Speech Communication 99 (2018) 144–160

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Single-channel speech enhancement using inter-component phase relations

Siarhei Y. Barysenka[a], Vasili I. Vorobiov[a], Pejman Mowlaee[*],[b],[c]

[a] *Belarusian State University of Informatics and Radioelectonics, Minsk, Belarus*
[b] *Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria*
[c] *Widex A/S, Nymøllevej 6, 3540 Lynge, Denmark*

ARTICLE INFO

Keywords:
Phase estimation
Phase invariance
Bi-phase
Speech enhancement
Speech quality
Speech intelligibility

ABSTRACT

Phase-aware processing has recently attracted lots of interest among researchers in speech signal processing field as successful results have been reported for various applications including automatic speech/speaker recognition, noise reduction, anti-spoofing and speech synthesis. In all these applications, the success of the applied phase-aware processing method is predominantly affected by the robustness and the accuracy of the provided estimate of the clean spectral phase to be obtained from noisy observation. Therefore, in this paper, we first consider the inter-component phase relations of poly-harmonic signals as speech captured by Phase Invariance, Phase Quasi-Invariance and Bi-Phase constraints. Then, relying on these constraints between harmonics as phase structure, we propose phase estimators. Throughout various experiments we demonstrate the usefulness of the newly proposed methods. We further report the achievable speech enhancement performance by the proposed phase estimators and compare them with the benchmark methods in terms of perceived quality, speech intelligibility and phase estimation accuracy. The proposed methods show improved performance averaged over different noise scenarios and signal-to-noise ratios.

1. Introduction

In many signal processing applications including radar, image and speech processing, the problem of interest is to detect the desired signal in a noisy observation. While many previous studies were dedicated to deriving new estimators for amplitude and frequency of signal components (harmonics) (Kay, 1993; Van Trees, 2004), the estimation of spectral phase has been less addressed.

In speech signal processing, the processing of spectral phase was historically reported perceptually unimportant follow up the early experiments by Wang and Lim (1982) and Vary (1985). In particular, Vary reported that human perceives phase distortion only below signal-to-noise ratio (SNR) of 6 dB, hence noisy spectral phase suffices for high enough SNRs. Later on, Aarabi (2006) and Alsteris and Paliwal (2007) reported that spectral phase could be helpful for speech applications including automatic speech recognition and noise reduction. More recently, overview on phase-aware signal processing for speech applications thoroughly demonstrated the advantages and potential of incorporating phase processing (Mowlaee et al., 2016a; Gerkmann et al., 2015; Mowlaee et al., 2016b).

The reasons why the research on phase-aware processing or in general studying the phase importance in speech applications was slow could be explained in following: (i) historically, the spectral phase of speech signals was believed to be unimportant as reported in the early studies (for a full review we refer to Mowlaee et al. (2016a, Ch. 1)), (ii) in contrast to the magnitude spectrum, the phase wrapping prevents an accessible pattern of phase spectrum in the Fourier domain which complicates the phase analysis of the given speech signal (Mowlaee et al., 2016b), (iii) phase processing is computationally complex and requires sophisticated algorithms with accurate prior statistics or fundamental frequency estimate (see e.g. Mowlaee and Kulmer, 2015b), (iv) little or no attention has been dedicated to the relations between harmonic components in speech, hence, the phase of harmonics has been estimated independently or relying on the phase of the fundamental harmonic.

It is important to note that an enhanced spectral phase obtained from noisy speech observation can be used directly for signal reconstruction and hence to enhance the noisy speech signal. Furthermore, an estimated clean spectral phase can also be used to derive improved spectral amplitude estimators in an iterative (Mowlaee et al., 2017; Mowlaee and Saeidi, 2013) or non-iterative (Gerkmann et al., 2015; Krawczyk and Gerkmann, 2016) configuration[1]. As the achievable improvement from a phase-aware processing framework is limited by the accuracy of the spectral phase estimator stage, therefore, a challenging research topic is to find novel approaches that provide accurate and robust estimators of the clean spectral phase from a noisy observation. The achievement of a robust and accurate spectral phase information opens up opportunities for further improved performance in other speech applications including automatic speech recognition

[*] Corresponding author at: Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria.
E-mail addresses: siarhei.barysenka@gmail.com (S.Y. Barysenka), viv314@gmail.com (V.I. Vorobiov), pejman.mowlaee@tugraz.at (P. Mowlaee).

[1] For a full review on phase-aware speech enhancement we refer to (Mowlaee et al., 2016a, Ch. 4).

https://doi.org/10.1016/j.specom.2018.03.009
Received 18 September 2017; Received in revised form 13 February 2018; Accepted 24 March 2018
Available online 28 March 2018

(Fahringer et al., 2016), speech synthesis (Espic et al., 2017), source separation (Mayer et al., 2017) and emotion recognition (Deng et al., 2016).

The previous attempts for spectral phase estimation can be divided into the following groups (Chacon and Mowlaee, 2014)[2]: (i) Griffin–Lim (GL) (Griffin and Lim, 1984) based methods which apply consistency of the short-time Fourier transform (STFT) spectrogram and iteratively reconstruct the spectral phase from an initial estimate of the spectral magnitude (see Mowlaee and Watanabe, 2013 for an overview), (ii) model-based short-time Fourier transform phase improvement (STFTPI) (Krawczyk and Gerkmann, 2014) relying on a harmonic model to predict the spectral phase across time using phase vocoder principle and across frequency by compensating for the analysis window phase response. Another model-based phase estimator is the geometry-based approach where additional time-frequency constraint (Mowlaee and Saeidi, 2014) is used to remove the ambiguity in the chosen spectral phase pairs. Three types of constraints were proposed in the geometry-based phase estimator: group delay deviation, instantaneous frequency deviation and relative phase shift (RPS) (Saratxaga et al., 2009). As another model-based approach, time-frequency smoothing of unwrapped harmonic phase was proposed by applying the harmonic model plus phase decomposition (Degottex and Erro, 2014b) followed by smoothing filter (Kulmer and Mowlaee, 2015b; Mowlaee and Kulmer, 2015b; 2015a), and (iii) statistical methods: maximum a posteriori harmonic (MAP) (Kulmer and Mowlaee, 2015a; Mowlaee et al., 2017), temporal smoothing of the unwrapped harmonic phase (TSUP) (Kulmer and Mowlaee, 2015b; Kulmer et al., 2014) and least-squares (LS) (Chacon and Mowlaee, 2014).

In all previous phase estimators, the underlying relation between harmonics phase or phase structure across harmonics is either not directly taken into account (Krawczyk and Gerkmann, 2014; Kulmer and Mowlaee, 2015a,b) or only relies on the phase of the fundamental frequency used as the reference (Mowlaee and Saeidi, 2014; Mowlaee and Kulmer, 2015a,b). For example, in geometry-based phase estimator with RPS constraint (Mowlaee and Saeidi, 2014) the relation between the harmonic phases with the fundamental frequency phase is taken into account. Also, smoothing across RPS has been considered in Mowlaee and Kulmer (2015b). The phase estimation performance relies on the accuracy of the fundamental frequency phase which relies itself on the fundamental frequency estimation accuracy. This limits the performance for low-frequency noise scenarios. Furthermore, the underlying phase structure across harmonics is not taken into account, therefore, the harmonic phases are estimated independently.

In this paper, we argue that the two aforementioned issues: (i) relying on the fundamental frequency phase, and (ii) neglecting the phase structure across harmonics in speech signal limit the achievable performance by the existing spectral phase estimators. Therefore, in this paper, we propose new phase estimators that rely on the **inter-component phase relations (ICPR)** for a polyharmonic signal like speech. In our earlier publication (Pirolt et al., 2017), we reported preliminary results on the usefulness of applying phase quasi-invariant constraint for phase estimation and speech enhancement. In this paper, we present the ICPR in details for a polyharmonic signal (here speech) and report their usefulness in speech enhancement for different noise scenarios. The three phase relations are: Phase Invariance (PI), Phase Quasi-Invariance (PQI), and Bi-Phase (see Section 2 for an overview). We will apply these phase relations as constraints to derive the harmonic phase estimators. The so-derived estimators are then applied for speech enhancement whereby a phase-enhanced speech signal is provided. Throughout the experiments, we demonstrate that the newly derived phase estimators result in improved perceived quality and speech intelligibility and a lower phase estimator error versus the benchmark methods.

The rest of the paper is organized as follows. Section 2 presents some background on the ICPR for polyharmonic signals in general. In particular, we will focus on three phase relations: Phase Invariance, Phase Quasi-Invariance and Bi-Phase. In Section 3, we propose details on the proposed phase estimators relying on each of the three constraints (PI, PQI and Bi-Phase). Section 4 presents proof-of-concept experiments and speech enhancement results. A comparative study of phase estimation performance is presented by comparing the achievable speech enhancement results versus the relevant benchmark methods followed up by discussions. Section 5 concludes on the work.

2. Background on inter-component phase relations in polyharmonic signals

In this section, we review the theory and applications of phase processing techniques that exploit the following underlying principle: the parameters of particular harmonic are considered in relation to parameters of other harmonics of the same oscillation process. This principle provides a basis for a number of inter-component phase processing methods and reveals the special properties of signals, that are failed to be observed by conventional magnitude and power spectrum analysis methods. Additionally, inter-component phase measurements are less sensitive to noise and signal magnitude variations in comparison with magnitude measurements. Since the natural speech is originated by a single material system represented by human vocal tract, an investigation into the impact of the inter-component relations (including the phase ones) in speech signal could be a promising research direction.

2.1. Phase invariant

The first attempt (to our knowledge based on the literature research) of exploiting the aforementioned principle originates back in 1953, when Zverev formulated and described the notion of phase invariance for a modulated oscillation (Zverev, 1953). We consider such oscillation $u(t)$ in Zverev notation from Zverev (1956):

$$u(t) = B(t)\sin(\omega_0 t + \Phi(t)) = A_0\sin(\omega_0 t - \phi_0) + A_1\sin(\omega_1 t - \phi_1) + A_2\sin(\omega_2 t - \phi_2). \tag{1}$$

During the experiments in ultrasonic dispersion measurements of acoustic waves, it was noted that for an oscillation (1) the special combination Θ of initial phase values remains invariant to the time coordinate:

$$\Theta = \phi_0 - \frac{\phi_1 + \phi_2}{2}. \tag{2}$$

The combination Θ was called *Phase Invariant*. The notion of phase invariance was successfully applied later in hydrodynamics by Tatarskii (2004), non-linear acoustics by Gavrilov (2009), radio-wave propagation by Galayev and Kivva (2009), and briefly discussed by Vorobiov and Barysenka (2014) with a proof-of-concept experiment for rotary machines vibration analysis.

To our knowledge, there are no known attempts to apply the relation of Phase Invariant in speech processing. In order to discuss the points that motivated us to do the research regarding the benefits of such application, we consider a polyharmonic speech signal $s(t)$ with time index t consisting of H_t harmonics. Given the fundamental frequency $F_0(t)$, each of the harmonics is characterized by the harmonic index $h \in [1, H_t]$ and the corresponding amplitude $A(h, t)$ and phase $\Phi(h, t)$, both assumed to be slowly varying in time:

$$\begin{aligned} s(t) &= \sum_{h=1}^{H_t} s(h, t) = \sum_{h=1}^{H_t} A(h, t)\cos\Psi(h, t) \\ &= \sum_{h=1}^{H_t} A(h, t)\cos(2\pi h F_0(t) t + \Phi(h, t)). \end{aligned} \tag{3}$$

Unlike a signal of form (1) studied in Zverev (1953), Zverev (1956), Tatarskii (2004), Gavrilov (2009), Galayev and Kivva (2009), Vorobiov and Barysenka (2014), a speech signal (3) contains more than

[2] For a full review on spectral phase estimation methods we refer to (Mowlaee et al., 2016a, Ch. 3).

three components ($H_t > 3$). Thus, we consider a signal (3) consisting of triplets of the following configurations:

$$\begin{cases} f_1 = K_1 F_0, & \text{where } K_1 = 1, 2, \ldots \\ f_2 = K_2 F_0, & \text{where } K_2 = K_1 + 1, K_1 + 2, \ldots \\ f_3 = K_3 F_0, & \text{where } K_3 = 2K_2 - K_1. \end{cases} \tag{4}$$

This consideration allows us to define a constraint similar to original Phase Invariant (2) on an arbitrary triplet that satisfy equations (4) of a speech signal (3):

$$PI(K_1, K_2, K_3, t) = \frac{\Phi(K_1, t) + \Phi(K_3, t)}{2} - \Phi(K_2, t). \tag{5}$$

In all further notation, we refer to constraint (5) as *Phase Invariant* (PI) to admit its origin from the Zverev Eq. (2). For the speech processing application, an important feature to note is that PI may be calculated based on unwrapped $\Psi(h, t)$ functions, since the dependency on fundamental frequency $F_0(t)$ is ignored during the algebraic manipulation on Eqs. (3) and (5):

$$\begin{aligned} PI(K_1, K_2, K_3, t) &= \frac{\Psi(K_1, t) + \Psi(K_3, t)}{2} - \Psi(K_2, t) \\ &= \frac{\Phi(K_1, t) + \Phi(K_3, t)}{2} - \Phi(K_2, t). \end{aligned} \tag{6}$$

In the following, the phase unwrapping problem is resolved using the one-dimensional phase unwrapping algorithm (Itoh, 1982).

2.2. Bi-Phase

The notion of Bi-Phase is known from the theory of higher-order spectra (Nikias and Mendel, 1993). The higher-order spectra retains the phase information in contrast with the power spectrum estimation, where the phase relations between the harmonic components are ignored.

For a finite energy real deterministic signal $x(n)$, $n \in \mathbb{Z}$ a particular case of higher-order spectrum called *third-order spectrum* (also referred to as *bi-spectrum*) is defined as follows:

$$M_3^x(\omega_1, \omega_2) = X(\omega_1)X(\omega_2)X^*(\omega_1 + \omega_2), \tag{7}$$

where $X(\omega)$ and $X^*(\omega)$ denote a Fourier transform of $x(n)$ and its complex conjugation, respectively. In physical meaning, the bi-spectrum describes the correlations between the components with harmonically related set of frequencies $\{\omega_1, \omega_2, \omega_1 + \omega_2\}$.

For an overview of the existing applications of bi-spectrum in signal processing we refer to Nikias and Mendel (1993). Whereas the higher-order spectra techniques have a long-lasting history of successful applications in radar, telecommunications and image processing (Bochkov and Gorokhov, 1995; Totsky et al., 2014), considerably less attention has been engaged to speech processing. Boyanov et al. (1990, 1991) summarized the properties and features of speech signal that are potentially discoverable by means of bi-spectral analysis in laryngeal pathology detection and speaker recognition applications. A few experiments were conducted by Fulchiero and Spanias (1993) to incorporate bi-spectrum for speech enhancement in Gaussian noise, where the positive results of enhancement were reported for SNR values less than 6 dB. Nevertheless, the authors concluded that invention of more reliable estimators may improve the performance of proposed bi-spectrum enhancer. Other works (Wells, 1985; Seetharaman and Jernigan, 1988; Azarov et al., 2011) outlined the possible benefits of bi-spectrum application for voiced/unvoiced decision, signal reconstruction and speaker identification. Therefore, the higher-order spectra estimation for speech processing applications remains a promising topic for thorough research.

In this work, we focus on phase component of bi-spectrum (namely, *Bi-Phase*) for particular triplets in a speech signal (3). The considered triplet frequencies satisfy the following condition of harmonically related frequencies:

$$\begin{cases} f_1 = K_1 F_0, & \text{where } K_1 = 1, 2, \ldots \\ f_2 = K_2 F_0, & \text{where } K_2 = K_1 + 1, K_1 + 2, \ldots \\ f_3 = K_3 F_0, & \text{where } K_3 = K_1 + K_2. \end{cases} \tag{8}$$

The Bi-Phase of bi-spectrum (7) is determined as $\angle M_3^x(\omega_1, \omega_2)$, and for harmonically related triplet (8) is given by:

$$BiPh(K_1, K_2, K_3, t) = \Phi(K_1, t) + \Phi(K_2, t) - \Phi(K_3, t). \tag{9}$$

Similar to PI, the Bi-Phase constraint (9) may be calculated based on unwrapped $\Psi(h, t)$ functions, since fundamental frequency items are subsequently discarded:

$$\begin{aligned} BiPh(K_1, K_2, K_3, t) &= \Psi(K_1, t) + \Psi(K_2, t) - \Psi(K_3, t) \\ &= \Phi(K_1, t) + \Phi(K_2, t) - \Phi(K_3, t). \end{aligned} \tag{10}$$

The Phase Distortion (PD) constraint that has been employed for the estimation of glottal model parameters and other speech processing applications (Degottex and Erro, 2014b) represents a special case of Bi-Phase:

$$PD(h, t) = -BiPh(1, h, h + 1, t). \tag{11}$$

2.3. Phase quasi-invariant

The idea of measuring the phase shift between two harmonics with frequencies $\{f_0, K f_0\}$ (where K is a positive integer, $K \neq 1$) originates from a problem of radar object recognition (Sletten et al., 1973), followed by discussion in G. (1994) and Vorobiov and Klimov (1986). Later on, a few experiments on phase measurement of multiple frequency signals in sonar were conducted by Vorobiov et al. (1986). Subsequently, the results of those experiments were theoretically generalized by Aksionov et al. for frequencies $\{K_1 f_0, K_2 f_0\}$ (where K_1 and K_2 are positive integers, $K_1 < K_2$) as an approach of phase analysis of ultra wide band digital signals (Aksionov et al., 1994). Finally, the aforementioned approach was applied to speech analysis by Vorobiov (2006), Vorobiov et al. (2006), Vorobiev and Davydov (2008) and Vorobiov and Davydov (2012), where it was first referred to the proposed phase relation as *Phase Quasi-Inavariant* (PQI).

Considering Eq. (3), the PQI constraint is defined for components with frequencies $\bar{h}F_0(t)$ and $hF_0(t)$, where $\bar{h} < h$, and given by:

$$PQI(\bar{h}, h, t) = \Psi(\bar{h}, t) - \frac{\Psi(h, t) \cdot \bar{h}}{h} = \Phi(\bar{h}, t) - \frac{\Phi(h, t) \cdot \bar{h}}{h}. \tag{12}$$

A vector representation of PQI as a combination of $PQI(\bar{h}, h, t)$ constraints for different pairs $\{\bar{h}, h\}$ is potentially useful for the multi-dimensional analysis of features contained in a speech signal (Vorobiov, 2006).

The RPS constraint (Saratxaga et al., 2009) that is used in phase processing of speech signals (Degottex and Erro, 2014b), (Mowlaee et al., 2016a, Ch. 2), represents a special case of PQI (Pirolt et al., 2017):

$$RPS(h, t) = -hPQI(1, h, t). \tag{13}$$

It is important to note that RPS defines the phase relation solely between the fundamental frequency $\bar{h} = 1$ and its higher harmonics, whereas PQI constraint does not limit $\bar{h}$ to any particular harmonic number.

2.4. Unambiguous definition range

In all previous equations, $\Phi(h, t)$ is a multi-valued function of any arbitrary harmonic h:

$$\Phi(h, t) = Princ\{\Phi(h, t)\} + 2\pi N, \tag{14}$$

where $N \in \mathbb{Z}$ and $Princ\{\Phi(h, t)\}$ denotes the principal value of $\Phi(h, t)$ in the interval $[-\pi, \pi)$. In other words, any estimate of $\Phi(h, t)$ can be given

Table 1
Relations between PI, Bi-Phase and PQI.

PQI configuration	Phase invariant, $PI(K_1, K_2, K_3, t)$	Bi-phase, $BiPh(K_1, K_2, K_3, t)$
$PQI(K_1, K_2, t)$ $PQI(K_2, K_3, t)$	$\frac{1}{2}PQI(K_1, K_2, t) - \frac{K_3}{2K_2}PQI(K_2, K_3, t)$	$PQI(K_1, K_2, t) + \frac{K_3}{K_2}PQI(K_2, K_3, t)$
$PQI(K_1, K_2, t)$ $PQI(K_1, K_3, t)$	$\frac{K_2}{K_1}PQI(K_1, K_2, t) - \frac{K_3}{2K_1}PQI(K_1, K_3, t)$	$-\frac{K_2}{K_1}PQI(K_1, K_2, t) + \frac{K_3}{K_1}PQI(K_1, K_3, t)$
$PQI(K_1, K_3, t)$ $PQI(K_2, K_3, t)$	$\frac{1}{2}PQI(K_1, K_3, t) - PQI(K_2, K_3, t)$	$PQI(K_1, K_3, t) + PQI(K_2, K_3, t)$

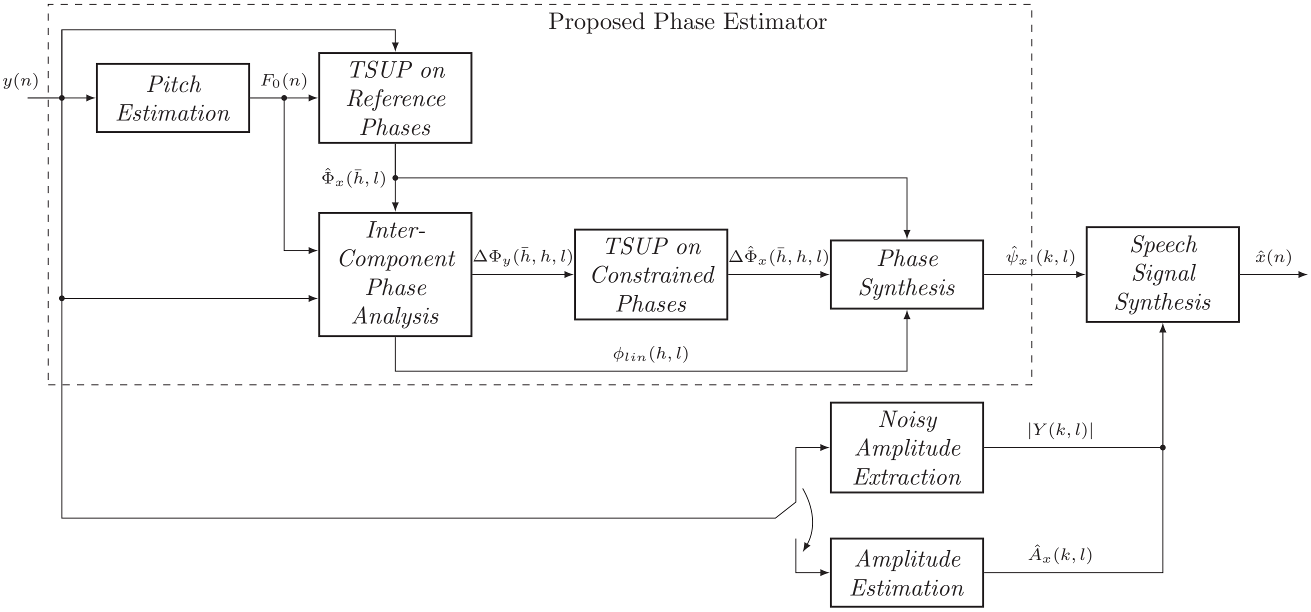


Fig. 1. Block diagram showing the proposed phase estimation framework for speech enhancement.

up to an additive constant divisible by 2π. This fact has to be considered when exploiting the constraints defined in Sections 2.1, 2.2 and 2.3.

Let us consider the case of PQI. From Eq. (12) we have:

$$\begin{aligned} PQI(\bar{h}, h, t) &= Princ\{\Phi(\bar{h}, t)\} + 2\pi N_1 - \frac{\bar{h}}{h}\cdot(Princ\{\Phi(h, t)\} + 2\pi N_2) \\ &= Princ\left\{\Phi(\bar{h}, t) - \frac{\Phi(h, t)\cdot\bar{h}}{h}\right\} + \underbrace{2\pi N_1}_{\Delta_1} - \underbrace{\frac{2\pi\bar{h}}{h}N_2}_{\Delta_2}, \end{aligned} \quad (15)$$

where $N_1, N_2 \in \mathbb{Z}$. In comparison with Eq. (14), the decrease of PQI range is influenced by $|\Delta_2|$ rather than $|\Delta_1|$, since $|\Delta_2| < |\Delta_1|$ for any $N_1 = N_2$. Therefore, Δ_1 can be ignored, as a result Eq. (15) reduces into the following form:

$$PQI(\bar{h}, h, t) = Princ\left\{\Phi(\bar{h}, t) - \frac{\Phi(h, t)\cdot\bar{h}}{h}\right\} + \frac{2\pi\bar{h}}{h}N, \quad (16)$$

The range $\left[\frac{-\pi\bar{h}}{h}, \frac{\pi\bar{h}}{h}\right)$ is called *unambiguous definition range* of $PQI(\bar{h}, h, t)$. It defines the range of principal values of PQI for any given combination of $\bar{h}$ and h.

A similar reasoning is applied to determine the unambiguous definition range of PI and Bi-Phase. It can be shown that in case of PI the range is $[\frac{-\pi}{2}, \frac{\pi}{2})$, and in case of Bi-Phase the range is $[-\pi, \pi)$ regardless of the configuration of harmonic numbers for both constraints.

2.5. Relations between phase invariant, Bi-phase and phase quasi-invariant

The relations between PI, Bi-Phase and PQI are given in Table 1. The indices K_1, K_2, K_3 denote the left-most, center and the right-most harmonic numbers, respectively, in considered PI in Eq. (4) or Bi-Phase in Eq. (8) constraints.

In a special case of Bi-Phase configuration where $K_1 = K_2$, a triplet of frequencies (8) confluents to a pair $\{\bar{h}, h\}$, where $h = 2\bar{h}$, therefore reducing the Bi-Phase to PQI:

$$\begin{aligned} BiPh(\bar{h}, \bar{h}, 2\bar{h}, t) &= \Phi(\bar{h}, t) + \Phi(\bar{h}, t) - \Phi(2\bar{h}, t) \\ &= 2\cdot\left(\Phi(\bar{h}, t) - \frac{\Phi(2\bar{h}, t)\cdot\bar{h}}{2\bar{h}}\right) = 2\cdot PQI(\bar{h}, 2\bar{h}, t). \end{aligned} \quad (17)$$

3. Proposed phase estimators

In this section, we present the proposed phase estimators relying on ICPR[3] Fig. 1 shows the block diagram of the speech enhancement setup that uses the proposed phase estimation framework.

Let $x(n)$ and $y(n)$ denote the clean and noisy signal, respectively, in time domain. The noisy signal $y(n) = x(n) + \nu(n)$ represents a mixture of the clean signal and the noise $\nu(n)$. The harmonic amplitudes of clean and noisy signal are referred to as $A_x(h, l)$ and $A_y(h, l)$, respectively, where h is the harmonic index and l denotes the time frame. Let $X(k, l)$ and $Y(k, l)$ denote the STFT of clean and noisy signal, respectively, at frequency bin k and time frame l. The spectral amplitudes of clean and noisy signal are referred to as $|X(k, l)|$ and $|Y(k, l)|$, respectively. The instantaneous STFT phases of clean and noisy signal are denoted by

[3] While in the preliminary work (Pirolt et al., 2017) we briefly discussed PQI-based phase estimator, here we present PI- and Bi-Phase-based estimators for the first time.

$\psi_x(k, l) = \angle X(k, l)$ and $\psi_y(k, l) = \angle Y(k, l)$, respectively.

The noisy signal $y(n)$ and its pitch frequency estimate $F_0(n)$ are supplied to *TSUP on Reference Phases* block, which outputs the pre-enhanced reference phases $\widehat{\Phi}_x(\bar{h}, l)$. The $\bar{h}$ denotes the set of reference harmonics. Then, *Inter-Component Phase Analysis* block outputs the linear phases $\phi_{lin}(h, l)$ and the noisy differential phases $\Delta\Phi_y(\bar{h}, h, l)$, that denote one of the considered constraints: PQI, PI or Bi-Phase. Then smoothing filter is applied in the *TSUP on Constrained Phase* block resulting in a smoothed estimate of differential phases $\Delta\widehat{\Phi}_x(\bar{h}, h, l)$. Finally, the *Phase Synthesis* block calculates the harmonic phase estimate $\widehat{\Phi}_x(h, l)$ based on $\Delta\widehat{\Phi}_x(\bar{h}, h, l)$ and $\widehat{\Phi}_x(\bar{h}, l)$, and subsequently combines the resulting $\widehat{\Phi}_x(h, l)$ with $\phi_{lin}(h, l)$ in order to synthesize the enhanced instantaneous STFT phase $\widehat{\psi}_x(k, l)$. The output $\widehat{\psi}_x(k, l)$ is used along with the spectral amplitude estimate to eventually synthesize the enhanced speech signal $\hat{x}(n)$. The spectral amplitude estimate is either extracted directly from the noisy observation (denoted as $|Y(k, l)|$) in a *phase-only enhancement* scenario, or obtained by a conventional speech enhancement technique (denoted as $\widehat{A}_x(k, l)$) in a *phase enhancement combined with enhanced magnitude* scenario. For more details, the Algorithm 1 describes the proposed phase enhancement procedure shown in Fig. 1.

3.1. Signal model

Since the considered phase constraints define the relationship between the phase components of particular harmonics, the harmonic model is used for phase estimation and signal reconstruction in this work.

A clean speech signal $x(n)$ is decomposed into segments windowed by prototype window $w(n')$, such that each segment $x_w(n', l)$ at frame l is defined as follows:

$$x_w(n', l) = x(n' + t(l))\cdot w(n'), \tag{18}$$

where $t(l)$ denotes the time instant at each frame l, and n' denotes the STFT time index $n' \in [-(N_l - 1)/2, (N_l - 1)/2]$ where N_l is the analysis window length. To prevent the influence of the number of periods during calculation of $t(l)$, the suggestions given in Degottex and Erro (2014b) are used during the calculation. In particular, 4 analysis instants per period are used in order to compute a reliable short-term phase variance:

$$t(l) = t(l-1) + \frac{1}{4}\cdot\frac{1}{F_0(l-1)} \quad \text{with } t(0) = 0. \tag{19}$$

Finally, the modeled noisy speech signal $y(n)$ is denoted by sum of segments $y(n', l)$:

$$\begin{aligned} y(n', l) = x_w(n', l) + \nu(n', l) &= w(n')\cdot \sum_{h=1}^{H_l} A_x(h, l)\cos\underbrace{(\phi_{lin}(n', h, l) + \Phi_x(h, l))}_{\Psi_x(n',h,t)} + \nu(n', l) \\ &= w(n')\cdot \sum_{h=1}^{H_l} A_y(h, l)\cos\underbrace{(\phi_{lin}(n', h, l) + \Phi_y(h, l))}_{\Psi_y(n',h,t)}, \end{aligned} \tag{20}$$

where H_l denotes the number of harmonics at frame l; $h \in [1, H_l]$ denotes the harmonic index; $\phi_{lin}(n', h, l) = 2\pi h n' F_0(l)/f_s$ denotes the linear phase; $\Phi_x(h, l)$ and $\Phi_y(h, l)$ denote the clean and noisy harmonic phase, respectively; $\Psi_x(n', h, l)$ and $\Psi_y(n', h, l)$ denote the clean and noisy instantaneous phase, respectively; $\nu(n', l)$ denotes the noise.

3.2. Inter-component phase analysis based on PQI constraint

Calculation of PQI constraint is based on Eq. (16) with appropriate unambiguous definition range discussed in Section 2.4. In order to convert the range to $[-\pi, \pi)$, the result of calculation is multiplied by scaling factor $h/\bar{h}$ after wrapping:

$$PQI(\bar{h}, h, l) = \frac{h}{\bar{h}}\left(\Phi(\bar{h}, l) - \frac{\Phi(h, l)\cdot\bar{h}}{h}\right)\Bigg|_{\frac{2\pi\cdot\bar{h}}{h}}. \tag{21}$$

The harmonic index $\bar{h}$ is referred to as PQI reference harmonic in all further notation. The harmonic phase of any arbitrary component with harmonic index $h \in [1, H_l]$ is derived from Eq. (21) as follows:

$$\Phi(h, l) = \frac{h\cdot\Phi(\bar{h}, l)}{\bar{h}} - PQI(\bar{h}, h, l). \tag{22}$$

3.3. Inter-component phase analysis based on PI and Bi-Phase constraints

Calculation of PI and Bi-Phase constraints are based on Eq. (5) and Eq. (9), respectively. Similar to case of PQI, the unambiguous definition range of PI differs from $[-\pi, \pi)$ according to discussion in Section 2.4. Hence in order to convert the range to $[-\pi, \pi)$, the result of calculation by Eq. (5) is multiplied by 2 after wrapping:

$$PI(K_1, K_2, K_3, l) = 2\cdot\left(\frac{\Phi(K_1, l) + \Phi(K_3, l)}{2} - \Phi(K_2, l)\right)\Bigg|_{\pi}. \tag{23}$$

Unlike the case with PQI, both Bi-Phase and PI describe relations between three components. Therefore, two harmonics are considered to be the reference ones in this scenario. Moreover, there exist three different combination of reference harmonics for each constraint, which yields in three equations for obtaining the harmonic phase of component with harmonic index $h \in [1, H_l]$. In all further notation, indices K_1, K_2 and K_3 denote the left-most, center and the right-most reference harmonics, respectively, in a given constraint configuration.

For Bi-Phase, the harmonic phase equations of any arbitrary component with harmonic index $h \in [1, H_l]$ are denoted as follows:

$$\Phi(h, l) = BiPh(h, K_2, K_3, l) - \Phi(K_2, l) + \Phi(K_3, l), \tag{24}$$

$$\Phi(h, l) = BiPh(K_1, h, K_3, l) - \Phi(K_1, l) + \Phi(K_3, l), \tag{25}$$

$$\Phi(h, l) = -BiPh(K_1, K_2, h, l) + \Phi(K_1, l) + \Phi(K_2, l). \tag{26}$$

For PI calculated by Eq. (23), the harmonic phase equations of any arbitrary component with harmonic index $h \in [1, H_l]$ are denoted as follows:

$$\Phi(h, l) = PI(h, K_2, K_3, l) + 2\cdot\Phi(K_2, l) - \Phi(K_3, l), \tag{27}$$

$$\Phi(h, l) = \frac{1}{2}\cdot(\Phi(K_1, l) + \Phi(K_3, l) - PI(K_1, h, K_3, l)), \tag{28}$$

$$\Phi(h, l) = PI(K_1, K_2, h, l) + 2\cdot\Phi(K_2, l) - \Phi(K_1, l). \tag{29}$$

Eqs. (26) and (29) will be used for harmonic phase calculations for Bi-Phase and PI scenarios respectively. Nevertheless, the rest of the equations can be potentially useful under the special noise conditions, as well as in other speech processing applications.

In order to obtain the whole set $\{\Phi(1, l), \Phi(2, l), \ldots, \Phi(H_l, l)\}$ of harmonic phases based on two reference harmonics, the iterative procedure is considered for PI and Bi-Phase constraints and depicted in Fig 2. At the $I = 0$ iteration, the $\Phi(3, l)$ is calculated based on pure reference phases $\Phi(1, l)$ and $\Phi(2, l)$. The next $I = 1$ iteration involves the calculated $\Phi(3, l)$ phase to obtain the following $\Phi(4, l)$ phase, and so on. The iterative algorithm requires $N_i = H_l - 2$ iterations in order to obtain all the H_l harmonic phases. Note that starting from $I = 2$, the PI iterative algorithm operates solely on harmonic phases that are calculated on previous iterations, whereas Bi-Phase iterative algorithm relies on pure reference phase $\Phi(1, l)$ at every iteration step.

3.4. Temporal smoothing

In order to reduce the noise component in unwrapped harmonic phase, the considered estimators exploit the principle of temporal smoothing of unwrapped phase (TSUP) proposed in Kulmer and

```
$y(n) \leftarrow$ Load noisy speech
$(\mathrm{ICPR}, \bar{h}) \leftarrow$ Define ICPR method (PQI, PI or Bi-Phase) and initialize a set of the reference harmonics $\bar{h}$ for the corresponding ICPR method
$(F_0(n), p_v(n)) \leftarrow$ Evaluate fundamental frequency and voicing probability
$H_l \leftarrow$ Set the number of harmonics for analysis/synthesis
Convert signal $y(n)$ into windowed STFT frames of $N_{frames}$ count with pitch-synchronous frame shift based on $F_0(n)$, $p_v(n)$ and Eq. (19)
for $l = 1$ to $N_{frames}$ do {Estimate the instantaneous parameters of noisy speech}
    Compute DFT of $l$-th frame
    for $h = 1$ to $H_l$ do
        Evaluate harmonic frequencies, amplitudes $A_y(h, t)$ and instantaneous phases $\Psi_y(h, l)$ from DFT
    end for
end for
for all $h \in \bar{h}$ do {Apply temporal smoothing on unwrapped reference phases}
    for $l = 1$ to $N_{frames}$ do
        Perform the phase unwrapping of $\Psi_y(\bar{h}, l)$
        $\Phi_y(\bar{h}, l) \leftarrow \Psi_y(\bar{h}, l) - \phi_{lin}(\bar{h}, l)$
        $\hat{\Phi}_x(\bar{h}, l) \leftarrow$ TSUP on the reference harmonic phase $\Phi_y(\bar{h}, l)$
    end for
end for
for $h = 1$ to $H_l$, $h \notin \bar{h}$ do {Calculate enhanced harmonic phases based on smoothed ICPR}
    for $l = 1$ to $N_{frames}$ do
        Perform the phase unwrapping of $\Psi_y(h, l)$
        $\Phi_y(h, l) \leftarrow \Psi_y(h, l) - \phi_{lin}(h, l)$
        $\Delta\Phi_y(\bar{h}, h, l) \leftarrow$ ICPR calculation based on $\Phi_y(h, l)$ and $\hat{\Phi}_x(\bar{h}, l)$ by Eq. (21), or (23), or (9)
        $\Delta\hat{\Phi}_x(\bar{h}, h, l) \leftarrow$ TSUP on computed ICPR $\Delta\Phi_y(\bar{h}, h, l)$ by Eq. (30)
        $\hat{\Phi}_x(h, l) \leftarrow$ Calculate enhanced harmonic phase based on $\Delta\hat{\Phi}_x(\bar{h}, h, l)$ and $\hat{\Phi}_x(\bar{h}, l)$ by Eq. (22), or (26), or (29).
    end for
end for
$\hat{\psi}_x(k, l) \leftarrow$ Transform $\hat{\Phi}_x(h, l) + \phi_{lin}(h, l)$ into STFT domain by Eq. (31)
if Phase-only enhancement scenario then
    $|\hat{X}(k, l)| \leftarrow |Y(k, l)|$ taken directly from the noisy signal
else
    $|\hat{X}(k, l)| \leftarrow \hat{A}_x(k, l)$ obtained by conventional amplitude enhancement
end if
return $\hat{x}(n) \leftarrow \mathrm{IDFT}(|\hat{X}(k, l)|, \hat{\psi}_x(k, l))$, followed by the overlap-and-add procedure
```

Algorithm 1. Speech Enhancement Using Inter-Component Phase Relations

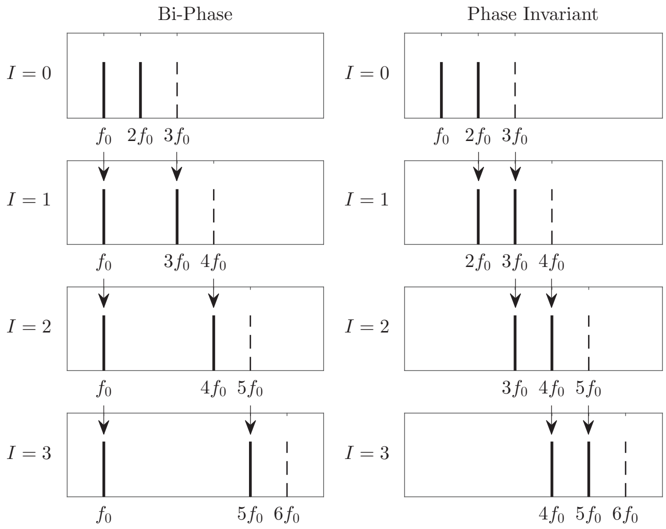


Fig. 2. First four iterations of harmonic phase calculations for Bi-Phase (left) and Phase Invariant (right). I denotes an iteration index; solid line denotes the parameters of reference harmonic at particular iteration; dashed line denotes the parameters of estimated harmonic at particular iteration; arrow denotes the harmonics whose parameters are inherited from previous iteration.

Mowlaee (2015b); Mowlaee and Kulmer (2015b). The original motivation for the usefulness of the temporal smoothing of phase is that the glottal source shape of voiced segment varies slowly across time (Degottex and Erro, 2014a) and the glottal pulse shape of a clean speech signal presents smooth changes in time (low variance) which is justified by the small phase distortion deviation values (Koutsogiannaki et al., 2014). Therefore, the unwrapped phase of voiced segment in the clean speech varies slowly too. Since the noise component increases the variance of unwrapped phase, application of temporal smoothing will decrease this variance, matching better to that in the clean spectral phase, which eventually results in an enhanced reconstructed speech signal.

For all the considered estimators, the reference phases $\Phi(\overline{h}, l)$ are of high importance, as their corruption by noise leads to erroneous results for the corresponding harmonic phases throughout the harmonics. Therefore, before calculating the enhanced harmonic phases $\widehat{\Phi}_x(h, l)$, the reference phases $\Phi_y(\overline{h}, l)$ are pre-enhanced by means of TSUP, resulting in $\widehat{\Phi}_x(\overline{h}, l)$.

Let differential phase $\Delta\Phi_y(\overline{h}, h, l)$ denotes the selected ICPR: either $PQI(\overline{h}, h, l)$ in (21), or $PI(K_1, K_2, K_3, l)$ in (23), or $BiPh(K_1, K_2, K_3, l)$ in (9). Accordingly, $\overline{h}$ in $\Delta\Phi_y(\overline{h}, h, l)$ denote the set of reference harmonics. The reference harmonic phases in formulas (21), (23) and (9) correspond to pre-enhanced $\widehat{\Phi}_x(\overline{h}, l)$, whereas the non-reference ones correspond to $\Phi_y(h, l)$. Differential phases $\Delta\Phi_y(\overline{h}, h, l)$ are then smoothed across time by mean averaging:

$$\Delta\widehat{\Phi}_x(\overline{h}, h, l) = \angle\frac{1}{|\mathscr{W}|}\sum_{\tilde{l}\in\mathscr{W}} e^{j\Delta\Phi_y(\overline{h}, h, \tilde{l})}, \tag{30}$$

where $\mathscr{W}$ denotes the set of frames that lie within a range of 100 ms around frame l. This filter length was chosen empirically, after analyzing the behavior of PQI, PI and Bi-Phase over time. Note that differential phases for PQI-based estimators are smoothed only for configurations where $h = N\overline{h}$, $N \in \mathbb{N}$.

Finally, the enhanced harmonic phases $\widehat{\Phi}_x(h, l)$ are calculated by Eq. (22) for PQI, Eq. (26) for Bi-Phase and Eq. (29) for PI, where $\Phi(\overline{h}, l)$ and $\Phi(h, l)$ are substituted by $\widehat{\Phi}_x(\overline{h}, l)$ and $\widehat{\Phi}_x(h, l)$, respectively, and ICPR in each formula is substituted by its corresponding smoothed estimate $\Delta\widehat{\Phi}_x(\overline{h}, h, l)$.

3.5. Synthesis of phase enhanced speech

The synthesis of the enhanced instantaneous STFT phase $\widehat{\psi}_x(k, l)$ requires the transformation of enhanced harmonic phase $\widehat{\Phi}_x(h, l)$ in the STFT domain. It is achieved by modification of the STFT frequency bins, that contained within the main lobe width of the analysis window (Kulmer and Mowlaee, 2015a). The enhanced instantaneous STFT phase is then given by:

$$\widehat{\psi}_x(\lfloor h\omega_0(l)K\rceil + i, l) = \widehat{\Phi}_x(h, l) + \phi_{lin}(h, l),$$
$$\forall\, i \in [-N_p(l)/2, N_p(l)/2], \tag{31}$$

where K denotes the DFT length with $k \in [0, K-1]$, and $N_p(l)$ denotes the minimum value of either the main lobe width of the analysis window N_w or frequencies close to neighboring harmonic, and is given by $N_p(l) = \min(N_w, \omega_0(l)K/(2\pi))$. Finally, the enhanced speech signal in the STFT domain is given by:

$$\widehat{X}(k, l) = |Y(k, l)|e^{j\widehat{\psi}_x(k,l)}, \tag{32}$$

The corresponding time domain signal $\hat{x}(n)$ is obtained by the inverse DFT of $\widehat{X}(k, l)$ followed by the overlap-add procedure.

The proposed speech enhancement technique is based solely on the phase enhancement, therefore spectral amplitude estimate in Eq. (32) is directly assigned to be the noisy spectral amplitude $|Y(k, l)|$. In the following we call this a *phase-only enhancement* scenario. Nevertheless, the discussed algorithms can be combined together with the conventional amplitude enhancement techniques, giving an enhanced spectral amplitude $\widehat{A}_x(k, l)$. Then, $|\widehat{X}(k, l)|=\widehat{A}_x(k, l)$. In the following we call this a *phase enhancement combined with enhanced magnitude* scenario.

4. Results

4.1. Experiment setup

We randomly choose 50 utterances of 20 speakers (10 male and 10 female) from GRID corpus (Cooke et al., 2006). The utterances were corrupted by the following noise types: white, babble and factory noise files taken from NOISEX-92 database (Varga et al., 1992); car and street noise files taken from NOIZEUS database (Hu and Loizou, 2007). The SNR levels ranged from − 5 to 10 dB with 5 dB step.

In order to quantify the error introduced between the estimated versus the clean ICPR values, we define the unwrapped mean square error (UnMSE) criterion, similar to one defined in Mowlaee et al. (2016a, Ch. 6):

$$\text{UnMSE}(\overline{h}, h) = \frac{1}{|\mathscr{L}|}\sum_{l\in\mathscr{L}}\left(\cos(\Delta\Phi(\overline{h}, h, l) - \Delta\widehat{\Phi}(\overline{h}, h, l))\right)^2, \tag{33}$$

where $\mathscr{L}$ denotes the set of voiced frames, $|\mathscr{L}|$ denotes its cardinality, $\overline{h}$ denotes the set of reference harmonics, $\Delta\Phi(\overline{h}, h, l)$ and $\Delta\widehat{\Phi}(\overline{h}, h, l)$ denote the clean and estimated observation, respectively, defined for the selected ICPR by Eqs. (21) for PQI, (23) for PI, and (9) for Bi-Phase. The closer the UnMSE value to 1, the lower the estimation error. In order to eliminate the impact of unvoiced regions in the resulting UnMSE outcome, only the voiced frames are considered in $\mathscr{L}$.

The speech enhancement performance achieved by the phase estimation methods was evaluated using the following evaluation criteria: perceptual evaluation of speech quality (PESQ) (Rix et al., 2001), short-time objective intelligibility measure (STOI) (Taal et al., 2011), and unwrapped root mean square estimation error (UnRMSE) (Gaich and Mowlaee, 2015). PESQ and STOI instrumentally predict the perceived speech quality and speech intelligibility, respectively, while UnRMSE quantifies the spectral phase estimation accuracy. As our benchmarks, we report results obtained by MAP (Kulmer and Mowlaee, 2015a) and STFTPI (Krawczyk and Gerkmann, 2014) phase estimators. Finally, we include the clean spectral phase results demonstrating the upper-bound performance achievable by a phase estimator in a speech enhancement

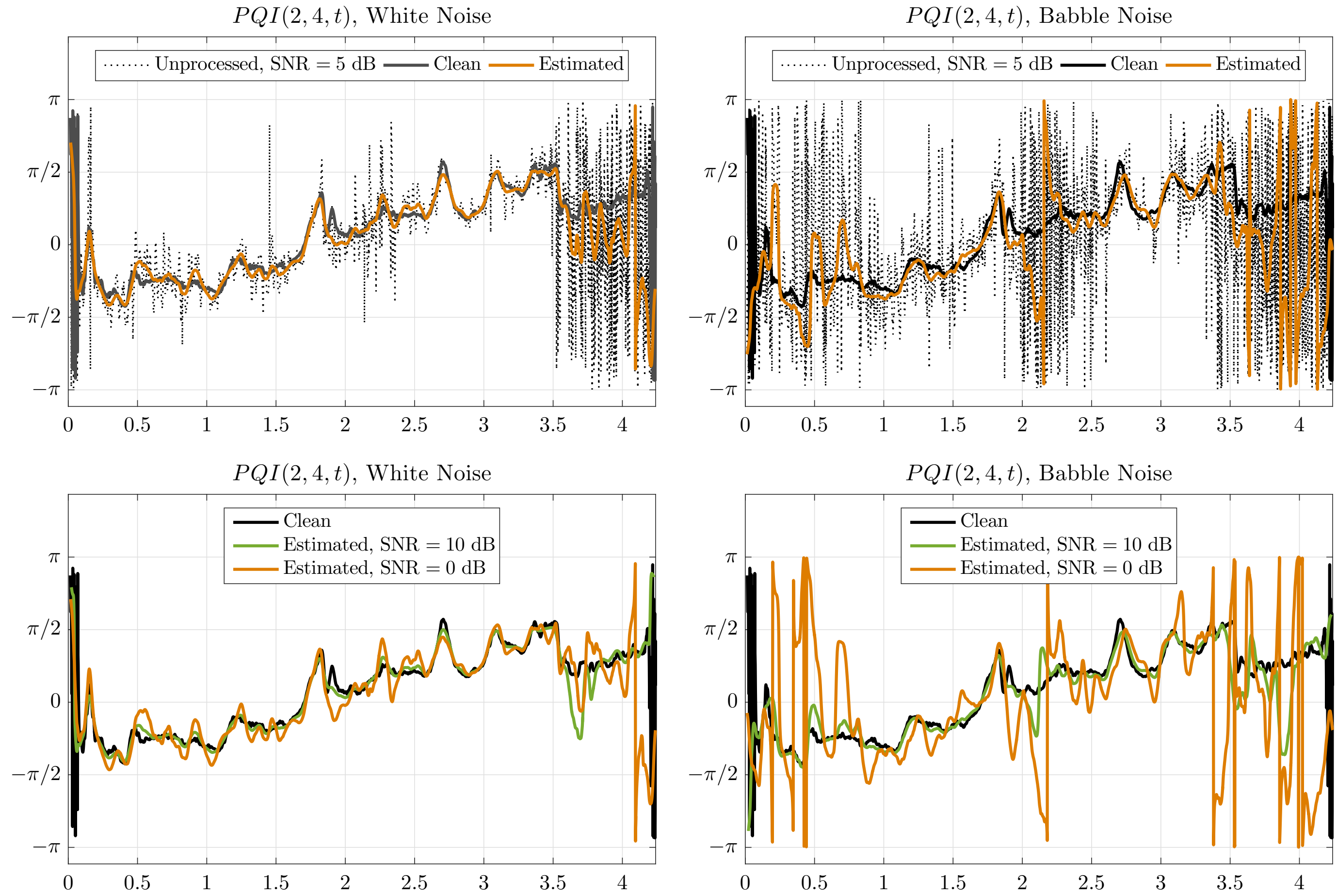


Fig. 3. Estimation accuracy of *PQI*(2, 4, *t*) for sustained \aeiou\\ sequence in white (left) and babble (right) noise. Top: estimated *PQI*(2, 4, *t*) trajectory using PQI-based estimator with $\bar{h} = 2$ reference harmonic compared to the unprocessed trajectory (lower-bound) and clean trajectory (upper-bound) obtained for SNR = 5 dB. Bottom: estimated *PQI*(2, 4, *t*) trajectories at 0 dB and 10 dB SNR levels compared to their corresponding trajectories obtained from the clean signal (denoting the upper-bound performance).

framework.

The discussed phase estimators require evaluation of fundamental frequency across noisy utterances. In this setup PEFAC (Gonzalez and Brookes, 2014) was chosen as a fundamental frequency evaluation algorithm due to its robustness to high levels of noise. To evaluate the robustness of the proposed phase estimators against fundamental frequency estimation errors, we consider two scenarios: (i) f_0-oracle scenario setup where f_0 information is obtained from the clean speech signal, (ii) f_0-blind scenario where f_0 information is obtained from the noisy speech input.

4.2. Proof-of-concept I: inter-component phase relations estimation

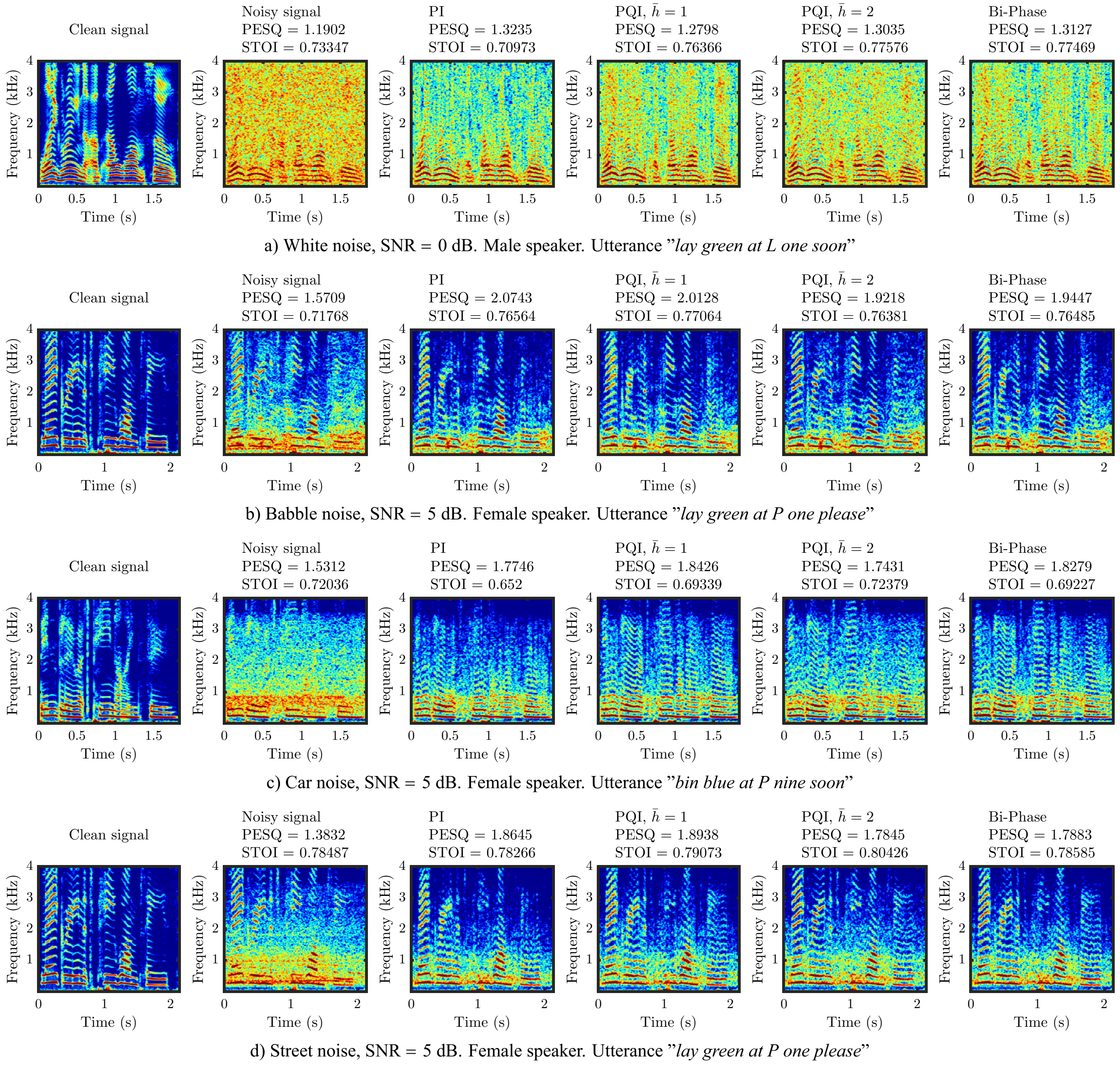
Here we present proof-of-concept results for *PQI*(2, 4, *t*) estimation using PQI-based estimator with $\bar{h} = 2$ reference harmonic. The results are shown for a blind setup, where f_0 is estimated from the noisy observation. We consider a sustained \aeiou\\ sequence uttered by male speaker, corrupted by white and babble noise at SNRs of 0, 5 and 10 dB. The estimated *PQI*(2, 4, *t*) trajectories and their clean and noisy variants are depicted in Fig 3.

In both noise scenarios the proposed PQI-based estimator results in a smoother PQI trajectory compared to the unprocessed one, getting to a PQI pattern closer to the observed one from the clean signal. At lower SNR levels, the fluctuations in the resulting PQI trajectory are increased, yielding considerable increase of variance versus the desired PQI pattern achievable from the clean speech. The variance increase is more pronounced for babble noise versus the white noise scenario, due to the nature of the non-stationary noise statistics in the latter. Nevertheless, the smoothed pattern of the estimated PQI trajectory obtained by the proposed PQI-based estimator is visible at all SNR levels and for both noise types.

4.3. Proof-of-concept II: speech enhancement

In this section, as the proof-of-concept, we present the results of experiments to evaluate the performance of the proposed phase estimators in f_0-blind scenario. We consider male and female utterances, corrupted by white (0 decibels), babble (5 dB), car (5 dB) and street (5 dB) noise. The spectrograms of utterances enhanced by the proposed estimators are depicted in Fig 4. For a perceptual comparison we refer to Barysenka et al. (2017) where several listening examples are found.

In all noise scenarios, the harmonic structure of enhanced speech is visually improved for all three proposed phase estimators in comparison to the harmonic structure obtained for noisy speech. At the same time, the PESQ scores are also improved for all scenarios. Another promising observation is that some of the estimators yield to joint improvement of perceived quality (PESQ) and speech intelligibility (STOI) at particular noise scenarios. This is an important finding since speech enhancement methods are often reported to degrade speech intelligibility or not being capable of improving perceived quality and intelligibility jointly (Loizou and Kim, 2011).



a) White noise, SNR = 0 dB. Male speaker. Utterance "*lay green at L one soon*"

b) Babble noise, SNR = 5 dB. Female speaker. Utterance "*lay green at P one please*"

c) Car noise, SNR = 5 dB. Female speaker. Utterance "*bin blue at P nine soon*"

d) Street noise, SNR = 5 dB. Female speaker. Utterance "*lay green at P one please*"

Fig. 4. Spectrograms of clean, noisy and enhanced (f_0-blind estimation) speech in (a) white, (b) babble, (c) car and (d) street noise scenarios. The PESQ and STOI scores obtained by each method are shown above each spectrogram.

4.4. *Evaluation of inter-component phase relations estimation accuracy*

The accuracy of ICPR estimation is evaluated using UnMSE as defined in (33). For a better representation on the achievable ICPR estimation improvement, in the following we report the results as the relative UnMSE improvement in the estimated ICPR given by:

$$\delta\text{UnMSE} = \left(\frac{\text{UnMSE}_{\text{estimated}}}{\text{UnMSE}_{\text{unprocessed}}} - 1\right)\cdot 100\%. \tag{34}$$

The positive and negative values of δUnMSE denote the improved and degraded estimation accuracy, respectively, compared to the noisy (unprocessed) signal.

The results of *PQI*(2, 4), *BiPh*(1, 2, 3) and *PI*(1, 2, 3) estimation accuracy obtained by the proposed and benchmark estimators are tabulated in Tables 2–4. The results are grouped by different noise scenarios and SNR levels. The top-performing results at each SNR value are marked in boldface, whereas the values below the noisy UnMSE scores at each SNR are highlighted in italic. Following observations are made:

- *ICPR estimation error:* The improvement in UnMSE is observed for all considered noise scenarios and SNR levels by all three proposed estimators. The most notable δUnMSE improvement is obtained for mid-SNR ratios of 0 and 5 dB for all considered trajectories and noise scenarios. For the lowest SNR = −5 dB, the most notable improvement of UnMSE is obtained for *BiPh*(1, 2, 3) and *PI*(1, 2, 3) for white and factory noise scenarios.
- *Comparison with benchmarks:* Based on the evaluated trajectories, the ICPR estimation accuracy is obviously higher at all considered SNR levels for all proposed estimators compared to MAP and STFTPI. The

Table 2
Estimation accuracy of *PQI*(2, 4) for PQI-based (with $\bar{h} = 2$), MAP and STFTPI estimators.

SNR level (dB)	UnMSE				δUnMSE (%)			
	−5	0	5	10	−5	0	5	10
White Noise								
Noisy (Unprocessed)	0.712	0.752	0.797	0.835	—	—	—	—
PQI, $\bar{h} = 2$ (Proposed)	**0.740**	**0.785**	**0.820**	**0.846**	**3.93**	**4.39**	**2.89**	**1.32**
MAP (Kulmer and Mowlaee, 2015a)	0.723	0.768	0.810	0.843	1.54	2.13	1.63	0.958
STFTPI (Krawczyk and Gerkmann, 2014)	*0.688*	*0.685*	*0.699*	*0.703*	*−3.37*	*−8.91*	*−12.3*	*−15.8*
Factory noise								
Noisy (Unprocessed)	0.693	0.720	0.768	0.814	—	—	—	—
PQI, $\bar{h} = 2$ (Proposed)	**0.710**	**0.756**	**0.806**	**0.835**	**2.45**	**5.00**	**4.95**	**2.58**
MAP (Kulmer and Mowlaee, 2015a)	0.698	0.732	0.783	0.827	0.722	1.67	1.95	1.60
STFTPI (Krawczyk and Gerkmann, 2014)	*0.686*	*0.684*	*0.697*	*0.698*	*−1.01*	*−5.00*	*−9.24*	*−14.3*
Babble noise								
Noisy (Unprocessed)	0.688	0.706	0.753	0.810	—	—	—	—
PQI, $\bar{h} = 2$ (Proposed)	**0.703**	**0.734**	**0.794**	**0.841**	**2.18**	**3.97**	**5.44**	**3.83**
MAP (Kulmer and Mowlaee, 2015a)	0.691	0.714	0.768	0.823	0.436	1.13	1.99	1.60
STFTPI (Krawczyk and Gerkmann, 2014)	0.689	*0.693*	*0.691*	*0.695*	0.145	*−1.84*	*−8.23*	*−14.2*
Car Noise								
Noisy (Unprocessed)	0.694	0.714	0.748	0.784	—	—	—	—
PQI, $\bar{h} = 2$ (Proposed)	**0.711**	**0.743**	**0.780**	**0.812**	**2.45**	**4.06**	**4.28**	**3.57**
MAP (Kulmer and Mowlaee, 2015a)	0.698	0.725	0.763	0.797	0.576	1.54	2.01	1.66
STFTPI (Krawczyk and Gerkmann, 2014)	*0.689*	*0.692*	*0.699*	*0.699*	*−0.720*	*−3.08*	*−6.55*	*−10.8*
Street noise								
Noisy (Unprocessed)	0.689	0.703	0.734	0.773	—	—	—	—
PQI, $\bar{h} = 2$ (Proposed)	**0.697**	**0.731**	**0.765**	**0.808**	**1.16**	**3.98**	**4.22**	**4.53**
MAP (Kulmer and Mowlaee, 2015a)	0.692	0.713	0.745	0.785	0.435	1.42	1.50	1.55
STFTPI (Krawczyk and Gerkmann, 2014)	0.693	*0.680*	*0.682*	*0.695*	0.581	*−3.27*	*−7.08*	*−10.1*

only exception occurs in white noise scenario at SNR = 10 dB for *BiPh*(1, 2, 3) and *PI*(1, 2, 3) trajectories, where MAP estimator slightly outperforms the proposed ones (less than 0.15% difference of δUnMSE in both cases). The results show that STFTPI improves δUnMSE only for babble and street noise at SNR = −5 dB for *PQI*(2, 4) trajectory, while not improving this metric for other scenarios.

Based on these results, we conclude that the proposed estimators are capable of improving ICPR estimation accuracy under the various noise conditions, which motivates us to conduct the phase-aware speech enhancement experiments with exploitation of the proposed estimators. The results of these experiments are presented in the following sections.

Table 3
Estimation accuracy of *BiPh*(1, 2, 3) for Bi-Phase-based, MAP and STFTPI estimators.

SNR level (dB)	UnMSE				δUnMSE (%)			
	− 5	0	5	10	− 5	0	5	10
White noise								
Noisy (Unprocessed)	0.761	0.822	0.873	0.910	—	—	—	—
Bi-Phase (Proposed)	**0.817**	**0.863**	**0.899**	0.914	**7.36**	**4.99**	**2.98**	0.440
MAP (Kulmer and Mowlaee, 2015a)	0.779	0.839	0.883	**0.915**	2.37	2.07	1.15	**0.549**
STFTPI (Krawczyk and Gerkmann, 2014)	*0.705*	*0.703*	*0.711*	*0.714*	*−7.35*	*−14.4*	*−18.6*	*−21.5*
Factory noise								
Noisy (Unprocessed)	0.720	0.778	0.843	0.891	—	—	—	—
Bi-Phase (Proposed)	**0.764**	**0.832**	**0.880**	**0.906**	**6.11**	**6.94**	**4.39**	**1.68**
MAP (Kulmer and Mowlaee, 2015a)	0.730	0.792	0.857	0.900	1.39	1.80	1.66	1.01
STFTPI (Krawczyk and Gerkmann, 2014)	*0.697*	*0.707*	*0.721*	*0.711*	*−3.19*	*−9.13*	*−14.5*	*−20.2*
Babble noise								
Noisy (Unprocessed)	0.690	0.734	0.804	0.865	—	—	—	—
Bi-Phase (Proposed)	**0.713**	**0.782**	**0.849**	**0.890**	**3.33**	**6.34**	**5.60**	**2.89**
MAP (Kulmer and Mowlaee, 2015a)	0.695	0.751	0.820	0.876	0.725	2.32	1.99	1.27
STFTPI (Krawczyk and Gerkmann, 2014)	*0.680*	*0.687*	*0.701*	*0.708*	*−1.45*	*−6.40*	*−12.8*	*−18.2*
Car noise								
Noisy (Unprocessed)	0.723	0.773	0.823	0.867	—	—	—	—
Bi-Phase (Proposed)	**0.758**	**0.820**	**0.862**	**0.889**	**4.84**	**6.08**	**4.74**	**2.54**
MAP (Kulmer and Mowlaee, 2015a)	0.734	0.792	0.840	0.877	1.52	2.46	2.07	1.15
STFTPI (Krawczyk and Gerkmann, 2014)	*0.695*	*0.717*	*0.723*	*0.714*	*−3.87*	*−7.24*	*−12.2*	*−17.6*
Street noise								
Noisy (Unprocessed)	0.704	0.747	0.799	0.853	—	—	—	—
Bi-Phase (Proposed)	**0.734**	**0.793**	**0.843**	**0.887**	**4.26**	**6.16**	**5.51**	**3.99**
MAP (Kulmer and Mowlaee, 2015a)	0.712	0.762	0.815	0.865	1.14	2.01	2.00	1.41
STFTPI (Krawczyk and Gerkmann, 2014)	*0.684*	*0.694*	*0.690*	*0.711*	*−2.84*	*−7.10*	*−13.6*	*−16.6*

Table 4
Estimation accuracy of *PI*(1, 2, 3) for PI-based, MAP and STFTPI estimators.

SNR level (dB)	UnMSE				δUnMSE (%)			
	− 5	0	5	10	− 5	0	5	10
					White noise			
Noisy (Unprocessed)	0.736	0.795	0.854	0.897	—	—	—	—
PI (Proposed)	**0.788**	**0.840**	**0.884**	0.902	**7.07**	**5.66**	**3.51**	0.557
MAP (Kulmer and Mowlaee, 2015a)	0.751	0.813	0.867	**0.903**	2.04	2.64	1.52	**0.669**
STFTPI (Krawczyk and Gerkmann, 2014)	*0.695*	*0.701*	*0.711*	*0.704*	*−5.57*	*−11.8*	*−16.7*	*−21.5*
					Factory noise			
Noisy (Unprocessed)	0.703	0.750	0.817	0.873	—	—	—	—
PI (Proposed)	**0.733**	**0.810**	**0.861**	**0.892**	**4.27**	**8.00**	**5.39**	**2.18**
MAP (Kulmer and Mowlaee, 2015a)	0.709	0.767	0.833	0.883	0.853	2.27	1.96	1.15
STFTPI (Krawczyk and Gerkmann, 2014)	*0.690*	*0.694*	*0.705*	*0.711*	*−1.85*	*−7.47*	*−13.7*	*−19.2*
					Babble noise			
Noisy (Unprocessed)	0.689	0.713	0.771	0.838	—	—	—	—
PI (Proposed)	**0.694**	**0.755**	**0.821**	**0.874**	**0.726**	**5.89**	**6.49**	**4.30**
MAP (Kulmer and Mowlaee, 2015a)	0.689	0.724	0.790	0.853	0.000	1.54	2.46	1.79
STFTPI (Krawczyk and Gerkmann, 2014)	*0.683*	*0.687*	*0.699*	*0.698*	*−0.871*	*−3.65*	*−9.34*	*−16.7*
					Car noise			
Noisy (Unprocessed)	0.707	0.746	0.802	0.853	—	—	—	—
PI (Proposed)	**0.739**	**0.797**	**0.844**	**0.877**	**4.53**	**6.84**	**5.24**	**2.81**
MAP (Kulmer and Mowlaee, 2015a)	0.720	0.769	0.822	0.865	1.84	3.08	2.49	1.41
STFTPI (Krawczyk and Gerkmann, 2014)	*0.682*	*0.702*	*0.697*	*0.712*	*−3.54*	*−5.90*	*−13.1*	*−16.5*
					Street noise			
Noisy (Unprocessed)	0.693	0.724	0.767	0.820	—	—	—	—
PI (Proposed)	**0.718**	**0.767**	**0.808**	**0.857**	**3.61**	**5.94**	**5.35**	**4.51**
MAP (Kulmer and Mowlaee, 2015a)	0.703	0.741	0.783	0.832	1.44	2.35	2.09	1.46
STFTPI (Krawczyk and Gerkmann, 2014)	*0.688*	*0.690*	*0.685*	*0.696*	*−0.722*	*−4.70*	*−10.7*	*−15.1*

4.5. *Speech enhancement performance evaluation*

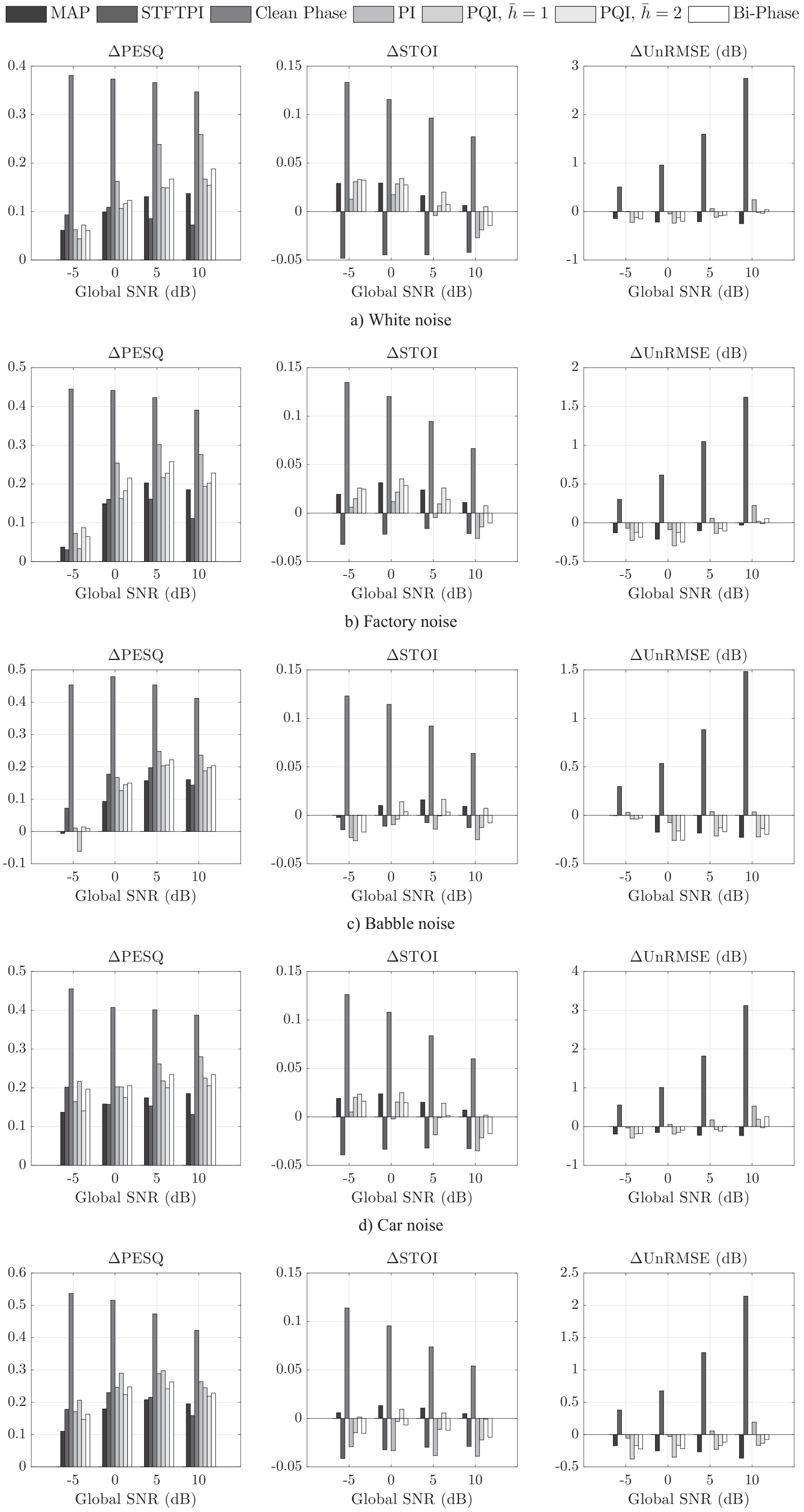
Fig. 5 shows the comparative speech enhancement performance evaluation between the proposed and benchmark phase estimators for f_0-blind scenario. The speech enhancement results are reported in terms of PESQ, STOI and UnRMSE measured as the differences between the scores obtained for the enhanced speech versus the noisy (unprocessed), therefore quantifying the improvement compared to the noisy speech. Note that the clean phase results are excluded from ΔUnRMSE charts. Following observations are made:

- *Speech quality:* The improvement of PESQ is observed at all SNRs for all considered noise scenarios by all proposed estimators, except the case of PQI-based estimator with $\bar{h} = 1$ reference harmonic in babble noise, − 5 dB SNR. The most notable PESQ improvement is obtained by PI-based estimator at SNR ≥ 0 dB for white, factory, babble and car noise scenarios. However, for the lowest considered SNR case of − 5 dB the PQI-based estimators performs better: $\bar{h} = 1$ in car and street noise scenarios, and $\bar{h} = 2$ in white, factory and babble noise scenarios.
- *Speech intelligibility:* In terms of STOI improvement, the PQI-based estimator with $\bar{h} = 2$ reference harmonic is superior at all SNRs for all considered noise scenarios among all the proposed estimators. The other proposed estimators also improve STOI in white, factory and car noise scenarios at SNR ≤ 0 dB, while increase of SNR level leads to intelligibility degradation (initially, by PI-based estimator at SNR = 5 dB, and later by Bi-Phase-based and PQI-based with $\bar{h} = 1$ reference harmonic estimators at SNR = 10 dB). In general, all the proposed estimators show the capability of joint improvement of STOI and PESQ scores under the certain SNR and noise conditions.
- *Phase estimation error:* PQI- and Bi-Phase based estimators show the decrease of phase estimation error in terms of UnRMSE at SNR ≤ 5 dB for all considered noise scenarios, whereas PI-based estimator either unnoticeable reduce the error at lower SNRs, or leads to increased phase estimation error at higher SNRs.
- *Comparison with benchmarks:* In comparison with MAP and STFTPI, the proposed estimators perform better in terms of PESQ in most of the combinations of noise types and SNR levels. As of STOI, only PQI-based estimator with $\bar{h} = 2$ reference harmonic outperforms MAP in white, factory, babble and car noise scenarios at SNR ≤ 5 dB, whereas the MAP estimator remains superior at the highest SNR = 10 dB for all noise scenarios, and in particular at all SNRs in street noise scenario. In terms of UnRMSE, the PQI- and Bi-Phase estimators outperform MAP at SNR ≤ 0 dB in all noise scenarios, whereas the MAP estimator is superior for higher SNRs. In terms of STOI and UnRMSE, the worst case performance of the proposed estimators can be compared to STFTPI performance, since none of the scenarios shows the improvement of STOI and UnRMSE scores by STFTPI estimator. Nevertheless, PI-based estimator performs worse than STFTPI in terms of STOI for some of the noise scenarios.

The results have shown that **PQI-based estimator with** $\bar{h} = 2$**reference harmonic** shows the most balanced and consistent joint improvement of speech quality, speech intelligibility and phase estimation error among all the proposed estimators, and in most cases, among all the benchmarks as well. It is important to emphasize that in contrast to the previous phase estimators, this estimator does not depend on the phase of fundamental frequency harmonic.

4.6. *Robustness to f_0 estimation error*

In this section, we present the performance comparison between the f_0-oracle and f_0-blind scenarios. The results are depicted in Fig. 6 showing result of each of the proposed phase estimators twice, once for f_0-oracle and once for f_0-blind scenario. In the white and street noise scenarios, the proposed estimators show the lowest sensitivity to the f_0 estimation error among the other noise scenarios. On the other hand, the proposed estimators have the highest sensitivity to f_0 estimation error in the babble noise scenario, where the improved f_0 estimation accuracy yields to considerable improvement of PESQ, STOI and UnRMSE scores at SNR ≤ 0 dB in comparison with f_0-blind scenario. Finally, the proposed estimators show the moderate level of f_0 estimation error sensitivity in the factory and car noise scenarios compared to the scenarios that show the lowest (white and street) and the highest (babble) sensitivity. On the whole, the presented results justify the



a) White noise

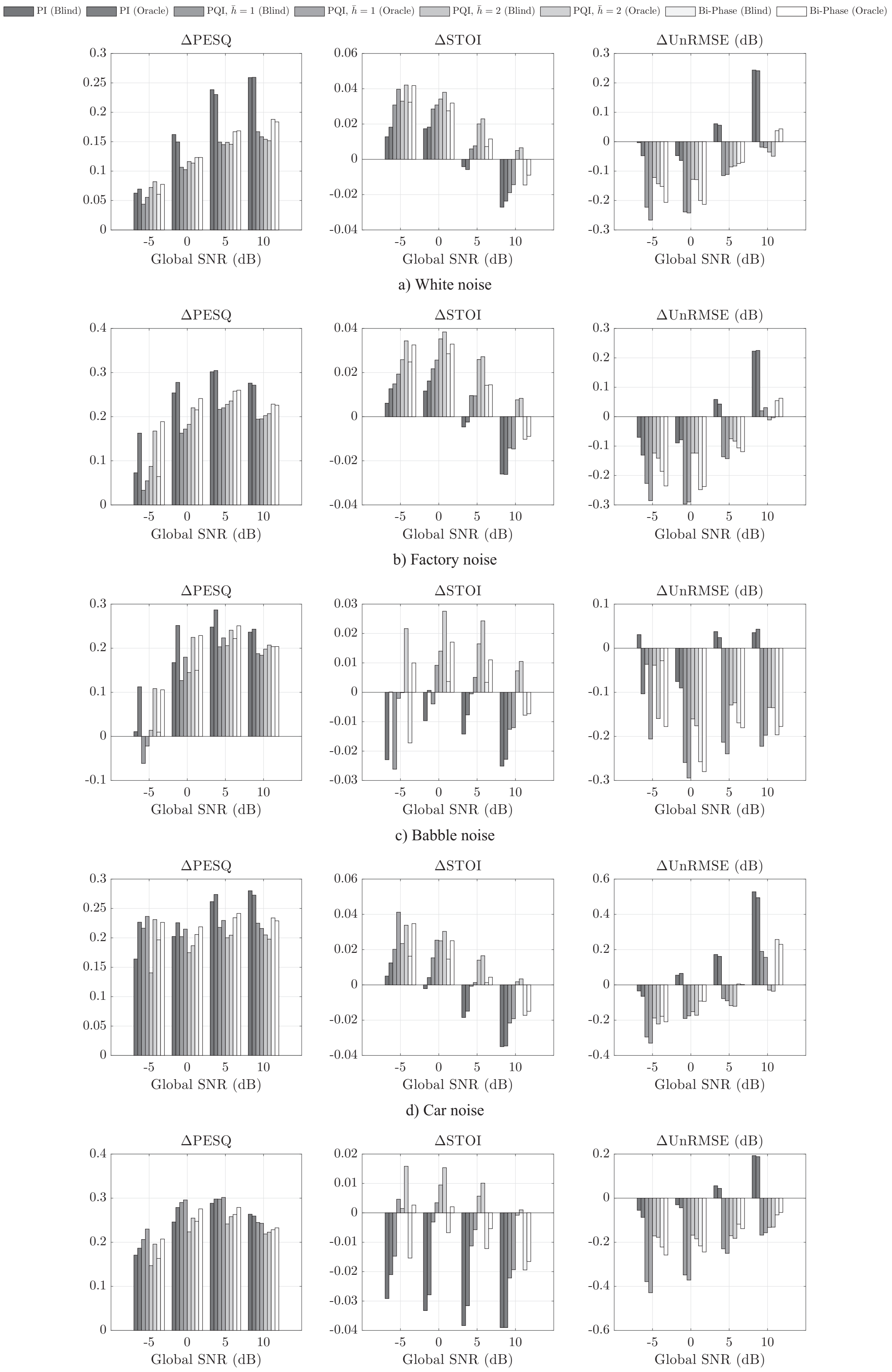
b) Factory noise

c) Babble noise

d) Car noise

e) Street noise

Fig. 5. Speech enhancement in f_0-blind setup in terms of relative improvement in PESQ, STOI and UnRMSE for (a) white, (b) factory, (c) babble, (d) car and (e) street noise scenarios.



a) White noise

b) Factory noise

c) Babble noise

d) Car noise

e) Street noise

Fig. 6. Phase estimation performance as the relative difference between the f_0-oracle and f_0-blind scenarios reported in terms of ΔPESQ, ΔSTOI and ΔUnRMSE for (a) white, (b) factory, (c) babble, (d) car and (e) street noise.

Table 5
PESQ and STOI results for the proposed phase enhancement methods combined with MMSE-LSA for white noise.

SNR level (dB)	PESQ				STOI			
	− 5	0	5	10	− 5	0	5	10
Noisy (Unprocessed)	1.26	1.40	1.60	1.87	0.55	0.64	0.72	0.80
LSA (Ephraim and Malah, 1985)	1.47	1.70	2.00	2.33	0.57	0.65	**0.74**	**0.81**
LSA + PI	**1.58**	**1.91**	**2.24**	**2.55**	0.57	0.65	0.72	*0.78*
LSA + PQI, $\bar{h}$ = 1	1.54	1.85	2.16	2.47	**0.59**	**0.67**	0.73	*0.79*
LSA + PQI, $\bar{h}$ = 2	1.54	1.82	2.14	2.46	0.58	**0.67**	**0.74**	**0.81**
LSA + Bi-Phase	1.55	1.86	2.19	2.51	**0.59**	**0.67**	0.73	*0.79*
Clean Phase	1.79	2.01	2.29	2.62	0.66	0.74	0.81	0.87

Table 6
PESQ and STOI results for the proposed phase enhancement methods combined with MMSE-LSA for factory noise.

SNR level (dB)	PESQ				STOI			
	− 5	0	5	10	− 5	0	5	10
Noisy (Unprocessed)	1.20	1.44	1.75	2.09	0.50	0.61	0.73	0.83
LSA (Ephraim and Malah, 1985)	1.39	1.75	2.12	2.48	*0.49*	0.63	0.75	**0.84**
LSA + PI	**1.51**	**1.99**	**2.35**	**2.65**	0.50	0.63	0.74	*0.82*
LSA + PQI, $\bar{h}$ = 1	1.44	1.89	2.28	2.60	0.50	0.64	0.75	0.83
LSA + PQI, $\bar{h}$ = 2	**1.51**	1.93	2.29	2.62	**0.51**	**0.65**	**0.76**	**0.84**
LSA + Bi-Phase	1.50	1.95	2.33	2.63	**0.51**	0.64	0.75	0.83
Clean Phase	1.79	2.12	2.47	2.82	0.59	0.71	0.81	0.87

Table 7
PESQ and STOI results for the proposed phase enhancement methods combined with MMSE-LSA for babble noise.

SNR level (dB)	PESQ				STOI			
	− 5	0	5	10	− 5	0	5	10
Noisy (Unprocessed)	**1.24**	1.48	1.82	2.19	**0.49**	**0.60**	0.73	0.83
LSA (Ephraim and Malah, 1985)	*1.19*	1.61	2.04	2.45	*0.46*	*0.59*	0.73	0.83
LSA + PI	*1.21*	**1.72**	**2.19**	**2.58**	*0.44*	*0.58*	*0.72*	*0.81*
LSA + PQI, $\bar{h}$ = 1	*1.14*	1.68	2.15	2.53	*0.44*	*0.59*	*0.72*	*0.82*
LSA + PQI, $\bar{h}$ = 2	*1.23*	1.71	2.17	2.56	*0.45*	**0.60**	**0.74**	**0.84**
LSA + Bi-Phase	*1.20*	1.71	2.17	2.55	*0.45*	*0.59*	0.73	0.83
Clean Phase	1.64	2.00	2.40	2.78	0.54	0.68	0.80	0.88

Table 8
PESQ and STOI results for the proposed phase enhancement methods combined with MMSE-LSA for car noise.

SNR level (dB)	PESQ				STOI			
	− 5	0	5	10	− 5	0	5	10
Noisy (Unprocessed)	1.31	1.50	1.71	1.98	0.55	0.66	0.76	0.84
LSA (Ephraim and Malah, 1985)	1.41	1.71	2.06	2.42	**0.58**	0.69	**0.79**	**0.86**
LSA + PI	1.52	**1.91**	**2.28**	**2.59**	0.57	0.67	0.76	*0.82*
LSA + PQI, $\bar{h}$ = 1	**1.54**	1.88	2.23	2.54	**0.58**	0.68	0.77	*0.83*
LSA + PQI, $\bar{h}$ = 2	1.49	1.83	2.20	2.54	**0.58**	**0.70**	0.78	0.85
LSA + Bi-Phase	**1.54**	1.89	2.24	2.56	**0.58**	0.68	0.77	*0.83*
Clean Phase	1.78	2.05	2.39	2.74	0.67	0.76	0.84	0.90

Table 9
PESQ and STOI results for the proposed phase enhancement methods combined with MMSE-LSA for street noise.

SNR level (dB)	PESQ				STOI			
	− 5	0	5	10	− 5	0	5	10
Noisy (Unprocessed)	1.12	1.39	1.72	2.07	0.59	0.70	0.79	**0.86**
LSA (Ephraim and Malah, 1985)	1.28	1.66	2.06	2.42	**0.61**	**0.72**	**0.80**	**0.86**
LSA + PI	1.34	1.81	2.19	2.53	*0.57*	*0.67*	*0.76*	*0.82*
LSA + PQI, $\bar{h}$ = 1	**1.49**	**1.90**	**2.26**	**2.55**	*0.58*	0.70	*0.78*	*0.84*
LSA + PQI, $\bar{h}$ = 2	1.36	1.80	2.19	2.52	0.60	0.71	**0.80**	**0.86**
LSA + Bi-Phase	1.39	1.82	2.20	2.51	0.59	0.70	*0.78*	*0.84*
Clean Phase	1.76	2.11	2.44	2.76	0.69	0.78	0.85	0.90

robustness of phase measurements to the increasing level of a noise nuisance.

4.7. Phase enhancement combined with enhanced magnitude

So far we considered the impact of the proposed phase estimators for phase-only speech enhancement where the noisy spectral amplitude was not processed. Here in this section, we present the effectiveness of the proposed phase estimators when combined with the conventional spectral amplitude speech enhancement. We choose the conventional method amplitude enhancement minimum mean square error log-spectral amplitude estimator (MMSE-LSA) (Ephraim and Malah, 1985). For noise power spectral density estimation we use the unbiased noise estimator (Gerkmann and Hendriks, 2012). For MMSE-LSA implementation we used (Hendriks et al., 2013). Tables 5–9 show the PESQ and STOI results obtained for white, factory, babble, car and street noise scenarios, respectively. Also, we included the speech enhancement result obtained from the noisy (unprocessed) and clean-phase (upper-bound) scenarios. The top-performing results (except the upper-bound) at each SNR value are marked in boldface, whereas the values that less than noisy score at each SNR value are typed in italic. Following observations are made:

- For white and factory noise scenarios, the PI method outperforms others in terms of PESQ. For babble and car noise, a similar trend for PI top-performing PESQ performance is observed except the SNR = −5 dB case. The Bi-Phase method achieves the second best performance in PESQ for white noise scenario. In these noise scenarios, the PQI phase estimators achieve a PESQ performance close to PI. In particular, for street noise, the PQI with the first harmonic reference ($\bar{h}$ = 1) achieves the best PESQ score, 0.1 more than others on average.
- In terms of speech intelligibility, the PQI-based estimators with the first and the second reference harmonic achieve the top STOI scores. In particular, the PQI-based phase estimator with the second harmonic reference ($\bar{h}$ = 2) presents the best performance in STOI. Additionally, the PQI with $\bar{h}$ = 2 is the only method that provides improved speech intelligibility in babble noise scenario. For the other noise scenarios, for low-medium SNRs, the PQI with $\bar{h}$ = 2 provides the best STOI results. Bi-Phase method presents its best scores for white noise over all SNRs and at SNR = −5 dB for other noise scenarios. This observation encourages to conduct a future work regarding the application of bi-spectrum for joint magnitude and phase enhancement of noisy speech.

A comparison between the phase enhancement combined with enhanced magnitude with the phase only results reported earlier is insightful here. When compared to conventional MMSE-LSA method, the combination of the proposed phase enhancement methods with the so-called amplitude-only enhancement (MMSE-LSA) results in a joint improvement of speech quality and intelligibility for low-medium SNR levels of -5 and 0 decibels in stationary noise (white, factory and car). Also, PQI with $\bar{h} = 2$ combined with MMSE-LSA method shows the joint improvement of speech quality and intelligibility for medium-high SNR levels of 0, 5 and 10 decibels in non-stationary (babble) noise, where speech enhancement is challenging in general.

4.8. Discussion on the potential of proposed phase estimators

4.8.1. PQI-based estimators

When comparing PQI-based estimators with $\bar{h} = 1$ and $\bar{h} = 2$ reference harmonics, clear advantage of PQI-based estimator with $\bar{h} = 2$ is observed for all noise scenarios in terms of STOI improvement at all SNR levels. This phenomenon can be explained from the following point of view: for some noise scenarios the fundamental frequency harmonic becomes more corrupted by the noise compared to higher harmonics. This leads to the erroneous estimate of all other PQI combinations with $\bar{h} = 1$, as a result, the restored phase information of higher harmonics becomes corrupted too. In general, the possibility to restore noisy phase information by exploiting the harmonic information other than the fundamental one, pushes the limits of these kinds of estimators. Under the special noise conditions, the reference harmonic can be chosen with consideration of lower noise impact at particular frequencies, therefore making a room for better speech enhancement.

4.8.2. PI-based estimator

PI-based estimator shows considerable improvement of PESQ score in all the noise scenarios. On the other hand, its resulting ΔSTOI score and UnRMSE are not improved but degraded in most cases. The reason for these outcomes lies in the strategy of phase estimation followed in the PI-based estimator as depicted in Fig. 2. At each iteration starting from $I = 2$, the reference harmonics considered for PI calculation include only the restored harmonics on previous iterations. Therefore, the phase estimation error accumulates at each iteration, resulting into degradation of STOI and UnRMSE. Additional impact of this phenomenon is depicted in Fig. 4, where the loss of information at higher harmonics is seen for the PI-based estimator compared to other estimators. The possible ways of resolving this issue are following: (i) add additional reference harmonics at higher frequencies to support accuracy of phase estimation; (ii) discard $\bar{h} = 1$ as initial reference harmonic and choose $\bar{h} = 3$ instead. The latter consideration is established based on the observations we made for the PQI-based estimators, where selection of a higher reference harmonic results in improved intelligibility of the enhanced speech outcome.

4.8.3. Bi-Phase-based estimator

Similar to PI-based estimator, Bi-Phase-based estimator also shows PESQ score improvement in all noise scenarios. However, unlike the PI-based method, the Bi-Phase-based estimator shows also the improvement of STOI at SNR ≤ 5 dB for the white, factory and car noise scenarios, as well as UnRMSE improvement at the same SNRs for all noise scenarios. The difference between the strategies of phase estimation for these two methods is that the Bi-Phase constraint relies explicitly on $\bar{h} = 1$ reference harmonic, which is not restored on previous iterations, and therefore, is not impacted by accumulating estimation error. However, for street noise scenario, Bi-Phase-based estimator could not also gain STOI score similar to PI-based estimator. In order to improve the performance, similar considerations proposed for PI-based estimator can be applied for Bi-Phase-based one. For a more detailed discussion on the potential of Bi-Phase-based estimators for speech enhancement, future studies on the application of higher order spectra in speech enhancement are required.

4.8.4. Joint improvement of speech quality and intelligibility

The proposed PQI-based and Bi-Phase-based estimators show the capability of joint improvement of speech quality and intelligibility. While the improvement in PESQ is consistently achieved for most of the considered scenarios, the improvement in STOI is not that consistent, and is achieved mainly in stationary noise (white, factory and car), in particular for low and medium SNR levels. In order to achieve an improved STOI performance in non-stationary noise (babble and street), often the prior knowledge of f_0 is required as suggested by the results shown in Fig. 6. It can be seen that the joint improvement in non-stationary noise in a blind scenario is either quite restricted or not offered mainly due to the substantial errors in f_0 estimation, which is required by the proposed phase estimators. In contrast, the achievable improvement in oracle f_0 scenarios shown in Fig. 6 reveals the potentially achievable performance of the proposed phase enhancement methods, once the pitch estimation is ideal. This encourages for future study on investigating the impact of f_0 estimators on phase enhancement performance and the phase estimation accuracy.

The achieved joint improvement in speech quality and intelligibility is an important observation as it contrasts with the conventional speech enhancement methods relying on magnitude-only modification reported offering a degraded the speech intelligibility performance (Loizou and Kim, 2011). This observation is well supported by the numerous studies regarding the incorporation of phase information in speech at signal reconstruction with reported improved speech intelligibility results (Oppenheim and Lim, 1981; Mannell, 1990; Paliwal et al., 2011). Furthermore, the recent studies report performance improvement of conventional short-time spectral amplitude MMSE estimator, when some prior knowledge regarding the spectral phase of the signal is taken into account (Mowlaee et al., 2016a, Ch. 4.4.2). These observations are well aligned with the results reported here in this work for the combination of the proposed phase enhancement methods with an enhanced spectral amplitude.

5. Conclusion

In this paper, after an overview on the importance of spectral phase estimation and the recent phase estimation methods in speech signal processing, we proposed new harmonic phase estimators from noisy speech that rely on the relations between the harmonics in a polyharmonic signal as speech. The phase structure was defined by three constraints: Phase Invariance, Phase Quasi-Invariance, and Bi-Phase. These constraints were used to derive estimators for the clean spectral phase from a noisy signal. The proposed phase estimators rely on harmonic model and temporal smoothing of the harmonic phase characterized by the aforementioned constraints. The so-derived enhanced harmonic phase is used in the overlap-and-add routine to output the enhanced-phase speech signal.

Through proof-of-concept and speech enhancement experiments, it was observed that the newly derived phase estimators outperform the existing benchmark methods in terms of perceived quality, speech intelligibility and phase estimation accuracy. The proposed methods were less sensitive to the fundamental frequency estimation errors and also showed higher robustness to the low-frequency noise scenarios where the fundamental frequency estimation is arguably more challenging.

While in this work we only focused on the spectral phase estimation accuracy and its impact on the achievable speech enhancement performance, the usefulness of the proposed estimators could be evaluated for other speech applications including automatic speech recognition, speech synthesis, emotion recognition and speaker recognition. Specifically, the proposed spectral phase estimators can be adapted for single-channel source separation and time-frequency masking whereby further harmonic structure could help to refine the quality of the estimated time-frequency masks and eventually improve the quality of the separation outcome. Another important direction to follow is to explore the potential of the proposed inter-component phase relations in a filter-bank setup rather than the current STFT-based framework. This will clarify how much improvement is achieved by the proposed ICPR constraints alone versus the achievable improvement introduced because of replacing STFT filter-bank with a time-domain filter-bank, that takes into account the ICPR in the time domain.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable comments that helped to considerably improve this paper. The work of Pejman Mowlaee was supported by the Austrian Science Fund (project number P28070-N33).

References

Aarabi, P., 2006. Phase-Based Speech Processing. World Scientific Publishing.

Aksionov, V.A., Vorobiov, V.I., Klimov, A.V., Maslakova, N.A., Sirotko, I.I., 1994. Digital phase processing methods of ultra wide band signals (in Russian). J. Radioeng. Electron. (Signal Gener. Trans. Recept. Radio Syst.) 22, 99–104.

Alsteris, L., Paliwal, K., 2007. Short-time phase spectrum in speech processing: a review and some experimental results. Elsevier Signal Process. 17 (3), 578–616.

Azarov, I.S., Vorobiov, V.I., Davydov, A.G., Petrovsky, A.A., 2011. Studying the connection between quasi-harmonic components of a speech signal. Proceedings of the Twenty-Fourth Session of the Russian Acoustical Society. 3. Russian Acoustical Society, pp. 514–517.

Barysenka, S. Y., Vorobiov, V. I., Mowlaee, P., 2017. Single-channel speech enhancement using inter-component phase relations. www2.spsc.tugraz.at/people/pmowlaee/ICPR.html.

Bochkov, G.N., Gorokhov, K.V., 1995. A synthesis approach of bi-spectral organized signals (in Russian). Tech. Phys. Lett. 21 (16), 27–32.

Boyanov, B., Hadjitodorov, S., Ivanov, T., 1991. Analysis of voiced speech by means of bispectrum. Electron. Lett. 27 (24), 2267–2268.

Boyanov, B., Hadjitodorov, S., Ivanov, T., 1991. Analysis of voiced speech by means of bispectrum. Electronics Letters 27 (24), 2267–2268. http://dx.doi.org/10.1049/el:19911402.

Chacon, C., Mowlaee, P., 2014. Least squares phase estimation of mixed signals. Proceedings of the International Speech Communication Association Interspeech. pp. 2705–2709.

Cooke, M., Barker, J., Cunningham, S., Shao, X., 2006. An audio-visual corpus for speech perception and automatic speech recognition. J. Acoust. Soc. Am. 120, 2421.

Degottex, G., Erro, D., 2014a. A measure of phase randomness for the harmonic model in speech synthesis. Proceedings of the International Speech Communication Association Interspeech. Singapore, pp. 1638–1642.

Degottex, G., Erro, D., 2014b. A uniform phase representation for the harmonic model in speech synthesis applications. EURASIP J. on Audio Speech Music Process. 2014 (1), 38.

Deng, J., Xu, X., Zhang, Z., Frühholz, S., Schuller, B., 2016. Exploitation of phase-based features for whispered speech emotion recognition. IEEE Access 4, 4299–4309.

Ephraim, Y., Malah, D., 1985. Speech enhancement using a minimum mean square error log-spectral amplitude estimator. IEEE Trans. Audio Speech Lang. Process. 33, 443–445.

Espic, F., Botinhao, C.V., King, S., 2017. Direct modelling of magnitude and phase spectra for statistical parametric speech synthesis. Proceedings of the International Speech Communication Association Interspeech. pp. 1383–1387.

Fahringer, J., Schrank, T., Stahl, J., Mowlaee, P., Pernkopf, F., 2016. Phase-aware signal processing for automatic speech recognition. Proceedings of the International Speech Communication Association Interspeech.

Fulchiero, R., Spanias, A.S., 1993. Speech enhancement using the bispectrum. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. 4. pp. 488–491.

Nebabin, V.G., 1994. Methods and Techniques of Radar Recognition (an english translation of a book originally published in Russian in 1984). Artech House Publishers.

Gaich, A., Mowlaee, P., 2015. On speech intelligibility estimation of phase-aware single-channel speech enhancement. Proceedings of the International Speech Communication Association Interspeech. pp. 2553–2557.

Galayev, Y., Kivva, F., 2009. Phase invariant method in radio-wave propagation experiments (in Russian). Prikladnaja Radioelektronika, Kharkiv National University of Radioelectronics, Kharkiv, Ukraine 8 (2), 124–130.

Gavrilov, A.M., 2009. Phase Related Processes of Nonlinear Acoustics: Modulated Waves (in Russian). State Technological University of Taganrog, Russia.

Gerkmann, T., Hendriks, R.C., 2012. Unbiased MMSE-based noise power estimation with low complexity and low tracking delay. IEEE Trans. Audio Speech Lang. Process. 20 (4), 1383–1393.

Gerkmann, T., Krawczyk, M., Le Roux, J., 2015. Phase processing for single-channel speech enhancement: history and recent advances. IEEE Signal Process. Mag. 32 (2), 55–66.

Gonzalez, S., Brookes, M., 2014. PEFAC - a pitch estimation algorithm robust to high levels of noise. IEEE Trans. Audio Speech Lang. Process. 22 (2), 518–530.

Griffin, D., Lim, J., 1984. Signal estimation from modified short-time fourier transform. IEEE Trans. Audio Speech Lang. Process. 32 (2), 236–243.

Hendriks, R.C., Gerkmann, T., Jensen, J., 2013. DFT-domain based single-microphone noise reduction for speech enhancement. Synthesis Lectures on Speech and Audio Processing. Morgan & Claypool Publishers.

Hu, Y., Loizou, P.C., 2007. Subjective comparison and evaluation of speech enhancement algorithms. Speech Commun. 49 (7), 588–601. http://dx.doi.org/10.1016/j.specom.2006.12.006.

Itoh, K., 1982. Analysis of the phase unwrapping algorithm. Appl. Opt. 21 (14) 2470–2470.

Kay, S.M., 1993. Fundamentals of statistical signal processing, volume i: Estimation theory. Prentice Hall.

Koutsogiannaki, M., Simantiraki, O., Degottex, G., Stylianou, Y., 2014. The importance of phase on voice quality assessment. Proceedings of the International Speech Communication Association Interspeech. pp. 1653–1657.

Krawczyk, M., Gerkmann, T., 2014. STFT Phase reconstruction in voiced speech for an improved single-channel speech enhancement. IEEE Trans. Audio, Speech Lang. Process. 22 (12), 1931–1940.

Krawczyk, M., Gerkmann, T., 2016. On MMSE-based estimation of amplitude and complex speech spectral coefficients under phase-uncertainty. IEEE Trans. Audio Speech Lang. Process. 24 (12), 2251–2262.

Kulmer, J., Mowlaee, P., 2015a. Harmonic phase estimation in single-channel speech enhancement using von mises distribution and prior SNR. Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5063–5067.

Kulmer, J., Mowlaee, P., 2015b. Phase estimation in single channel speech enhancement using phase decomposition. IEEE Signal Process. Lett. 22 (5), 598–602.

Kulmer, J., Mowlaee, P., Watanabe, M., 2014. A probabilistic approach for phase estimation in single-channel speech enhancement using von mises phase priors. Proceedings of the IEEE Workshop on Machine Learning for Signal Processing. pp. 1–6.

Loizou, P.C., Kim, G., 2011. Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions. IEEE Trans. Audio Speech Lang. Process. 19 (1), 47–56.

Mannell, R.H., 1990. The effects of phase information on the intelligibility of channel vocoded speech. Proceedings of the Third Australian International Conference on Speech Science and Technology, Melbourne.

Mayer, F., Williamson, D.S., Mowlaee, P., Wang, D., 2017. Impact of phase estimation on single-channel speech separation based on time-frequency masking. J. Acoust. Soc. Am. 141 (6), 4668–4679.

Mowlaee, P., Kulmer, J., 2015a. Harmonic phase estimation in single-channel speech enhancement using phase decomposition and SNR information. IEEE Trans. Audio Speech Lang. Process. 23 (9), 1521–1532.

Mowlaee, P., Kulmer, J., 2015b. Phase estimation in single-channel speech enhancement: limits-potential. IEEE Trans. Audio Speech Lang. Process. 23 (8), 1283–1294.

Mowlaee, P., Kulmer, J., Stahl, J., Mayer, F., 2016a. Phase-Aware signal processing in speech communication: History, Theory and Practice. John Wiley & Sons.

Mowlaee, P., Saeidi, R., 2013. Iterative closed-loop phase-aware single-channel speech enhancement. IEEE Signal Process. Lett. 20 (12), 1235–1239.

Mowlaee, P., Saeidi, R., 2014. Time-frequency constraints for phase estimation in single-channel speech enhancement. Proceedings of the International Workshop on Acoustic Signal Enhancement. pp. 338–342.

Mowlaee, P., Saeidi, R., Stylianou, Y., 2016b. Advances in phase-aware signal processing in speech communication. Speech Commun 81, 1–29.

Mowlaee, P., Stahl, J., Kulmer, J., 2017. Iterative joint map single-channel speech enhancement given non-uniform phase prior. Speech Commun 86, 85–96.

Mowlaee, P., Watanabe, M., 2013. Iterative sinusoidal-based partial phase reconstruction in single-channel source separation. Proceedings of the International Speech Communication Association Interspeech. pp. 832–836.

Nikias, C.L., Mendel, J.M., 1993. Signal processing with higher-order spectra. IEEE Signal Process Mag. 10 (3), 10–37.

Oppenheim, A.V., Lim, J.S., 1981. The importance of phase in signals. Proc. IEEE 69 (5), 529–541.

Paliwal, K., Schwerin, B., Wójcicki, K., 2011. Role of modulation magnitude and phase spectrum towards speech intelligibility. Speech Commun. 53 (3), 327–339.

Pirolt, M., Stahl, J., Mowlaee, P., Vorobiov, V.I., Barysenka, S.Y., Davydov, A.G., 2017. Phase estimation in single-channel speech enhancement using phase invariance constraints. Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5585–5589. http://dx.doi.org/10.1109/ICASSP.2017.7953225.

Rix, A.W., Beerends, J.G., Hollier, M.P., Hekstra, A.P., 2001. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. 2. pp. 749–752.

Saratxaga, I., Hernaez, I., Erro, D., Navas, E., Sanchez, J., 2009. Simple representation of signal phase for harmonic speech models. Electron. Lett. 45 (7), 381–383.

Seetharaman, S., Jernigan, M.E., 1988. Speech signal reconstruction based on higher order spectra. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing. 1. pp. 703–706.

Sletten, C. J., Schell, A. C., Mack, R. B., Goggins, W. B., Blacksmith, P., 1973. Radar phase comparison method and system for object recognition. US 3725917.

Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J., 2011. An algorithm for intelligibility prediction of time-frequency weighted noisy speech. IEEE Trans. Audio Speech Lang. Process. 19 (7), 2125–2136.

Tatarskii, V., 2004. On the possibility of measuring the phase velocity, group velocity, and dispersion parameter of surface waves by means of coherent amplitude-modulated RADA. J. Electromagn. Waves Appl. 18, 429–435.

Totsky, A., Zelensky, A., Kravchenko, V., 2014. Bispectral Methods of Signal Processing. Applications in Radar, Telecommunications and Digital Image Restoration. De

Gruyter.

Van Trees, H.L., 2004. Detection, Estimation, and Modulation Theory. John Wiley and Sons.

Varga, A., Steeneken, H.J.M., Tomlinson, M., Jones, D., 1992. The NOISEX–92 study on the effect of additive noise on automatic speech recognition. Technical Report. DRA Speech Research Unit.

Vary, P., 1985. Noise suppression by spectral magnitude @@estimation mechanism and theoretical limits. Elsevier Signal Process. 8 (4), 387–400.

Vorobiev, V.I., Davydov, A.G., 2008. Complex cepstrum and inter-component processing of speech. Proceedings of the Twentieth Session of the Russian Acoustical Society. 3. Russian Acoustical Society, pp. 9–12.

Vorobiov, V.I., 2006. Inter-component phase processing of speech signals for their recognition and identification of announcers. Proceedings of the Eighteenth Session of the Russian Acoustical Society. 3. Russian Acoustical Society, pp. 48–51.

Vorobiov, V.I., Barysenka, S.Y., 2014. Application of inter-component phase processing methods in non-stationary vibration analysis (in Russian). Proceedings of the Twenty-Seventh Session of the Russian Acoustical Society. pp. 1–6.

Vorobiov, V.I., Davydov, A.G., 2012. Study of the relations between quasi-harmonic components of speech signal in Chinese language. Proceedings of the Twenty-Fifth Session of the Russian Acoustical Society. 3. Russian Acoustical Society, pp. 11–14.

Vorobiov, V.I., Davydov, G.V., Shamgin, Y.V., 2006. Phase relation between fundamental tones and vowel sounds obertones (in Russian). Reports of the Belarusian State University of Informatics and Radioelectronics (BSUIR) 14 (2), 64–68.

Vorobiov, V.I., Klimov, A.V., 1986. Phase characteristics of objects reflection in multiple frequency radar with multipath propagation (in russian). J. Radioeng. 2, 19–22.

Vorobiov, V.I., Klimov, A.V., Sirotko, I.I., 1986. Analysis of phase characteristics of multiple frequency hydroacoustic signals with dual path propagation (in Russian). Proceedings of the Reports of the Fourteenth All-union statistical Hydroacoustics workshop. 1. Acoustic Institution of the USSR Academy of Sciences, pp. 105–107.

Wang, D., Lim, J., 1982. The unimportance of phase in speech enhancement. IEEE Trans. Audio Speech Lang. Process. 30 (4), 679–681.

Wells, B., 1985. Voiced/unvoiced decision based on the bispectrum. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. 10. pp. 1589–1592.

Zverev, V.A., 1953. Modulation method of ultrasonic dispersion measurements (in Russian). Pap. USSR Acad. Sci. 91/4, 791–794.

Zverev, V.A., 1956. Modulation method of ultrasonic dispersion measurements (in russian). Acoust. Phys. 2 (2), 142–145.