds, these approaches do not provide valid spectral phase estimates for unvoiced sounds, like fricatives or plosives.

For a single sinusoid, $\sin(\Omega n + \varphi)$, with normalized angular frequency Ω, the phase difference between two samples $n_2 = n_1 + R$ is given by $\Delta\phi = \phi(n_2) - \phi(n_1) = \Omega R$. For a harmonic signal, H sinusoids at integer multiples of the normalized angular fundamental frequency Ω_0, i.e., $\Omega^h = (h+1)\Omega_0 \in [0, 2\pi)$, are present at the same time:

$$s(n) = \sum_{h=0}^{H-1} A^h(n) \cos(\Omega^h(n) \cdot n + \varphi^h), \tag{5}$$

with real-valued amplitude A^h and initial time-domain phase φ^h for harmonic component h. Due to the fixed relation between the frequencies, (5) is also referred to as the *harmonic model*, which is a special case of the more general sinusoidal model. The harmonic frequencies and amplitudes are assumed to be slowly changing over time with respect to the length N of an STFT signal segment and we define $A_\ell^h = A^h(\ell R + N/2)$ and $\Omega_\ell^h = \Omega^h(\ell R + N/2)$ as the representative harmonic amplitudes and frequencies for the ℓth signal segment.

In speech enhancement, the sinusoidal model has, for instance, been employed in [23], where the model parameters are iteratively estimated from a noisy observation in the STFT domain, and the enhanced signal is synthesized using (5). In the absence of noise, synthesis results are reported to be almost indistinguishable from the clean speech signal, underlining the capability of (5) to accurately model voiced human speech. In contrast to [23], we now discuss how the sinusoidal model (5) can be employed to directly reconstruct the STFT phase. If the frequency resolution of the STFT is high enough to resolve the harmonic frequencies Ω^h in (5), in each frequency band k only a single harmonic component is dominant. The normalized angular frequency Ω_ℓ^h of the harmonic that dominates frequency band k is denoted as

$$\widetilde{\Omega}_{k,\ell} = \underset{\Omega_\ell^h}{\operatorname{argmin}} \{|2\pi k/N - \Omega_\ell^h|\}, \tag{6}$$

i.e., the harmonic frequency that is closest to the center frequency $2\pi k/N$ of the kth frequency band. Interpreting the STFT of a signal as the output of a complex filter bank subsampled by the hop size R, the spectral phase $\phi_{k,\ell}^{\mathrm{S}}$ changes from segment to segment according to

$$\phi_{k,\ell}^{\mathrm{S}} = \underset{2\pi}{\operatorname{mod}}\left(\phi_{k,\ell-1}^{\mathrm{S}} + \widetilde{\Omega}_{k,\ell} R\right) = \underset{2\pi}{\operatorname{mod}}\left(\phi_{k,\ell-1}^{\mathrm{S}} + \Delta\phi_{k,\ell}^{\mathrm{S}}\right), \tag{7}$$

where the modulo operator $\underset{2\pi}{\operatorname{mod}}(\cdot)$ wraps the phase to values between 0 and 2π.

When the clean signal $s(n)$ is deteriorated by noise, the spectral phases and thus the temporal phase differences $\Delta\phi_{k,\ell}^{\mathrm{S}}$ are deteriorated as well. With an estimate of the fundamental frequency at hand, however, the temporal phase relations in each band can be restored using (7) recursively from segment to segment.

Almost 50 years ago, a similar approach for the propagation of the spectral phase along time was taken in the phase vocoder [5] for time-scaling or pitch-shifting of acoustic signals. The temporal STFT phase difference is modified according to

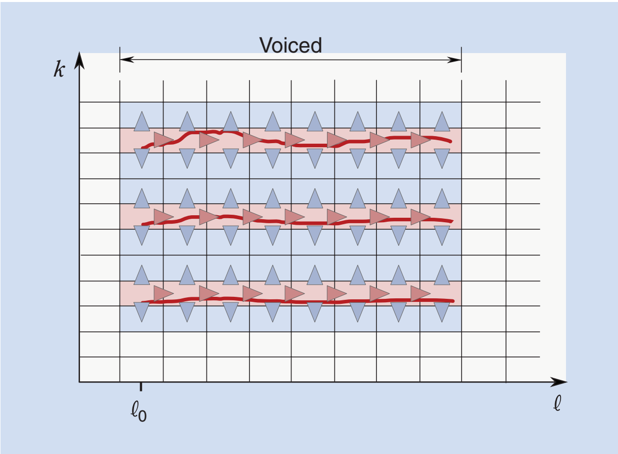
$$\hat{\phi}_{k,\ell}^{\mathrm{S}} = \hat{\phi}_{k,\ell-1}^{\mathrm{S}} + \alpha\Delta\phi_{k,\ell}^{\mathrm{S}}, \tag{8}$$

where in this context, $\Delta\phi_{k,\ell}^{S}$ is often referred to as the IF. By scaling $\Delta\phi_{k,\ell}^{S}$ with the positive real-valued factor α, the IF of the signal component is either increased $(\alpha > 1)$ or decreased $(\alpha < 1)$. Comparing (7) to (8), the phase estimation along time for speech enhancement can be expressed in terms of a phase vocoder with a scaling factor of $\alpha = 1$. However, the application is completely different: instead of deliberately modifying the original phase, the clean speech phase is estimated from a noisy observation. It is worth noting that for the original phase vocoder, in contrast to phase estimation in speech enhancement, no fundamental frequency estimate is needed, as the phase difference $\Delta\phi_{k,\ell}^{\mathrm{S}} = \phi_{k,\ell}^{\mathrm{S}} - \phi_{k,\ell-1}^{\mathrm{S}}$ can be taken directly from the clean original signal.

For an accurate estimation of the clean spectral phase along segments using (7) a proper initialization is necessary [4]. In voiced sounds, the bands between spectral harmonics contain only little signal energy and, in the presence of noise, these bands are likely to be dominated by the noise component, i.e., $\phi_{k,\ell}^{\mathrm{Y}} \approx \phi_{k,\ell}^{\mathrm{N}}$, where $\phi_{k,\ell}^{\mathrm{Y}}$ and $\phi_{k,\ell}^{\mathrm{N}}$ are the spectral phases of the noisy mixture and the noise, respectively. Even though the phase might be set consistent within each band, the spectral relations across frequency bands are distorted already at the initialization stage. Directly applying (7) to every frequency band therefore does not necessarily yield phase estimates that could be employed for phase-based speech enhancement [4].

In the phase vocoder, this problem can be alleviated by aligning phases of neighboring frequency bands relative to each other, which is known as *phase locking*, e.g., [24]. There, the phase is evolved along time only in frequency bands that directly contain harmonic components. The phase in the surrounding bands, which are dominated by the same harmonic, is then set relative to the modified phase. For this, the spectral phase relations of the original signal are imposed on the modified phase spectrum.

In the context of speech enhancement, the same principle has been incorporated to improve the estimation of the clean speech spectral phase [4]. However, since only a noisy signal is observed, the clean speech phase relations across frequency bands are not readily available. To overcome this limitation, again the sinusoidal model is employed. The spectrum of a harmonic signal segment is given by the cyclic convolution of a comb-function with the transfer function of the analysis window, which causes spectral leakage. The spectral leakage induces relations not only between the amplitudes, but also between the phases of neighboring bands. It can be shown that phases of bands that are dominated by the same



[FIG3] Symbolic spectrogram illustrating the sinusoidal model-based phase estimation [4]. Starting from the noisy phase at the onset of a voiced sound in segment ℓ_0, in bands containing harmonic components (red) the phase is estimated along segments. Based on the temporal estimates, the spectral phase of bands between the harmonics (blue) is then inferred across frequency.

harmonic are directly related to each other through the phase response of the analysis window ϕ_k^{W}; see, e.g., [4] for more details. Accordingly, starting from a phase estimate at a band that contains a spectral harmonic, possibly obtained using (7), the phase of the surrounding bands can be inferred by accounting for the phase shift introduced by the analysis window. For this, only the fundamental frequency and the phase response ϕ_k^{W} are required, of which the latter can be obtained offline either from the window's discrete-time Fourier transform (DTFT) or from its DFT with a large amount of zero padding. The complete setup of [4] is illustrated in Figure 3.

THE PHASE OF TRANSIENT SOUNDS IS NOT ONLY RELEVANT FOR DETECTION, BUT ALSO FOR THE REDUCTION OF TRANSIENT NOISE.

It can be argued that for speech enhancement, the phase reconstruction across frequency bands between harmonics is more important than the temporal reconstruction on the harmonics: on the one hand, the local SNR in bands that directly contain harmonics is rather large for many realistic SNR situations, i.e., $\phi_{k,\ell}^{\mathrm{Y}} \approx \phi_{k,\ell}^{\mathrm{S}}$. Thus, the temporal alignment of the harmonic components is maintained rather well in the noisy signal. Further, the noisy phase $\phi_{k,\ell}^{\mathrm{Y}}$ in these bands typically yields a good starting point for the phase reconstruction along frequency. On the other hand, frequency bands between harmonics are likely to be dominated by the noise, i.e., $\phi_{k,\ell}^{\mathrm{Y}} \approx \phi_{k,\ell}^{\mathrm{N}}$, and the clean phase relations across bands are strongly disturbed. Here, the possible benefit of the phase reconstruction is much larger.

Even though the employed model is simple and limited to purely voiced speech sounds, the obtained phase estimates yield valuable information about the clean speech signal that can be employed for advanced speech enhancement algorithms. Interestingly, even the sole enhancement of the spectral phase can lead to a considerable reduction of noise between harmonic components of voiced speech after overlap-add [4]. This is because the speech components of successive segments are adding up constructively after the phase modifications, while the noise components suffer from destructive interference, since the phase relations of the noise have been destroyed. However, speech distortions are also introduced, which are substantially reduced when the estimated phase is combined with an enhanced magnitude, as, e.g., in [25]. Besides its value for signal reconstruction, the estimated phase can also be utilized as additional information for phase-aware magnitude estimation [25] and even for the estimation of clean speech complex coefficients [12], which will be discussed in more detail later.

GROUP DELAY AND TRANSIENT PROCESSING

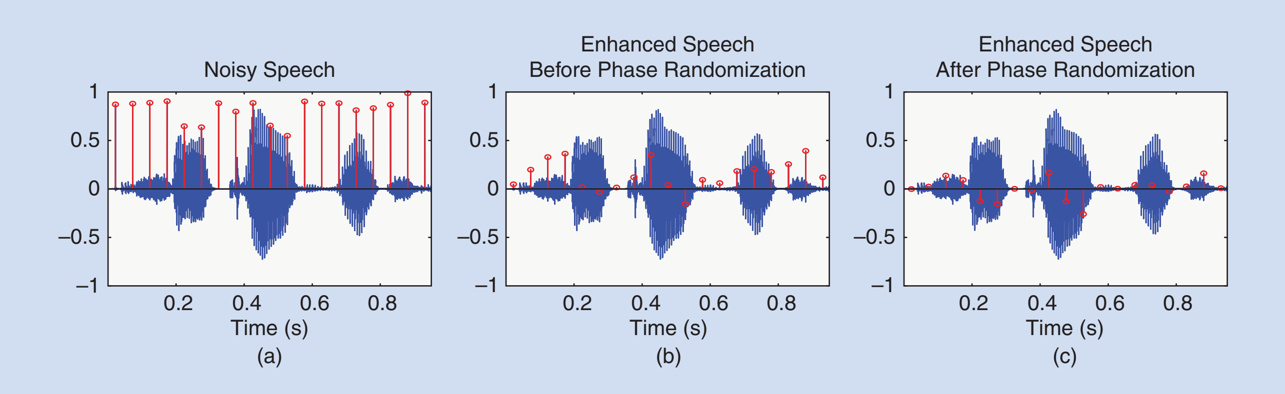
Structures in the phase are not limited to voiced sounds, but are also present for other sounds, like impulses or transients. These structures are well captured by the group delay, which can be seen in Figure 1(c), rendering it a useful representation for phase processing. For example, the group delay has been employed to facilitate clean speech phase estimation in phase-sensitive noise reduction [26]. It can be shown geometrically that if the spectral magnitudes of speech and noise are known, only two possible combinations of phase values remain, both of which perfectly explain the observed spectral coefficients of the mixture. In [26] (and the references therein), Mowlaee and Saedi proposed to solve this ambiguity by choosing the phase combination that minimizes a function of the group delay.

Besides phase estimation, the group delay has successfully been employed for the detection of transients sounds, such as sounds of short duration and speech onsets. To illustrate the role of the phase for transient sounds, let us consider a single impulse as the simplest example. The DFT of such a pulse is $Ae^{-j2\pi\frac{n_0 k}{N}}$, where n_0 is the shift of the peak relative to the beginning of the current segment and A denotes the spectral magnitude. Hence, we observe a linear phase with a constant slope of $-2\pi\,(n_0/N)$. For impulsive signals, we accordingly expect a phase difference across frequency bands that is approximately constant, i.e., a constant group delay. That this is the case also for real speech sounds can be seen in Figure 1(c), where transient sounds show vertical lines with almost equal group delay.

For the detection of impulsive sounds, in [27] a linearity index $\mathrm{LI}_\phi(k)$ is defined, which measures the deviation of the observed phase difference across frequencies to the one that is expected for an impulse at n_0, i.e., $-2\pi(n_0/N)$. The observed phase differences are weighted with the spectral magnitude and averaged over frequency to obtain an estimate of the time domain offset n_0. Only if $\mathrm{LI}_\phi(k)$ is close to zero, i.e., the observed phase fits well to the expected linear phase, an impulsive sound is detected. The detection can be made either at a segment level or for each time-frequency point separately. While the former states if an impulsive sound is present in the current signal segment or not, the latter allows to localize frequency regions that are dominated by an impulsive sound, such as a narrowband onset.

Apart from the group delay, the IF, which corresponds to the temporal derivative of the phase, has also been employed for the detection of transient sounds, e.g., in [28] and the references therein. For steady-state signals, like voiced sounds, the IF is changing only slowly over time, due to the temporal correlation of the overlapping segments. When a transient is encountered, however, the most current segment differs significantly from previous segments and thus the IF also changes abruptly. This can be observed in Figure 1(d), where at speech onsets thin vertical lines appear in the IF deviation. Hence, the change of the IF from segment to segment—and its distribution—allow for the detection of transient sounds, such as note onsets [28].

The phase of transient sounds is not only relevant for detection, but also for the reduction of transient noise. In low SNR time-frequency regions, the observed noisy phase is close to the approximately linear phase of the transient noise. This can lead to artifacts in the enhanced signal if only the spectral magnitude is improved and the noisy phase is used for signal reconstruction: usage of the phase of the transient noise reshapes the enhanced time-domain signal in an uncontrolled way, such that it may again depict an undesired transient behavior. Even for a perfect magnitude estimate, the interfering noise is not perfectly suppressed if the phase



[FIG4] (a) Speech degraded by a click train. (b) Signal obtained by combination of the clean speech spectral magnitude with the noisy phase. (c) Signal after supplemental phase randomization. Samples that contain a click are highlighted in red.

is not processed alongside. To illustrate this, let us consider a speech signal degraded by an impulse train with a period length of T_0, which is nonzero every $N_0 = T_0 f_s$ samples. In Figure 4, the noisy signal (a) is presented together with the result obtained when combining the true clean speech STFT magnitudes with the noisy phase (b). Even though the clean magnitude is employed, which represents the best possible result for phase-blind magnitude enhancement, the time-domain signal still depicts residual impulses, which are caused by the noisy phase. In regions where the enhanced spectral magnitude is close to zero, i.e., in speech absence, the phase is not relevant and the peaks are well suppressed. During speech presence, however, the spectral magnitude is nonzero and the phase becomes important. Accordingly, the residual impulses are most prominent in regions with some speech energy at low local SNRs, where the noisy phase is close to the phase of the impulsive noise.

Recently, Sugiyama and Miyahara proposed the concept of phase randomization to overcome this issue; see, e.g., [27] and references therein. First, time-frequency points that are dominated by speech are identified by finding spectral peaks in the noisy signal. These peaks are excluded from the phase randomization to avoid speech distortions. To further narrow down time-frequency regions where randomization of the spectral phase is sensible, phase-based transient detection can be employed as well [27]. Then, the spectral phase in bins classified as dominated by transient noise is randomized by adding a phase term that is uniformly distributed between $-\pi$ and π. In this way, the approximately linear phase of the dominant noise component is neutralized. The effect of phase randomization is depicted in Figure 4(c), where a perfect magnitude estimate is combined with the modified phase for signal reconstruction. It can be seen that the residual peaks that are present when the noisy phase is employed are strongly attenuated, showing that phase randomization can indeed lead to a considerable increase of noise reduction, especially in low local SNRs. It is interesting to note that while the previously described iterative and sinusoidal model-based approaches aim at estimating the phase of the clean speech signal, the phase randomization approach merely aims at reducing the impact of the phase of the noise on the enhanced speech signal. Although the presented example is just a simple toy experiment, it still highlights the potential of phase randomization toward an improved suppression of transient noise, which has also been observed for real-world impulsive noise, like tapping noise on a touchscreen [27].

RELATION BETWEEN PHASE- AND MAGNITUDE ESTIMATION

So far, we have discussed phase estimation using iterative approaches, sinusoidal model-based approaches, and group delay approaches; we now address the question of how STFT phase estimation can best be employed to improve speech enhancement. The most obvious way to do this is to combine enhanced speech spectral magnitudes in the STFT domain with the estimated or reconstructed STFT phases. It is interesting to note that Wang and Lim [10] already stated that obtaining a more accurate phase estimate than the noisy phase is not worth the effort "if the estimate is used to reconstruct a signal by combining it with an independently estimated magnitude [...]. However, if a significantly different approach is used to exploit the phase information such as using the phase estimate to further improve the magnitude estimate, then a more accurate estimation of phase may be important" [10]. However, at that point it was not clear how a phase estimate could be employed to improve magnitude estimation.

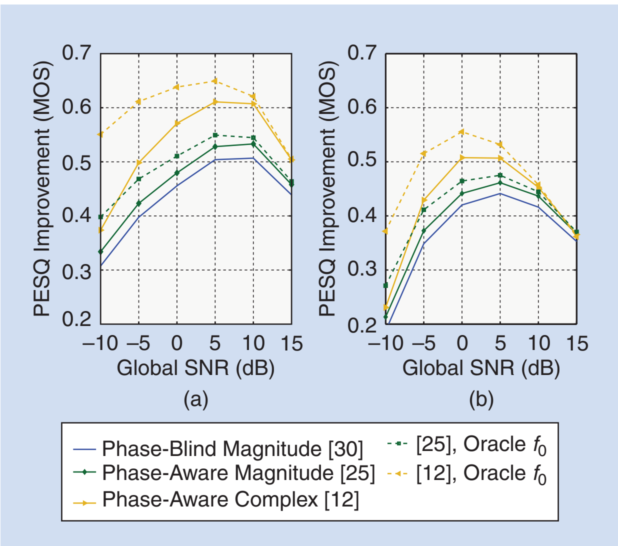
Gerkmann and Krawczyk [25] derived an MMSE estimator of the spectral magnitude when an estimate of the clean speech phase is available, referred to as *phase-sensitive* or *phase-aware* magnitude estimation. They were able to show that the information of the speech spectral phase can be employed to derive an improved magnitude estimator that is capable of reducing noise outliers that are not tracked by the noise PSD estimator. In babble noise, in a blind setup, the PESQ MOS can be improved by 0.25 points in voiced speech at 0 dB input SNR [25]. Further experimental results are given in the following section.

Instead of estimating phase and magnitude separately, one may argue that they should ideally be jointly estimated. The first step in this direction was proposed by Le Roux and Vincent [29] and references therein in the context of Wiener filtering for speech enhancement. As a classical Wiener filter only changes the magnitudes in the STFT domain, the modified spectrum $\widetilde{X}$ is inconsistent, meaning that $\mathrm{STFT}(\mathrm{iSTFT}(\widetilde{X})) \neq \widetilde{X}$. In contrast to this, in [29] the relationship between STFT coefficients across time and frequency is taken into account, leading to the consistent Wiener filter [29], which modifies both the magnitude and the phase of the noisy observation to obtain the separated speech. Wiener filter optimization is formulated as a maximum a posteriori problem under Gaussian assumptions, and a consistency-enforcing term is added either through a hard constraint or a soft penalty. Optimization is respectively performed directly on the signal in the time domain or jointly on phase and magnitude in the complex time-frequency domain, through a conjugate gradient method with a well-chosen preconditioner. Thanks to this joint optimization, the consistent Wiener filter was shown to lead to an improved separation performance compared to the classical Wiener filter and other methods that attempt to use phase information in combination with variance estimates [9], [21], [22], in an oracle scenario as well as in a blind scenario where the speech spectrum is obtained by spectral subtraction from a stationary estimate of the noise spectrum.

To combine phase-sensitive magnitude estimation and iterative approaches, Mowlaee and Saeidi [26] proposed placing the phase-sensitive magnitude estimator into the loop of an iterative approach that enforces consistency. Starting with an initial group-delay-based phase estimate, they proposed to estimate the clean speech spectral magnitude using a phase-sensitive magnitude estimator similar to [25]. After computing the iSTFT and the STFT they reestimated the clean speech phase, and from this reestimate the magnitudes. With this approach, convergence is reached after only few iterations.

WHEN AN INITIAL PHASE ESTIMATE IS ALSO EMPLOYED AS UNCERTAIN PRIOR INFORMATION WHEN IMPROVING THE SPECTRAL PHASE AS PROPOSED IN THE PHASE-AWARE COMPLEX ESTIMATOR CUP, THE PERFORMANCE CAN BE IMPROVED FURTHER.

Another way to jointly estimate magnitudes and phases is to derive a joint MMSE estimator of magnitudes and phases directly in the STFT domain when an uncertain initial phase estimate is available. This phase-aware complex estimator is referred to as the *complex estimator with uncertain phase* (*CUP*) [12]. The initial phase estimate can be obtained by an estimator based on signal characteristics, such as the sinusoidal model-based approach [4]. Using this joint MMSE estimator [12], no STFT iterations are required. The resulting magnitude estimate is a nonlinear tradeoff between a phase-blind and a phase-aware magnitude estimator, while the resulting phase is a tradeoff between the noisy phase and the initial phase estimate. These tradeoffs are controlled by the uncertainty of the initial phase estimate, avoid processing artifacts, and lead to an improvement in predicted speech quality [12]. Experimental results for the CUP estimator are given in the following section.



[FIG5] The PESQ improvement over the noisy input. The results are averaged over four noise types. Evaluated (a) on voiced speech and (b) on the entire signal.

EXPERIMENTAL RESULTS

In this section, we demonstrate the potential of phase processing to improve speech enhancement algorithms. To focus only on the differences due to the incorporation of the spectral phases, we choose algorithms that employ the same statistical models and PSD estimators: for the estimation of the noise PSD we choose the speech presence probability-based estimator with fixed priors (see [6, Sec. 6.3] and references therein) while for the speech PSD we choose the decision-directed approach [7]. We assume a complex Gaussian distribution for the noise STFT coefficients and a heavy-tailed χ-distribution for the speech magnitudes. Furthermore, we use an MMSE estimate of the square root of the magnitudes to incorporate the compressive character of the human auditory system. These models are employed in the phase-blind magnitude estimator [30], the phase-aware magnitude estimator [25], and the phase-aware CUP [12]. We use a sampling rate of 8 kHz and 32 ms spectral analysis windows with 7/8th overlap to facilitate phase estimation. To assess the speech quality, we employ PESQ as an instrumental measure that has been originally proposed for speech coding applications but has been show to correlate with subjective listening tests also for enhanced speech signals. The results are averaged over pink noise modulated at 0.5 Hz, stationary pink noise, babble noise, and factory noise, where the latter three are obtained from the NOISEX-92 database. To have a fair balance between male and female speakers, per noise type, the first 100 male and the first 100 female utterances from dialect region 6 of the Texas Instruments and Massachusetts Institute of Technology (TIMIT) training database are employed. The initial phase estimate is obtained based on a sinusoidal model [4], which only yields a phase estimate in voiced speech. The fundamental frequency is estimated using PEFAC from the voicebox toolkit (http://www.

ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html). Because with [4] we only have a phase estimate in voiced sounds, we show the improvement in voiced segments alongside the overall improvement for entire utterances in Figure 5. When the fundamental frequency estimator detects unvoiced speech segments, the estimators fall back to a phase-blind estimation. Thus, if evaluated over entire signals, the results of the phase-aware estimators will get closer to the phase-blind approaches while the general trends remain.

A PROMISING APPROACH FOR PERFORMANCE IMPROVEMENT IS TO JOIN THE DIFFERENT TYPES OF PHASE PROCESSING APPROACHES, SUCH AS BY INCLUDING MORE EXPLICIT SIGNAL MODELS INTO ITERATIVE PHASE ESTIMATION APPROACHES OR VICE VERSA.

It can be seen that employing phase information to improve magnitude estimation [25] can indeed improve PESQ. The dominant benefit of the phase-aware magnitude estimators is that the phase provides additional information to distinguish between noise outliers and speech. Thus, the stronger the outliers after processing with phase-blind approaches, the larger the potential benefit of phase-aware processing. While here we show the average result over four noise types, a consistent improvement for the tested nonstationary noise types has been observed. While in stationary pink noise the PESQ scores are virtually unchanged, the largest improvements are achieved in babble. This is because babble bursts are often of high energy and may result in large outliers in phase-blind magnitude estimation that can be reduced by exploiting the additional information in the phase.

When an initial phase estimate is also employed as uncertain prior information when improving the spectral phase as proposed in the phase-aware complex estimator CUP [12], the performance can be improved further. The CUP estimator [12] employs the probability of a signal segment being voiced to control the certainty of the initial phase estimate. In unvoiced speech, the uncertainty is largest, effectively resulting in a phase-blind estimator. Therefore, again, we can only expect a PESQ improvement in voiced speech. Compared to phase-blind magnitude estimation [30] in voiced speech and at an input SNR of 0 dB, an improvement in PESQ by 0.12 points is achieved when all parameters are blindly estimated, while 0.18 points are gained with an oracle fundamental frequency. Considering that the improvement of the phase-blind estimator improves PESQ by 0.46 points, the additional improvement of 0.18 points by incorporating phase information in voiced speech is remarkable (factor 1.4), and demonstrates the potential of phase processing for the improvement of speech enhancement algorithms. While the average improvements using phase processing are still moderate, in specific scenarios, e.g., in voiced sounds or impulsive noise, phase processing can help to reduce noise more effectively than using phase-blind approaches. Audio examples can be found at www.speech.uni-oldenburg.de/pasp.html.

FUTURE DIRECTIONS

While the majority of single-channel STFT domain speech enhancement algorithms only address the modification of STFT magnitudes, in this article we reviewed methods that also involve STFT phase modifications. We showed that phase estimation could be done mainly based on models of the signal or by exploiting redundancy in the STFT representation. Examples for model-based algorithms are sinusoidal model-based approaches, and approaches that employ the group delay. By contrast, iterative approaches mainly rely on the spectrotemporal correlations introduced by the redundancy of the STFT representation with overlapping signal segments. While the results of the instrumental evaluations indicate that a sophisticated utilization of phase information can lead to improvements in speech quality, for a conclusive assessment, formal listening tests are required, rendering the subjective evaluation of particularly promising phase-aware algorithms a necessity for future research.

Despite recent advances, there are still many open issues in phase processing. For instance, similar to magnitude estimation, phase estimation is still difficult in very low SNRs. A promising approach for performance improvement is to join the different types of phase processing approaches, such as by including more explicit signal models into iterative phase estimation approaches or vice versa. A first step in this direction is presented in [26]. As another example, while the consistent Wiener filter only exploits the phase structure of the STFT representation, an exciting challenge going forward is to integrate models of the phase structure of the signal itself into a joint optimization framework.

Modern machine-learning approaches such as deep neural networks, which have proven to be very successful in improving speech recognition performance, have recently been shown to lead to state-of-the-art performance for speech enhancement using a magnitude-based approach. The natural next step is to extend their use to phase estimation to further improve performance. On top of the fact that they are data driven, which reduces the necessity for modeling assumptions that may be inaccurate, a great advantage of such methods over the iterative approaches for phase estimation presented here or approaches based on nonnegative matrix factorization or Gaussian mixture models, is that they can typically be efficiently evaluated at test time.

Indeed, striving for fast, lightweight algorithms is critical in the context of assisted listening and speech communication devices, where special requirements with respect to complexity and latency persist. While more and more computational power will be available with improved technology, for economic reasons as well as to limit power consumption, it is always of interest to keep the complexity as low as possible. Thus, more research in reducing complexity remains of interest. Complexity reduction could be obtained, for instance, by decreasing the overlap of the STFT analysis, but its impact on performance of phase estimation algorithms is not well studied. On the other hand, the lower bound on the latency of the algorithms is dominated by the window lengths in

STFT analysis and synthesis. Further research could therefore also address phase estimation using low latency filter banks.

After many years in the shadow of magnitude-centric speech enhancement, phase-aware signal processing is now burgeoning and expanding quickly: with still many aspects to explore, it is an exciting area of research that is likely to lead to important breakthroughs and push speech processing forward. Supplemental material and further references can be found at www.speech.uni-oldenburg.de/pasp.html.

ACKNOWLEDGMENT

This work was supported by grant GE2538/2-1 of the German Research Foundation.

AUTHORS

Timo Gerkmann (timo.gerkmann@uni-oldenburg.de) received his Dipl.-Ing. and Dr.-Ing. degrees in electrical engineering and information technology in 2004 and 2010 from the Ruhr-Universität Bochum, Germany. In 2005, he spent six months with Siemens Corporate Research in Princeton, New Jersey, United States. From 2010 to 2011, he was a postdoctoral researcher at the Royal Institute of Technology, Stockholm, Sweden. Since 2011, he has been a professor for speech signal processing at the University of Oldenburg, Germany. His main research interests are digital speech and audio processing, including speech enhancement, dereverberation, modeling of speech signals, speech recognition, and hearing devices.

Martin Krawczyk-Becker (martin.krawczyk-becker@uni-oldenburg.de) studied electrical engineering and information technology at the Ruhr-Universität Bochum, Germany. His major was communication technology with a focus on audio processing, and he received his Dipl.-Ing. degree in August 2011. From January 2010 to July 2010, he was with Siemens Corporate Research in Princeton, New Jersey, United States. Since November 2011, he has been pursuing his Ph.D. degree in the field of speech enhancement and noise reduction at the University of Oldenburg, Germany.

Jonathan Le Roux (leroux@merl.com) completed his B.Sc. and M.Sc. degrees in mathematics at the Ecole Normale Supérieure, Paris, France, and his Ph.D. degree at the University of Tokyo, Japan, and the Université Pierre et Marie Curie, Paris, France. He is a principal research scientist at Mitsubishi Electric Research Laboratories in Cambridge, Massachusetts, United States, and was previously a postdoctoral researcher at Nippon Telegraph and Telephone Communication Science Laboratories. His research interests are in signal processing and machine learning applied to speech and audio. He is a Senior Member of the IEEE and a member of the IEEE Audio and Acoustic Signal Processing Technical Committee.

REFERENCES

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

[2] B. Yegnanarayana and H. Murthy, "Significance of group delay functions in spectrum estimation," *IEEE Trans. Signal Processing*, vol. 40, no. 9, pp. 2281–2289, Sept. 1992.

[3] A. P. Stark and K. K. Paliwal, "Speech analysis using instantaneous frequency deviation," in *Proc. ISCA Interspeech*, 2008, pp. 2602–2605.

[4] M. Krawczyk and T. Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Processing*, vol. 22, no. 12, pp. 1931–1940, Dec. 2014.

[5] J. L. Flanagan and R. M. Golden, "Phase vocoder," *Bell Syst. Tech. J.*, vol. 45, no. 9, pp. 1493–1509, 1966.

[6] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement: A Survey of the State-of-the-art*. San Rafael, CA: Morgan & Claypool, Feb. 2013.

[7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[8] T. F. Quatieri, "Phase estimation with application to speech analysis-synthesis," Ph.D. dissertation, Massachusetts Inst. Technol., 1979.

[9] N. Sturmel and L. Daudet, "Iterative phase reconstruction of Wiener filtered signals," in *Proc. ICASSP*, Mar. 2012, pp. 101–104.

[10] D. L. Wang and J. S. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 30, no. 4, pp. 679–681, 1982.

[11] P. Vary, "Noise suppression by spectral magnitude estimation – mechanism and theoretical limits," *Elsevier Signal Process.*, vol. 8, pp. 387–400, May 1985.

[12] T. Gerkmann, "Bayesian estimation of clean speech spectral coefficients given a priori knowledge of the phase," *IEEE Trans. Signal Processing*, vol. 62, no. 16, pp. 4199–4208, Aug. 2014.

[13] J. R. Hershey, S. J. Rennie, and J. Le Roux, "Factorial models for noise robust speech recognition," in *Techniques for Noise Robustness in Automatic Speech Recognition*, T. Virtanen, R. Singh, and B. Raj, Eds. Hoboken, NJ: Wiley, 2012, ch. 12.

[14] M. Kazama, S. Gotoh, M. Tohyama, and T. Houtgast, "On the significance of phase in the short term Fourier spectrum for speech intelligibility." *J. Acoust. Soc. Amer.*, vol. 127, no. 3, pp. 1432–1439, Mar. 2010.

[15] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *Elsevier Speech Commun.*, vol. 53, no. 4, pp. 465–494, Apr. 2011.

[16] D. W. Griffin, "Signal estimation from modified short-time Fourier transform magnitude," Master's thesis, *Dept. Electr. Eng. and Computer Sci.*, Massachusetts Inst. Technol., Dec. 1983.

[17] X. Zhu, G. T. Beauregard, and L. L. Wyse, "Real-time signal estimation from modified short-time Fourier transform magnitude spectra," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 15, no. 5, pp. 1645–1653, July 2007.

[18] J. Le Roux, N. Ono, and S. Sagayama, "Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction," in *Proc. ISCA Workshop Statistical Perceptual Audition (SAPA)*, Sept. 2008, pp. 23–28.

[19] J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama, "Phase initialization schemes for faster spectrogram-consistency-based signal reconstruction," in *Proc. Acoustical Society Japan Autumn Meeting*, paper no. 3-10-3, Sept. 2010.

[20] V. Gnann and M. Spiertz, "Improving RTISI phase estimation with energy order and phase unwrapping," in *Proc. Int. Conf. Digital Audio Effects (DAFx)*, Sept. 2010.

[21] D. Gunawan and D. Sen, "Iterative phase estimation for the synthesis of separated sources from single-channel mixtures," *IEEE Signal Process. Lett.*, vol. 17, no. 5, pp. 421–424, May 2010.

[22] N. Sturmel and L. Daudet, "Informed source separation using iterative reconstruction," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 21, no. 1, pp. 178–185, Jan. 2013.

[23] J. Jensen and J. H. Hansen, "Speech enhancement using a constrained iterative sinusoidal model," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 7, pp. 731–740, Oct. 2001.

[24] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Trans. Speech Audio Processing*, vol. 7, no. 3, pp. 323–332, May 1999.

[25] T. Gerkmann and M. Krawczyk, "MMSE-optimal spectral amplitude estimation given the STFT-phase," *IEEE Signal Process. Lett.*, vol. 20, no. 2, pp. 129–132, Feb. 2013.

[26] P. Mowlaee and R. Saeidi, "Iterative closed-loop phase-aware single-channel speech enhancement," *IEEE Signal Process. Lett.*, vol. 20, no. 12, pp. 1235–1239, Dec. 2013.

[27] A. Sugiyama and R. Miyahara, "Tapping-noise suppression with magnitude-weighted phase-based detection," in *Proc. IEEE WASPAA*, Oct 2013, pp. 1–4.

[28] J. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 1035–1047, Sept. 2005.

[29] J. Le Roux and E. Vincent, "Consistent Wiener filtering for audio source separation," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 217–220, Mar. 2013.

[30] C. Breithaupt, M. Krawczyk, and R. Martin, "Parameterized MMSE spectral magnitude estimation for the enhancement of noisy speech," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, Las Vegas, NV, Apr. 2008, pp. 4037–4040.

[SP]