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

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):

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

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:

$$\Theta = \phi_0 - \frac{\phi_1 + \phi_2}{2}. \quad (2)$$

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:

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

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