Advances in Phase-Aware Signal Processing in Speech Communication

Pejman Mowlaee∗†, Rahim Saeidi†, Yannis Stylianou‡

∗Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria
†Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland
‡Computer Science Dept. University of Crete, Crete, Greece

Abstract
During the past three decades, the issue of processing spectral phase has been largely neglected in speech applications. There is no doubt that the interest of speech processing community towards the use of phase information in a big spectrum of speech technologies, from automatic speech and speaker recognition to speech synthesis, from speech enhancement and source separation to speech coding, is constantly increasing. In this paper, we elaborate on why phase was believed to be unimportant in each application. We provide an overview of advancements in phase-aware signal processing with applications to speech, showing that considering phase-aware speech processing can be beneficial in many cases, while it can complement the possible solutions that magnitude-only methods suggest. Our goal is to show that phase-aware signal processing is an important emerging field with high potential in the current speech communication applications. The paper provides an extended and up-to-date bibliography on the topic of phase aware speech processing aiming at providing the necessary background to the interested readers for following the recent advancements in the area. Our review expands the step initiated by our organized special session and exemplifies the usefulness of spectral phase information in a wide range of speech processing applications. Finally, the overview will provide some future work directions.

Keywords: Phase-aware speech processing, phase-based features, signal enhancement, automatic speech recognition, speaker recognition, speech synthesis, speech coding, speech analysis.

1. Introduction
In many everyday-life applications a reliable speech communication system is in demand whose performance is expected to be robust enough to deliver a certain high quality of service to establish a reliable speech communication medium. For this purpose, it is highly important to verify the robustness of the underlying speech communication interface against impairments that occur due to some acoustic interference, background noise or those introduced by the failure in the communication channel, which can be modeled, for example, as the additive background noise or reverberations in the room, respectively. In a speech communication application, the design goal is to deliver enough flexibility and reliability such that it provides a certain quality of service that may differ depending on the specified application. A full end-to-end speech communication chain from microphone to receiver end involves several blocks including: speech analysis, multi-channel processing (beamformer), single-channel signal restoration entailing speech enhancement, source separation and artificial bandwidth extension. Depending on the desired application, one might be interested in playing back the reconstructed signal at the receiver end and, as a consequence, a speech synthesis block is required. Alternatively, in biometric and dictation applications, the goal is to recognize, or verify the identity of the speaker (speaker identification/verification) or to recognize the phonemes and words spoken by the transmitting party (automatic speech recognition). Another application is speech watermarking, for which certain cover data for copyright issues are required to be embedded in speech without modifying the speech quality of the watermarked

∗Corresponding author. Tel.: +43 316 873 4433; fax: +43 316 873 7941.
E-mail addresses: pejman.mowlaee@tugraz.at (P. Mowlaee), rahim.saeidi@aalto.fi (R. Saeidi), yannis@csd.uoc.gr (Y. Stylianou).
signal still be robust against spoofing attacks. Finally, in speech coding the phase information is important to fulfill the bits budget while delivering a transparent speech quality at the receiver end. This is feasible by keeping the phase quantization errors below a perceptual threshold, determined by just noticeable difference, not perceptible by the listener.

The common strategies described in the aforementioned speech processing applications mainly focus on spectral amplitude modification and signal processing methods that are used to filter the spectral amplitude of the speech signal. More recently, the research attention towards incorporating the spectral phase information has increased. In particular, individual steps have been taken by researchers through different applications from different sub-communities in speech signal processing. In order to unify the attempts made by the researchers in such diverse communities and to gather them together to share their findings, a special session at INTERSPEECH 2014 entitled Phase Importance in Speech Processing Applications [1] has recently been organized.

In this paper, we present an extended literature review of earlier viewpoints about the importance of phase information in speech signal processing. We exemplify the importance of phase information in several speech processing applications and further demonstrate the positive impact brought about by phase-aware signal processing in each of the aforementioned applications. Finally, recent and future directions towards applying phase-aware signal processing to tackle different speech applications are pointed out. Our work extends the initial step made at special session [1]. In contrast, here we present a broader picture of phase-aware signal processing and provide a detailed overview of the advances made by researchers in the field. Following the literature review, we present several useful representations for phase spectrum that illustrate certain structures in the phase domain. The common short-time Fourier transform (STFT) spectral phase representation exhibits no useful structure or distribution in the phase domain due to its cyclic wrapping nature. However, new presentations provide means to explore various properties and details of speech signals (e.g., continuity, formant, harmonicity, or statistical explanation captured by mean, variance and probability density function). We then describe the phase-based derived features appearing in the literature for different speech applications. To highlight the importance of phase-aware signal processing, we give examples in several speech applications: single-channel and multi-channel speech enhancement, source separation, automatic speech recognition, speaker recognition, speech coding, speech analysis/synthesis, digital speech watermarking, and speech quality estimation. For each of these applications, we present the impact of phase information and explain the state-of-the-art methods that neglect the phase information, as well as more recent ones that incorporate phase information into their proposed solutions.

Several recent attempts have presented partial overviews of some aspects of phase processing for speech applications [2–5]:

- In [2], Alsteris and Paliwal reviewed experimental results for short-time phase spectrum usage in speech processing. This work demonstrated the usefulness of phase spectrum for automatic speech recognition, human listening and speech intelligibility. In [6], Paliwal et al. studied the importance of phase information for speech enhancement and proposed a phase spectrum compensation (PSC) method using the conjugate symmetric property of the discrete Fourier transform (DFT). These subjects are covered in Section 4.1.3.

- In [3], Gerkmann et al. reviewed phase processing for single-channel speech enhancement. In contrast to this narrow focus on phase-aware speech enhancement, the current review presents a unified approach to phase-aware speech communication including other applications such as speech analysis, speech synthesis, speaker/speech recognition, speech coding, speech watermarking, source separation and multi-channel speech processing. These subjects are covered in Section 4.

- In [4, 5], Mowlaee and Kulmer reviewed phase estimation in speech enhancement and demonstrated the potential and limits of the phase estimation methods demonstrated with a comparative study. The main contribution in [4, 5] was on the phase estimation in noise and incorporate it for improved signal reconstruction. Efficient techniques for estimating phase in noisy environments are reviewed in Section 4.1.3.

In this paper, throughout the extended up-to-date bibliography we provide a guidance for our readers to a summary of what has been done with regard to phase processing of speech signal. For readers more interested in the specific solution, we refer to bibliography where a list of early to latest contributions can be found. Our goal is to help the interested readers to have a quick access to the novel papers, and get to know the earlier and current innovations and
efforts made to explore phase-aware signal processing. We also elaborate on why phase was believed to be unimportant in each application and why phase processing is recently attracting increasing interest in speech processing community. The usefulness of spectral phase information will be exemplified in a wide range of speech processing applications (speech analysis/synthesis, speech enhancement, source separation, speech coding, speech quality estimation, speech watermarking, automatic speech recognition, and speaker recognition), providing a reference for the researchers who are just starting with their research topic related to the phase-aware signal processing. The current paper is an extensive overview to date on the topic of phase-aware speech processing and provides a unique picture that depicts the importance of phase assessment in many aspects of speech technology.

The paper is organized as follows: section 2 presents the controversial standpoints in the literature with regard to the importance of phase information in speech signal processing; section 3 presents useful phase representations derived from the short-time Fourier transform phase; It also focuses on phase-based features used in speech processing applications; Section 4 we exemplify the importance of phase-aware signal processing in speech applications; and finally, section 5 concludes the work and presents a discussion.

2. History on Phase Processing

In 1843 [7, 8], Helmholtz and Ohm concluded from their experiments that the human ears are insensitive to phase perception. Several studies later demonstrated that phase information could contribute to the intelligibility of speech signals, disproving the earlier observations. For example, Schroeder [9] reviewed hearing models, studying the importance of the phase spectrum in human perception. Oppenheim and Lim [10] explored the importance of phase in signals. In [10–12], Oppenheim et al. also demonstrated the contribution of phase-only signal reconstruction on human perception with regard to the quality of the reconstructed signal, exemplified for audio, image, crystallographic and speech processing applications. Liu et al. [13] studied the perceptual importance of the phase spectrum specifically on the intervocalic stop consonants. They reported the strong dependency of the perception of intervocalic stops on phase information.

As another example that highlights the importance of phase information, the authors in [14] presented the statistical interpretation of phase information in signal and image reconstruction. Oppenheim et al. in 1983 [11] demonstrated signal synthesis and proposed signal reconstruction techniques using only partial information in the Fourier domain. In more detail, they investigated the conditions for reconstructing a signal only from its available partial information in the Fourier-domain. They also reported results on the exact signal reconstruction from partial STFT information, i.e., signed-magnitude-only or phase-only. Same authors then studied the conditions under which a signal can be reconstructed from one-bit phase-only and successfully applied them for image reconstruction [15]. Several studies later estimated a time-domain signal from its magnitude information of the short-time Fourier transform as well as studied how to use the phase-only information to achieve a unique reconstruction of a time-domain signal [15–17].

More specifically, for speech signal processing, Paliwal et al. demonstrated the importance of phase information in human listening [18, 19] and speech signal processing [2] exemplifying its contribution in the context of automatic speech recognition (ASR) [20], speech intelligibility [21], speech analysis [22, 23], and speech enhancement [6]. In multi-channel signal processing, phase difference information between sensors has been used to estimate direction of arrival [24, 25]. An overview on microphone array speech processing was presented by Brandstein and Ward in [26] where the phase difference between channels was shown as the main source of information for time-delay estimation and source localization applications. In [27], Aarabi et al. reviewed phase-based speech processing methods that were focused on improved automatic speech recognition or binaural speech enhancement using phase difference information between microphones.

The importance of phase information in noise reduction has been controversial. Wang and Lim [28] concluded that phase information was negligible in speech enhancement. Vary [29] derived the relationship between the phase distortion threshold and local signal-to-noise ratio to be 6 decibels, above which the noisy phase can be considered to be a decent estimate for the clean phase spectrum and the roughness in the synthesized speech is not audible. Ephraim and Malah [30] showed that the noisy spectral phase is the minimum mean square error (MMSE) estimate for the clean phase, assuming a uniform prior distribution on the phase. Similarly, Lotter and Vary showed that the noisy phase is the maximum a posteriori (MAP) estimate of the clean phase under the same assumptions [31]. Recently, a short review on phase importance in different speech processing applications was made [1]. Several scattered papers describe phase-aware processing mechanisms in specific applications such as: phase-aware processing in single-channel
speech enhancement history and advances [3], phase estimation in noise and its limits and potential in single-channel speech enhancement [4, 5], phase estimation for single-channel source separation [32], speech synthesis [33, 34], and automatic speech recognition [2].

One common observation in the literature with regard to the importance of phase in speech processing is the fact that the positive contribution of short-time Fourier phase spectrum becomes larger once an analysis window of a relatively long length is chosen for the spectral Fourier analysis of speech [35, 36]. The origin of the previously observed controversy was mainly due to the fact that, as reported in [13] the phase holds negligible information about speech intelligibility when a short window analysis is applied, e.g., 20 to 40 ms. This is in contrast to conventional magnitude processing where common signal processing for speech employs the short window analysis, leading to meaningful features e.g., spectral amplitude for speech enhancement or Mel frequency cepstral coefficients (MFCC) for automatic speech/speaker recognition.

3. Phase Analysis

When analyzing speech phase, we first need to mention that the phase term has been used in different contexts with different meanings. Table 1 provides a list of different phase terms (named on the first column) used throughout the speech signal processing literature found in different applications (exemplified on the second column). We explain each specific phase representation in the following Sections.

<table>
<thead>
<tr>
<th>Phase Representation</th>
<th>References</th>
<th>Related Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>STFT phase</td>
<td>signal enhancement [6, 37, 38], signal reconstruction [4, 5, 17, 32, 39, 40]</td>
<td>Section 4.1.7 and Section 4.1</td>
</tr>
<tr>
<td>inter-microphone phase</td>
<td>multi-channel noise reduction [26, 27, 41, 42], time-delay estimation [24]</td>
<td>Section 4.1.1</td>
</tr>
<tr>
<td>phase of the vocal tract filter</td>
<td>speech synthesis [33], glottal model estimation [43–45]</td>
<td>Section 4.6</td>
</tr>
<tr>
<td>phase in harmonic domain</td>
<td>speech synthesis [33, 46, 47], phase estimation [39, 48, 49]</td>
<td></td>
</tr>
<tr>
<td>phase in perceptual bands</td>
<td>speech coding [50, 51], speech watermarking [52–54]</td>
<td>Section 4.2 and Section 4.3</td>
</tr>
</tbody>
</table>

Table 1: The term phase has different meanings and usages in speech processing.

3.1. Fourier Analysis of Speech Signals

Let \( x(n) \) represent a time-domain speech signal with \( n \) being the time sample index. Having segmented the signal into short frames of length \( N \), we define \( x_l(n) \) as the short-time windowed segment with \( n \in [0, N - 1] \). Taking the Fourier transformation of \( x_l(n) \) we obtain

\[
X(k, l) = \mathcal{F}(x_l(n)),
\]

where \( X(k, l) \) is the DFT coefficient for the frequency bin index \( k \) and frame index \( l \). The complex DFT value can be inspected through its magnitude and phase representations given by

\[
X(k, l) = |X(k, l)| e^{j\phi_x(k, l)},
\]

with \( |X(k, l)| \) and \( \phi_x(k, l) = \angle X(k, l) \) representing the spectral amplitude and the instantaneous phase spectrum, respectively. While the STFT magnitude spectrum is known to follow a super-Gaussian distribution (see e.g., [55]), the spectral phase has often been reported to follow a uniform distribution.

Figure 1 (first column) shows the STFT representation for the magnitude (left) and phase (right) spectra. Unlike the magnitude spectrum, the phase spectrum provides no useful structure. In order to have a proper access to phase information, a similar clear harmonics structure in terms of speech formants and harmonicity as presented by the spectrogram could be quite helpful². The plots are shown for a female utterance from GRID corpus [56]. The utterances in this corpus are of command like structure consisting of six units, including a color, letter, and number. The chosen utterance for our experiment shown here is “bin blue at l four soon” consisting of a plosive [b], vowel [U:], and fricative [s].

In the following, we provide alternative phase representations that bring a more useful structure and are of more help when performing a phase-aware speech processing.

²Some implementations for the figures and audio samples are available at https://www.spsc.tugraz.at/PhaseLab
3.2. Phase Derivatives across Time and Frequency

Extracting useful features from a Fourier phase spectrum is not straightforward. This is due to the difficulties in phase wrapping, the dependency of the phase spectrum on window starting sample and rapid changes in the phase spectrum when the zeros of complex spectrum $X(z)$ lie near the unit circle in $z$-plane. The wrapping issue in the STFT analysis leads to cyclic wrapping that masks any useful structure in the spectral phase. In the literature, phase unwrapping methods [57] has been studied with a recent overview presented in [58]. In phase-based feature extraction, two important definitions play a major role in finding relevant features; group delay and instantaneous frequency. The group delay describes the local time behavior as a function of frequency. In a dual representation, the instantaneous frequency characterizes a local frequency behavior as a function of time. Taking the derivative of the spectral phase across frequency and time de-emphasizes the wrapping issue of phase and reveals the harmonic structure (for example see the group delay plots shown in Figure 2 and Figure 3).

The group delay representation has been long considered as an alternative method for spectrum estimation [59, 60]. The calculation of the group delay requires a continuous function, hence unwrapped phase. As a solution for the wrapping problem in calculating the group delay function, an averaging method was proposed in [61].

The first derivative of the instantaneous phase with respect to frequency called group delay (GD) is given by:

$$\tau(k, l) = -\Delta_k(\phi(k, l)), \quad (3)$$

where $\Delta_k$ is the discrete differentiator across frequency. Group delay representation has been reported as a useful analysis tool in sound classification [62], spectrum estimation [59], speech recognition [63], speaker identification [64], signal reconstruction [65], speech segmentation [66], waveform estimation [61], and formant extraction [67]. Stark and Paliwal proposed group delay deviation (GDD) as a useful feature for speech analysis [68]. The concept was also employed to estimate spectral phase of a desired signal in noise or a mixture of speakers [69, 70].

As an alternative for phase representation in the STFT domain, one can consider the first time-derivative called instantaneous frequency (IF), [71] given by

$$\text{IF}(k, l) = \text{princ}\{\phi(k, l) - \phi(k, l - 1)\}, \quad (4)$$

where princ[·] is the principal value operator that maps an input value to the range in $[-\pi, \pi]$. The instantaneous frequency has been reported to be useful for formant detection [72], speaker recognition [73], and source separation [74]. Furthermore, the authors in [22] proposed IF deviation (IFD). The IFD has been shown useful representation for speech analysis [22] as well as speech quality estimation [75, 76]. For more details on instantaneous frequency and different methods for its estimation, we refer to [77].

Speech signal is, in general, a mixed phase type of signal. In terms of its analysis, the group delay of a mixed phase signal is the sum of the group delay of its minimum phase and maximum phase components [65]. It is common to work with only the minimum phase part of speech signal [16, 66]. The definition of the group delay function as the derivative of the phase spectrum, results in a tight connection to the amplitude spectrum, implying that phase and amplitude spectrum convey complementary information about the underlying signal. By dropping the frame index $l$,
for a short-time signal \( x(n) \) of length \( N \), we have

\[
X(z) = \sum_{n=0}^{N-1} x(n)z^{-n} = x(0)z^{-(N-1)} \prod_{m=1}^{N-1} (z - z_m),
\]

\( X(\omega) = X(z)|_{z=e^{j\omega}} = |X(\omega)|e^{j\phi(\omega)}, \)

\[
\log(X(\omega)) = \log|X(\omega)| + j\phi(\omega),
\]

(5)

where \( z_m \) are the zeros of the signa. The group delay function in the continuous domain is then calculated as

\[
\tau_c(\omega) = -\tau(\omega) = -\left( \frac{d|X(\omega)|}{d\omega} \right) = \frac{X_R(\omega)D_R(\omega) + D_I(\omega)X_I(\omega)}{|X(\omega)|^2},
\]

(8)

where the \( R \) and \( I \) indices indicate the real and imaginary parts of a complex variable, respectively. The first derivative of \( X(\omega) \) with respect to \( \omega \) is shown as \( D(\omega) \). The zeros of \( X(\omega) \) can cause instability in group delay function. In order to deal with spurious peaks in the resulting group delay function (dips of amplitude spectrum), it was suggested to use a cepstrally smoothed version of \( |X(\omega)| \), called \( S(\omega) \), in the denominator of \( \tau(\omega) \) and further proceed with defining the modified group-delay (MGD) given by [78]

\[
\tau_m(\omega) = \frac{X_R(\omega)D_R(\omega) + D_I(\omega)X_I(\omega)}{|S(\omega)|^2},
\]

\[
\tau_m(\omega) = \left( \frac{\tau(\omega)}{|\tau(\omega)|} \right)^\gamma,
\]

(10)

where \( \gamma \) and \( \alpha \) are the tune parameters. Placing the group delay function in the form of \( \tau_m(\omega) \) is an established way to extract phase-based features for ASR and speaker recognition. A conventional way of extracting meaningful features from group delay is to convert them to cepstral representations [64] using discrete cosine transformation on group delay.

Several modifications of the group delay calculation are proposed in [81] to deal with the zeros of complex spectra when preparing phase-based features for ASR. Alternatively, it is possible to reduce the effect of zeros by smoothing the phase spectrum of a mixed-phase speech signal before treating at the group delay function and next performing cepstral smoothing [2]. Another method is to treat the group delay representation as the power spectrum and pass it through the Mel-scale filterbank before cepstral transformation [81]. Unlike the common method of application of log compression for Mel filterbank energies, the significance of log (root) compression in processing of the group delay representation is controversial. In the search for alternative ways to deal with dips in the amplitude spectrum in [82],

---

**Figure 2:** Group delay representation of a voiced speech segment. Three group delay functions are presented along with Fourier spectrum (FFT) and LP spectrum (LPC): modified group delay (MGD) [78], Chirp group delay (CGD) [79] and LP group delay (LPGD) [80] methods provide different properties of a group delay function, which can be utilized efficiently in specific recognition applications.
a product spectrum was used whereby the power spectrum and group delay were multiplied.

It has been shown that windowing is crucial to obtain reliable group delay functions for speech analysis [81, 83]. With a proper choice of window function [84], i.e., efficient treatment of zeros in the amplitude spectrum, group delay functions reveal formant structure more clearly. Zeros close to the unit circle in the \( z \)-plane result in spikes in the group delay function and consequently, mask the vocal tract information present in group delay function [85]. In order to obtain a spike-free group delay representation, a chirp group delay (CGD) was proposed in [79]. The CGD was defined as the negative derivative of the phase spectrum computed from chirp \( z \)-transform, that is \( z \)-transform computed on a circle/spiral other than the unit circle. Finally, in [80], the group delay function was calculated using the phase of a parametric all-pole modeling rather than the Fourier spectrum. To this end, the speech spectrum is decomposed into minimum-phase and all-pass components. The group delay is then extracted from the minimum-phase part which has been reported to be suitable for speaker recognition.

A demonstration of selected group delay representations is presented in Figure 2. The Time-frequency representation for group delay shown in Figure 3 (third column) reveals a harmonic structure in speech that follows the spectrogram shown in Figure 1 (first column).

### 3.3. Phase Representations in Harmonic Domain

McAulay and Quatieri [86] proposed sinusoidal model as an analysis/synthesis technique whereby a short-time segmented speech signal is represented as sum of sinusoids. While in sinusoidal model each sinusoid is represented by triple parameters of amplitude, frequency and phase, in case of harmonically related sinusoidal components, a harmonic model, as a special case of the more general sinusoidal model is used. The speech waveform is approximated by harmonics as follow:

\[
\hat{x}(n) = \sum_{h=1}^{H_t} A(h, l) \cos \phi_{x}(h, l),
\]

where \( h \) is the harmonic index, \( H_t \) is the number of harmonics at frame \( l \), and \( A(h, l) \) is the real-valued amplitude and \( \phi_{x}(h, l) \) is the phase both sampled at the \( h \)th harmonic, respectively. The instantaneous phase at harmonic \( h \), \( \phi_{x}(h, l) \) is wrapped between \([-\pi, \pi]\). Alternatively, a meaningful representation of phase is the unwrapped one where the phase changes continuously without jumps. Several different methods have been proposed to unwrap the phase of a signal (see, for example, [57, 87, 88]). A recent study by Drugman and Stylianou [58] showed a comparative evaluation of the existing phase unwrapping methods in terms of their accuracy and computational complexity. Furthermore, they proposed an efficient iterative phase unwrapping method that relied on the calculated number of zeros of the \( z \)-transform outside the unit circle.

Given a fundamental frequency at the \( l \)th frame denoted by \( f_0(l) \), a harmonic model plus phase distortion signal analysis/synthesis was proposed in [89]. The authors proposed a pitch-synchronous segmentation with time instant \( t(l) = t(l-1) + \frac{1}{\pi f_0(l)} \), whereby the instantaneous phase at harmonics was decomposed into a linear and unwrapped phase part as:

\[
\phi_{x}(h, l) = 2\pi h \sum_{l'=0}^{l} f_0(l') (t(l') - t(l'+1)) + \angle V(h, l) + \psi_{\phi}(h, l).
\]

The first term is the linear phase component, which depends only on the fundamental frequency at \( l \)th frame \( f_0(l) \). Degottex and Erro gave a detailed analysis on phase distortion analysis [89]. The second term \( \angle V(h, l) \), is the phase response of the vocal tract filter denoted by \( V(h, l) \), known to have a minimum phase response, assuming an all-pole model. The phase response of an envelope filter can be estimated through the Hilbert transform that links the phase response and its magnitude for a minimum-phase signal through cepstral representation (see [90] for more details). The last term, called source shape \( \psi_{\phi}(h, l) \), represents the pulse shape and captures the stochastic character of the phase spectrum of the speech signal. The term was also called phase dispersion used in speech coding using a vector quantizer [91] and was modeled using wrapped Gaussian distribution [92]. It is important to note that the linear phase removal results in an unwrapped phase composed of the minimum phase and dispersion phase parts. The unwrapped phase has been also used for phase-based speech enhancement [4, 5, 48].

Comparing the plots for magnitude spectrogram versus the instantaneous phase spectrum shown in Figure 1 reveals that the instantaneous STFT phase exhibits only a random pattern without a clear harmonic structure due to the
wrapping issue of the STFT phase. In fact, phase wrapping has been the main reason that phase-based signal processing has been considered less often in the literature on speech signal processing. As a consequence, the phase-based signal processing is believed to be more troublesome than signal processing methods relying on spectral amplitude only. As alternative useful representations, in the following, we highlight three recent representations, derived from harmonic phase \( \phi_x(h, l) \) which provide useful insights into phase interpretation and show a more useful structure, hence, help when performing phase-aware speech processing.

### 3.3.1. Phase Distortion (PD)

Phase distortion (PD) was first proposed by Degottex [93], who calculated it by removing the contributions of minimum phase and linear phase parts from the instantaneous phase as:

\[
PD(h, l) = \phi_x(h + 1, l) - \phi_x(h, l) - \phi_x(1, l).
\]

(13)

Phase distortion represents the shape of the glottal signal [93, 94] and the glottal model parameters [44]. As reported in [89], phase distortion can be statistically modeled using a wrapped Gaussian distribution determined by circular mean and variance parameters. As the mean value of phase was shown to be perceptually negligible in [95], Degottex et al. proposed to characterize the short-time phase features of phase distortion in terms of its standard deviation only [89] where the phase distortion is modeled as:

\[
PD(h, l) = W_G(0, \text{PDD}(h f_0(l), l)).
\]

(14)

In (14), \( \text{PDD}(h f_0(l), l) \) captures the variance of phase changes at the \( h \)th harmonic multiple [89] and \( W_G(\cdot) \) denotes the wrapped Gaussian distribution characterized by circular mean and variance parameters. In Figure 3 (fifth column), the PDD representation shows a binary pattern between low-mid harmonics with low PDD and levels a randomized pattern for high frequencies.

### 3.3.2. Relative Phase Shift (RPS)

Saratxaga [46] proposed relative phase shift (RPS) that relates the phase shift between the \( h \)th harmonic and the one at the fundamental frequency \( h = 1 \) given by:

\[
\text{RPS}(h f_0, l) = \phi_x(h f_0, l) - h \phi_x(f_0, l).
\]

(15)

As shown in [89], the RPS representation discards the linear phase contribution as only the source shape and minimum-phase envelope phase contributions remain at each harmonic relative to those of the first harmonic. Because both envelope and source shape are assumed to evolve smoothly over time, the RPS is known for its smooth trend across frames [46], without the requirement of estimating the pitch pulse onset. The RPS representations and derived features were used for spoofing detection in speaker recognition [34, 96]. More recently, the RPS representation was used for frequency-smoothing during phase estimation in single-channel speech enhancement [4, 97]. In Figure 3 (fourth column), the RPS uncovers the harmonic structure in the phase spectrum, which is not visible in the instantaneous phase. The RPS of clean signal displays small changes with a smooth trend at harmonics in a clean signal.

### 3.3.3. Unwrapped Harmonic Phase

Following harmonic model phase decomposition given in (12), by removing the linear phase component (defined as the integral part of the instantaneous frequency) from the instantaneous harmonic phase, an unwrapped harmonic phase is obtained

\[
\Psi(h, l) = \angle V(h f_0) + \psi_d(h, l),
\]

(16)

known for its smooth change across time [89] with \( \psi_d(h, l) \) as the phase dispersion already defined in (12). This property of unwrapped phase has been used to derive several interpolation techniques used in speech synthesis [98] using the harmonic model of speech signals. Recently, Agiomyrgiannakis in [99] presented an overview on several interpolation techniques applied to harmonic phase for speech synthesis. These interpolation techniques comprised of quadratic phase spline and phase-locked pitch-synchronous and were demonstrated to deliver a high quality synthesized speech quality when used in a harmonic signal modeling framework. Furthermore, using the smoothness property of the unwrapped phase in harmonics, temporal smoothing filters were proposed to enhance noisy speech in a single-channel
Figure 3: Time-frequency representations, instantaneous phase plots, from left to right: STFT phase, phase variance, group delay, relative phase shift (RPS), phase distortion standard deviation (PDD) shown for a female speech saying “bin blue at l four soon”. While the instantaneous phase presents no useful structure, both phase variance (second panel) and group delay (third panel) present a harmonic structure. At voiced segments, the shape of the glottal pulse smoothly changes justified by a close to zero PDD pattern [33, 95] and presents a randomized structure in non-harmonic or unvoiced segments (fourth panel). Finally, RPS shows stable patterns during the vowels showing the evolution along time of the RPS for each harmonic (fifth panel).

scenario. Certain improvements in perceived quality and speech intelligibility have been reported by reducing the phase variance of the noisy signal (see e.g., [5, 48, 97]).

A non-uniform representation of the unwrapped phase in harmonics has been employed to achieve a meaningful statistical representation of the underlying distribution of the harmonic phase features. To this end, the authors in [48, 97] proposed to fit a von Mises distribution [100, p. 191] to harmonic phase given as

\[ \text{VM} (\mu_c(h, l), \kappa(h, l)) = e^{\kappa(h, l) \cos(\Psi(h, l) - \mu_c(h, l))} / 2 \pi I_0(\kappa(h, l)), \] (17)

where \( I_\eta(\cdot) \) is a modified Bessel function of order \( \eta \) with \( \mu_c(h, l) \) and \( \kappa(h, l) \) denoting the mean direction (also called circular mean) and concentration parameters in the von Mises distribution, respectively. It is important to note that the von Mises distribution as phase prior ranges between a uniform distribution (maximum uncertainty) when \( \kappa = 0 \) and a Dirac delta (maximum deterministic) when \( \kappa = \infty \). The concentration parameter \( \kappa \) plays the key role to deliver the uncertainty around a given mean direction and is estimated indirectly from the phase variance using a non-linear equation in terms of Bessel functions [100]. Figure 3 (second column) shows the phase variance in time-frequency for a clean speech signal, revealing a harmonic structure similar to spectrogram shown in Figure 1 (first column).

4. Applications

4.1. Signal Enhancement

Signal enhancement is itself composed of two sub-groups: speech enhancement and source separation. While in the first category, one is interested in recovering the desired signal only by filtering out the undesirable interfering noise, in the latter category we are interested to recover all the signals in the mixture. Depending on the number of microphones used during the process of signal enhancement, two sub-groups are often considered: multi-channel and single-channel scenarios.

4.1.1. Multi-Channel Signal Enhancement

Array processing was first used in sonar and radar systems that were designed to detect and analyze narrowband signals. Extensions to speech signal processing (wideband signals) started by Brandstein and Ward, which were extended to microphone array processing for many speech applications, included noise reduction, acoustic source localization, source separation and target tracking [26].

In case of more than one microphone, the spatial diversity could be added in order to improve the signal acquisition in front-end unit. Microphone array processing allows for extracting the spatial and statistical characteristics of the signals. This is feasible by a combination of microphone signals in the array carried out with a beamformer. A beamformer is designed such that a desired spatial selectivity is met while the undesired sources from other directions are suppressed. The simplest beamformer is given by adding up the microphone signals, called delay-and-sum beamformer\(^3\). The delay is approximated as phase shifts instead of delay lines and the resulting beamformer is

\(^3\)as a special example of the more generalized filter-and-sum beamformer
referred to as phased-array, implemented using phase shifters [25, 35].

Let the phase of the harmonics arriving at the first and the second microphone be represented by $\phi_1$ and $\phi_2$ and we further define $c$ as the speed of sound, $d$ the inter-spacing between microphones, and $\theta_c$, the direction of arrival. Therefore, the phase shift between the two microphones is given by:

$$\Delta \phi_d = \phi_1 - \phi_2 = 2\pi f \Delta \tau = \frac{d}{c} \sin(\theta_c),$$

where $\Delta \tau = \frac{d \sin(\theta_c)}{c}$ is defined as the time delay of arrival. The phase shift defined in (18) is often used instead of time delay in sensory array processing (also called phased-array in radar [101] and antenna array processing [25, 35]), where $\beta = \frac{d}{c}$ is defined as the wave number with $\lambda$ as the wave length. The phase difference contains information about the direction of arrival which is useful for source localization applications [26].

Several cross-correlation based methods have been proposed in the literature and intensively used for the estimation of the time delay of arrival of a signal. The methods are referred to as generalized cross correlation (GCC) [24], where the time-delay estimate is given by searching for the peak of the cross-correlation function. The cross-correlation based on the magnitude spectrum is known as the basis for the GCC method. In order to filter out the spurious outlier peaks that often occur in the cross correlation, a GCC weighting function [24] or cepstral prefiltering step [102] has been recommended. In contrast, the cross-spectrum phase has been reported in [24] as a successful approach to resolve the peak for the dominant delay in a reverberant environment.

The use of phase information in microphone array processing has been mostly introduced for time-delay estimation, relying on the calculation the phase difference between the microphones, $\Delta \phi_d$ in (18). As an example, Aarabi et al. demonstrated robust speech enhancement [103] and improved automatic speech recognition [104]. Obaid et al. [41] also derived phase error filters, attempting to minimize noisy phase variance. They showed improved speech intelligibility in cochlear implant processing with more acceptable performance for cochlear implant recipients. Aarabi et al. demonstrated robust speech enhancement [103] and improved automatic speech recognition [104]. Obaid et al. [41] also derived phase error filters, attempting to minimize noisy phase variance. They showed improved speech intelligibility in cochlear implant processing with more acceptable performance for cochlear implant recipients. Aarabi et al. showed that their proposed phase-aware method preserves spatial cues contained in the phase difference for a clean non-reverberant speech signal, is equal after delay compensation. The phase difference was calculated as [105]:

$$\theta_l(\omega, l) = \phi_R(\omega, l) - \phi_L(\omega, l) - \omega \beta,$$

where the $R$ and $L$ are sub-indices refer to the right and left microphones in a binaural setup, $\beta$ refers to a selected time delay estimate upon which the phase difference is calculated. Aarabi et al., showed that the signal-to-noise ratio at each frame and frequency is bounded by a function of the phase error [103]. This principal was also used to derive a perceptually motivated phase-error filter for speech enhancement [103, 105]. In order to obtain the optimal phase transform (PHAT) they proposed a search to find the minimum phase error used to find the time-delay estimate between the left and right microphones. Given the optimal, a phase-mask was derived based on this error term. The mask, hence, was a phase-based time-varying filter, and was finally applied to either of the two microphones signals to get a signal with reduced noise. Later, Kim et al. [106] proposed a two-microphone approach where additional smoothing of phase estimates was performed over time and frequency by applying Gammatone channel weighting. This further improved the automatic speech recognition accuracy.

Deleforge and Kellerman [107] proposed a blind source separation method for an under-determined case where phase information was taken into account. The phase difference information carries the spatial information comprised of interaural time difference (ITD) and interaural level difference (ILD). Such spatial information are to be preserved due to the correlation in the speech-dominated signal components, which were dominated by the desired signal. Deleforge and Kellermann [107] generalized the K-singular value decomposition (KSVD) in terms of its sparse activation matrix in that, they proposed a sparse factorization relying on estimation of the instantaneous phase information. They showed that their proposed phase-aware method preserves spatial cues contained in the phase difference between the channels. The atoms in the dictionary are learned, using phase-optimized dictionary learning, where the source activations together with their corresponding instantaneous phase data are used to develop a phase-corrected dictionary.

---

[4] where the number of sources are more than the number of microphones
These obtained dictionaries are then used to recover the sources from a mixture. Their results showed that phase-optimized KSVD outperforms the conventional NMF or KSVD methods, where phase information is not taken into account.

4.1.2. Single-Channel Speech Enhancement

In single-channel speech enhancement, the aim is to extract the clean speech signal from a single-channel recorded, noise-corrupted speech. In particular, an optimal estimate of the spectral amplitude and phase in order to recover the underlying speech signal from background noise is desired. Conventional methods used within the last three decades (see [108] for an overview), have mostly focused on filtering the spectral amplitude of the noisy speech, while reconstructing the time-domain signal by copying the noisy spectral phase. Recent advances demonstrate that phase information could be of certain importance, extend the limited performance of the conventional enhancement methods that rely only on amplitude information. The importance of phase information in speech enhancement has been studied extensively in [6, 21] and more recently emphasized in [1, 4, 5, 32]. In this section, we provide an extended up-to-date bibliography to highlight the recent advances made towards phase-aware signal processing in single-channel speech enhancement.

Figure 4 (top row) shows a block diagram of the conventional single-channel speech enhancement, which is composed of two stages: 1) amplitude estimation and 2) signal reconstruction. As phase has an impact on both, in the following we consider three scenarios; in the first two, we study the impact of phase on each stage individually, while in the third we consider the joint estimation of the amplitude and phase information.

For the first stage, a variety of estimators of spectral amplitude are used, which differ in their efficacy, selected optimization criterion, and the assumptions they made about the speech and noise probability distribution (see e.g. [108, 109]). During signal reconstruction stage, the noisy phase spectrum has been a common choice. The noisy phase has been shown to provide the minimum mean square error (MMSE) [30] or maximum a posteriori (MAP) [31] estimate for a clean phase at individual frequency bins. However, this is only true when assuming that the Fourier coefficients are independent in time and frequency, which is obviously not the case in practice [2, 114].

Let $x(n)$ and $v(n)$ represent speech and noise signals, respectively, and let $y(n) = x(n) + v(n)$ represent the noisy observation in a discrete time domain, with $n$ as a time index. Let $Y(k, l) = |Y(k, l)|e^{j\phi_y(k, l)}$ be the complex Fourier representation of the noisy signal as defined for the $k$th frequency bin with $|Y(k, l)|$ and $\phi_y(k, l)$ as the noisy spectral amplitude and phase spectrum, respectively. Similarly, we define $|X(k, l)|$ and $|Y(k, l)|$ as the spectral amplitude for
clean and noisy speech signals, respectively with \(X(k,l) = |X(k,l)|e^{j\phi(k,l)}\) and \(V(k,l) = |V(k,l)|e^{j\beta(k,l)}\) as the complex spectra for speech and noise, and \(|V(k,l)|\) as the noise spectral amplitude. For the observed noisy signal:

\[
|Y(k,l)|e^{j\beta(k,l)} = |X(k,l)|e^{j\phi(k,l)} + |V(k,l)|e^{j\beta(k,l)}.
\]

The spectral amplitude of the noisy signal is the absolute value of the vector sum of the underlying components:

\[
|Y(k,l)| = \sqrt{|X(k,l)|^2 + |V(k,l)|^2 + 2|X(k,l)||V(k,l)|\cos \Delta \phi(k,l),
\]

where the phase difference is defined as \(\Delta \phi(k,l) = \phi_x(k,l) - \phi_y(k,l)\). It is obvious that \(\pm \Delta \phi(k,l)\) are both valid solutions for (21) and this is the source of ambiguity in phase estimation based on the speech and noise geometry [69].

The goal in spectral amplitude estimation is to find \(\hat{X}(k,l)\) in a noisy complex spectrum \(Y(k,l)\) by defining a cost function in the form of \(d = f(X(k,l), \hat{X}(k,l))\), which should be minimized. The form of cost function specifies the type of the estimator and optimality criterion used. Examples of distance functions are the mean square error in amplitude and log-amplitude spectral domain [30, 115], weighted squared error, or Itakura - Saito measure [110] (see [108] for a overview). For the sake of compact representation, in the rest of this subsection we omit the frame index \(l\) unless needed. The optimal estimate for the spectral amplitude of speech \(|\hat{X}(k)|\), in the sense of a minimum mean square error (MMSE), is derived in [109] by optimizing the MMSE criterion formulated as:

\[
|\hat{X}(k)| = \arg\min_{\hat{X}(k)} \left\{ E\left[ (|X(k)| - c(\hat{X}(k)))^2 \right] |\hat{Y}(k)|, \sigma_x^2(k), \xi(k) \right\},
\]

where \(\sigma_x^2(k) = E(|Y(k)|^2)\) is the noise power for frequency bin \(k\), \(\xi(k) = E(|X(k)|^2)/\sigma_x^2(k)\) and \(c(x) = x^\beta\) defines a compression function parameterized by \(\beta\), and \(E(\cdot)\) is the expectation operator. The MMSE estimation for the spectral amplitude was defined as a parametric estimator in [109] and was expressed in the form of a softmax function \(G\) multiplied by the observed signal as \(|\hat{X}(k)| = G(\xi(k), \zeta(k), \beta_a, \mu_a)|Y(k)|\). The parameters \(\beta_a\) and \(\mu_a\) determine the type of the estimator, where \(\xi(k)\) and \(\zeta(k) = |Y(k)|^2/\sigma_x^2(k)\) are defined as the a priori and the a posteriori signal-to-noise ratios (SNRs), respectively. Different choices of \(\beta_a\) and \(\mu_a\) lead to different amplitude estimators. For instance, \(\beta_a = 1, \mu_a = 1\) gives the short-time spectral amplitude (MMSE-STSA) estimator [30] and \(\beta_a \rightarrow 0, \mu_a = 1\) yields the log-spectral amplitude (MMSE-LSA) estimator [115]. By incorporating psychoacoustic considerations into the error function, the estimation of enhanced speech was driven by the perceptual perspective [110].

State-of-the-art speech enhancement techniques use the enhanced spectral amplitude together with the noisy phase in order to reconstruct the enhanced speech as denoted by \(\hat{X}(n)\). This process comprises a return back to the time domain by performing \(F^{-1}(|\hat{X}(k)|e^{j\hat{\phi}(k,l)})\), where \(F^{-1}(\cdot)\) is the inverse short-time Fourier transformation.

4.1.3. Phase Estimation for Signal Reconstruction

Although the noisy phase is a reasonable estimate (MMSE [30] or MAP [31]) for the clean phase spectrum, Vary showed that the phase distortion due to added noise is perceptible for signal components with a local SNR lower than 6 decibels [29]. While the choice of noisy phase at a high enough local signal-to-noise ratios is not critical, the choice of noisy phase spectrum for all signal components in signal reconstruction has been observed to introduce a certain degradation in the signal reconstruction performance [4, 5, 38-40, 48, 69, 113]. Therefore, a proper phase estimation method has the potential to improve perceived speech quality. In the following, we explain recent advances made towards phase estimation for signal reconstruction in speech enhancement and source separation.

**GL-Based:** GL-methods rely on Griffin-Lim iterative signal reconstruction, described as follows. The problem of estimating the spectral phase information dates back to 1980’s, when researchers took the first steps to address the issue of recovering the time-domain signal from STFT magnitude information only. For example, Hays et al. [15] proposed techniques to reconstruct a signal from its magnitude or phase information only. Quatieri et al. [16] derived solutions to estimate the time-domain signal for a minimum-phase sequence, in an iterative manner. Paliwal et al. in [116] presented a review of iterative signal reconstruction methods from short-time Fourier spectra. The key idea behind the exploration of the possibility of reconstructing a signal from its partial information in the STFT (either magnitude or phase) is the redundancy in the short-time Fourier representation, with its minimum half-length overlapped frames, which results in the magnitude and phase information not being independent. Griffin and Lim (GL) [17] derived an iterative solution (a minimum mean square error solution) for signal reconstruction given the
magnitude of STFT. They reported improved performance for large overlaps (e.g., 87.5%) and for a specific choice of window, and square root Hann, leading to the maximum improvement in signal reconstruction in the sense of MSE optimality criterion. The goal was to exploit the correlation between the magnitude and phase spectra.

Inspired by the Griffin-Lim (GL) iterative method for signal reconstruction, several different variants have been proposed in the literature. The original proposal by Griffin and Lim had some fundamental drawbacks: 1) it demanded large number of iterations with Fourier transformations (synthesis-analysis) for each frame in order to provide a high signal reconstruction quality, 2) the algorithm required the magnitude spectra of all frames in the audio data, 3) the GL algorithm was not appropriate for real-time applications due to the large number of iterations. Therefore, the authors in [118] proposed a practical real-time iterative magnitude spectrogram inversion.

In [119], Stursmael and Daudet attempted to extend the GL solution from a full-band to a partial phase reconstruction (PPR) where GL iteration was restricted to a selected time-frequency region. The idea was to replace the noisy phase at a confidence domain where the Wiener filter amplitude of the desired source is above a certain threshold in a static [119], or in a signal-dependent way [117, 120]. More recently, the authors in [121] extended the iterative GL signal reconstruction procedure to a blind source separation framework. The idea was to calculate a remixing error in time-frequency between the mixture and the estimated sources at each iteration of GL and distribute it over individual sources. Different ways for error re-distribution have been considered: uniformly [121], or depending on the activation of sources in the STFT [120], or via a sinusoidal model fit to each source [117]. As an alternative approach, the inconsistency between the valid STFT for a real signal versus the STFT obtained from the Wiener filter gain function and mixture phase was taken into account. Vincent and Le Roux [40] proposed a consistent Wiener filter solution where the inconsistency was incorporated as a constraint when deriving the mean square error solution.

Geometry-Based: In [69], Mowlaee et al. proposed a phase estimation method that relied on a geometry-derived analytic solution together with an additional constraint on group delay deviation. The additional constraint helped to resolve the ambiguity problem faced when choosing between two phase candidates that differed in their resulting sign of $\Delta \phi(k, l)$. The outcome of the phase-enhanced reconstructed signal was compared to benchmark methods, which showed a considerable improvement in blind source separation evaluation (BSS-EVAL) measures [122] including signal-to-distortion ratio (SDR) and signal-to-interference ratio (SIR) as well as the perceived quality measured in terms of perceptual evaluation of speech quality (PESQ) [123]. The geometry-based phase estimation method in [69] relies on the fact that, given the amplitude spectra of speech and noise, there are two phase candidates at each time-frequency cell. In order to remove the ambiguity in selecting the most likely phase value, a group delay constraint at sinusoids was employed. The group delay deviation is defined as the variation in group delay with respect to the fixed value due to the analysis window, denoted by $\tau_w = \frac{N-1}{2}$, with $N$ being the length of the window. The group delay deviation is given by:

$$\Delta \tau(k, l) = \tau_w - \tau(k, l).$$

The group delay deviation has been reported to present minima at sinusoidal components (spectral peaks) [59, 68]. Furthermore, in [38] it was shown that the group delay deviation is related to phase deviation, described as:

$$\Delta \tau(k, l) = \Delta \phi_\text{dev}(k, l),$$

where $\Delta$ is the discrete differentiation operator and we define $\phi_\text{dev}(k, l)$ as the phase deviation given by the difference between the noisy phase and clean phase

$$\phi_\text{dev}(k, l) = \phi(k, l) - \phi_s(k, l),$$

first suggested by Vary [29]. For large enough SNRs (e.g., SNRs above 6 decibels), the group delay deviation becomes small and is primarily contributed by a single sinusoid. The optimal phase value at a sinusoidal peak frequency was calculated by searching for the phase candidate in the ambiguous set, which minimized the group delay deviation [69].

In [70], Mowlaee and Saeidi further extended the geometry-based phase estimation idea to single-channel speech enhancement, where other constraints on phase such as instantaneous frequency deviation (IFD) and relative phase shift (RPS) were incorporated in order to reduce the phase ambiguity. Slight improvements in both perceived quality

\footnote{For a comparative study on the GL-based methods see [117].}
and speech intelligibility as predicted by instrumental measures were reported for phase-enhanced reconstructed signals. The geometry-based phase estimators were integrated into the iterative phase-aware closed-loop single-channel speech enhancement in [38].

**Model-Based:** The authors in [124] proposed reconstructing the spectral phase across the time-axis only and reported insignificant improvements in PESQ measure [123]. The idea of having predictable phase evolving across time for the first few harmonics has often been used in phase vocoders [125] and low-bit-rate speech coding, assuming that the phase of the current frame \( \hat{\phi}_s(k, l) \) is given by adding an error term to the phase of the previous frame \( \hat{\phi}_s(k, l - 1) \), and we have

\[
\hat{\phi}_s(k, l) = \hat{\phi}_s(k, l - 1) + \alpha_s \text{IF}(k, l),
\]

where \( \text{IF}(k, l) \) is the instantaneous frequency (IF) defined in (4) and \( \alpha_s \) representing a positive real-valued factor that indicate as an increasing or decreasing signals instantaneous frequency (phase vocoder case for \( \alpha_s = 1 \)). Following this principle, the authors in [111] proposed STFT phase reconstruction of noisy speech signal called STFT phase improvement (STFTPI). The method imposes the harmonic structure up to the Nyquist frequency. A comb filter harmonizes the noisy speech, by modification of noisy phase spectrum across the frequency axis in order to compensate the phase introduced by the window. The spectral phase components between harmonics were modified to the value of the closest harmonic. An improvement in PESQ was reported in [39, 126] but buzzy speech quality was introduced, in particular at high frequencies. A similar concept was employed in [126] to estimate noise in-between harmonics. Improvement in terms of PESQ [123], was reported when replacing the noisy STFT phase with the temporally reconstructed one using the phase prediction model given in (26). The improvement predicted by PESQ versus the informal listening results achieved for STFTPI phase-only enhanced signals invites questions about the reliability of the existing instrumental measures that were originally proposed for conventional amplitude-only speech enhancement, without considering capturing phase distortion. In Section 4.4, we will expand on this topic and present some recent findings for speech quality estimation when phase-aware signal processing is utilized [75, 76].

**Time-Frequency Smoothing of Unwrapped Phase:** With the recent advances made in speech synthesis using harmonic model plus phase distortion (HMPD) [89], it is possible to obtain an unwrapped phase from an instantaneous STFT phase. Mowlaee and Kulmer [48] proposed applying harmonic phase decomposition followed by a temporal smoothing filter in order to reduce the variance of the noisy phase. The phase variance has been shown to be a reliable metric for the quality assessment of synthesized speech [95], and was further reported to be the key to resolving the spectral phase information of two adjacent harmonics [5]. The phase variance given by the PDD calculated in (14) was shown to be a function of the selected window and, in particular, depends on two characteristics: i) side-lobe level rejection and ii) the coherent gain\(^6\). A study in [5] showed that a Blackman window choice provided a good balance between the joint improvement in quality predicted by PESQ [123] and short-time objective intelligibility (STOI) [128] measures. In [97] the idea of temporal smoothing of unwrapped noisy phase was extended to selective time-frequency smoothing by using a probabilistic approach, that relied on a pre-defined threshold on the circular variance of the unwrapped phase. The method led to a joint improvement in instrumental speech quality and intelligibility when used for phase-only enhancement or when combined with some conventional amplitude-only enhancement. Finally, Maly and Mowlaee [129] studied the importance of harmonic phase modification for signal reconstruction in speech enhancement reporting that an unwrapped phase combined with a linear phase suffice for an improved signal reconstruction in speech enhancement.

**Mismatched Window:** Wojcicki and Paliwal [130] suggested employing different windows in terms of their dynamic range in the analysis and synthesis stages. Paliwal et al. showed that employing a Hamming window for analysis, followed by choosing a Chebyshev window for the synthesis stage, resulted in improvement in PESQ.

**Phase Spectrum Compensation (PSC):** The DFT phase spectrum of the noisy speech signal is conjugate symmetric. In [131, 132], the authors proposed a speech enhancement approach by controlling the degree of conjugate symmetry enforced on the noise-corrupted spectrum. By modifying the noisy spectral phase, hence, the angular relationship between the speech and noise spectra, an anti-symmetry function was applied in order to compensate the noisy short-time phase spectrum before the signal reconstruction stage. The compensated phase spectrum was eventually used together with the noisy spectral amplitude to produce a modified phase-enhanced time-domain signal. The method improved the perceptual quality of noisy speech, in particular, and was reported to be well suited for mid-

---

\(^6\)The coherent gain is defined as [127]: \( \text{CG} = \sum_{n=0}^{N-1} w(n) \), where \( w(n) \) is the prototype window of length \( N \) with \( n \in [0, N - 1] \).
to high-SNR scenarios.

**Maximum A Posteriori (MAP):** Assuming a uniform distribution of spectral phase and the independence of the spectral amplitude and phase, it has been shown that an optimal estimate for the clean phase is equal to that of the noisy phase in terms of the minimum mean square error (MMSE) [30] or maximum a posteriori (MAP) sense [31]. In contrast, changing the prior distribution of phase from uniform to a non-uniform one results in a different MAP phase estimator rather than the observed noisy phase. For example, in [37], it was shown that the phase posterior for a known speech amplitude follows a von Mises distribution. For a given clean amplitude, a MAP phase estimator of the clean phase, was proposed relying on a grid search over a set of candidate phases in the possible range, i.e., $\phi(k, l) \in [-\pi, \pi]$. As reported in [37], the brute-force solution is difficult due to its increased computational complexity.

More recently, Kulmer and Mowlaee [49] proposed a closed-form solution for a MAP estimator of clean phase. The MAP harmonic phase estimate was derived by replacing the uniform prior with a von Mises distribution and under the assumption of a white Gaussian distribution for the noise signal. The MAP phase estimate is a function of the von Mises parameters (mean and concentration), SNR at the underlying harmonic\(^7\), and the observation data length. For large SNRs in the limiting scenario, the MAP estimator asymptotically approaches the ML estimate given by the noisy DFT phase sampled at the harmonic frequency [133, p. 168]. At such high SNR scenario, the noisy phase is more weighted rather than the mean value. This is made possible by incorporating a low concentration parameter in the von Mises prior considered in the MAP phase estimator derived in [49]. On the other hand, the additional dependency on the harmonic SNR serves as a reliability mechanism for the estimated phase, implying that at low SNRs, the proposed MAP estimator relies only on the circular mean. The circular mean and variance parameters in the von Mises distribution phase prior provide a flexible sweep between the maximum certainty (Dirac delta with infinite concentration parameter) and maximum uncertainty (uniform prior with zero concentration parameter) and statistically take into account the uncertainty in the estimated phase.

**Phase Randomization:** Miyahara and Sugiyama [134–136] proposed phase randomization as a new paradigm for speech enhancement. They reported improved speech enhancement for auto-focusing noise. The auto-focusing noise is a mechanical noise in audio-visual recording using digital still cameras. In particular when using zoom function during movie recording, the auto-focusing noise could become stronger than the desired signals from the environment, masking speech [136].

The main principle in phase randomization is to take into account the underlying auto-focusing or clicking noise and its linear phase character. Therefore, the structured patterns of the phase of noise signal could be de-emphasized via randomizing the observed noisy phase spectrum. Phase randomization contributes to change the noise corrupted regions and restricts the constructive addition (in-phase) of noise-dominated spectral components. Results in [136] demonstrated that, using the overlap-add at signal reconstruction, the consecutive segments with randomized phase brought improved noise reduction performance due to the removal of the noise phase pattern. The improvement was justified via subjective listening results using a comparative category rating (CCR) test [134–136].

**Minimum Variance Phase Estimation Error:** Mowlaee and Kulmer [5] presented a theoretical derivation for the phase variance in a single-channel speech enhancement and demonstrated how to optimize the window to minimize this variance. It was shown that the selected window impacted the variance in phase estimation directly, and the phase variance was a function of window type and signal-to-signal ratio of the adjacent harmonic to the current one. The choice of the Blackman window in phase estimation was shown to provide a good balance between a joint improvement in perceived quality and intelligibility, as predicted by instrumental measures. Throughout this analysis, the authors in [5] presented the potential and limits of the phase estimation methods in single-channel speech enhancement. Furthermore, they made a comparative study between representative methods and quantified their performance. The performance evaluation was carried out using instrumental predictors for perceived quality and speech intelligibility using PESQ and STOI, respectively, as reported in [4, 5]. In a subjective evaluation, MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) [75, 137] and comparison category rating (CCR)\(^8\) [4] tests were conducted to estimate the perceived speech quality achieved by enhancing the phase of a noisy speech signal. For a subjective evaluation of speech intelligibility, the listening test was conducted and results were reported in [4, 76] (see also section 4.4 for more details).

---

\(^7\)defined as the signal-to-noise ratio for each sinusoid sampled at each harmonic

\(^8\)the instructions provided in [138] was recommended for assessing speech processing methods that either degrade or improve the speech quality.
4.1.4. Phase-Aware Amplitude Estimators

An estimate of the clean phase could be used to derive an improved spectral amplitude. In this regard, the authors in [112, 113] derived the MMSE estimator of the spectral amplitude given the clean spectral phase information. The derived estimator is called phase-sensitive and showed an improved performance in the sense of a reduction in noise outliers and more noise reduction because of the phase contribution. The derivation showed dependency to the phase deviation which rotates the noisy phase coordination towards the clean phase one, the phase difference was interpreted as an additional noise reduction mechanism as it corrects the noisy phase coordination towards the coordination of the clean spectral coefficients. This, as a consequence, leads to additional noise reduction, apart from the MMSE phase-insensitive part. The phase deviation concept was first defined by Vary in [29], given in (25) whereby he studied the impact of phase modification in speech enhancement concluding that a phase deviation below 0.679 radians is not perceptually relevant to human. Vary also showed that the maximum phase deviation corresponds to a local signal-to-noise ratio for a harmonic above 6 decibels threshold. The phase deviation concept was further used for speech quality estimation [75], phase estimation, and iterative closed-loop phase-aware speech enhancement [38].

In [139], a combination of phase-aware minimum mean square error (PAMMSE) [112, 113] with a phase-insensitive amplitude estimator [109] was proposed. The combined method relied on the voicing probability, provided by a pitch estimator that was robust to high levels of noise (PEFAC) [140]. The enhanced spectral amplitude under conditions of voiced-unvoiced uncertainty was derived in [139]. The method made a trade-off between the degree of confidence and the amount of phase information to incorporate in voiced frames, while the conventional phase-insensitive method was used in unvoiced frames.

4.1.5. Joint Amplitude and Phase Estimation

So far we consider the individual estimators for amplitude and phase spectra of the clean signal. In the following, we consider approaches addressing the issue of jointly estimating the amplitude and phase spectra of clean speech from the noisy observation.

Complex Spectral Speech Coefficients Given Uncertain Phase Information (CUP): Following that the distribution of the clean speech phase around the prior phase information is characterized by a von Mises distribution given in (17), in [37] a joint estimator called Complex spectral speech coefficients given Uncertain Phase information (CUP) estimator, was proposed. The idea was to take into account the uncertainty in the spectral phase information given an uncertain clean phase estimate, in the derivation of a Bayesian estimator for clean speech coefficients [37].

Iterative Closed-loop Phase-Aware Speech Enhancement: Mowlaee and Saeidi in [38] proposed a closed-loop configuration that jointly improved the spectral amplitude and phase given a noisy speech observation. The block diagram for the method is shown in the bottom panel in Figure 4. The central idea was to incorporate the phase estimator followed by the MMSE phase-aware amplitude estimator [112, 113], and iterate until either a stopping criterion was satisfied or a maximum number of iterations was reached. Improved PESQ was reported through a few iterations. As the stopping criterion, it was proposed to check the number of iterations required in order to achieve a reasonable reduction in the inconsistency measure in the form of a relative error capturing the difference in the STFT inconsistency across iterations.

Phase Enhancement Combined with Amplitude Enhancement: The methods enlisted in Section 4.1.3 provide an estimate for the clean phase spectrum. They could be used as the post-processing step to improve the signal reconstruction performance in a conventional amplitude-only enhancement method. We refer to [5], regarding a comparative study that investigated how much of each amplitude or phase spectra is important to bring about improved speech quality. The phase information recovers missing harmonics and allows additional noise reduction between harmonics, with a special focus on onset (timing events). The phase enhancement further suppresses the musical noise outliers, which are introduced in the amplitude enhancement stage. An improvement in performance up to 0.2 in PESQ and 2% in STOI has been reported [5]. Similar joint improvements were reported using the MAP phase estimator proposed in [49] with a lower sensitivity to errors in fundamental frequency estimation.

4.1.6. Proof-of-Concept Example

As a proof-of-concept, we report the importance of incorporating phase information in two stages of single-channel speech enhancement: i) in the signal reconstruction stage, ii) in the phase-aware amplitude estimator. For the phase estimation, we choose the MAP phase estimator [49] and for the phase-aware amplitude and phase estimator
we consider the iterative phase-aware closed-loop method proposed in [38]. The conventional single-channel speech enhancement is set to MMSE-LSA [30] with noise estimation from minimum statistics [141]. Figure 5 shows the results obtained by phase-aware speech processing methods applied to enhance a noisy speech signal. The noisy speech file is created by combining a female clean speech selected from GRID corpus with babble noise set at 0 (dB) SNR (the same sentence as used in the phase representation, see Figure 1). The results are shown in terms of spectrogram (top), phase variance (middle) and group delay (bottom). The achievable improvement\(^9\) in speech enhancement performance is quantified using instrumental predictors of perceived quality (PESQ [123]) and speech intelligibility (STOI [128]), shown at the top of each panel. Consistent improvement is achieved when phase estimation is applied compared to when no phase information is applied in the conventional method (third column).

4.1.7. Phase-Aware Single-Channel Source Separation

Unlike with speech enhancement, there are applications which demand recovering all the underlying sources in an observed mixed signal. This scenario is categorized as source separation which is carried out either with a single microphone or multiple microphones. While the advances regarding phase-aware signal processing in multi-microphone scenario was earlier discussed in section 4.1.1, we will now summarize the advances made towards incorporating phase information to solve the source separation problem.

The single-channel source separation (SCSS) problem is formulated in the following: given a mixture of \(P\) audio sources denoted by \(y(n) = \sum_{p=1}^{P} x_p(n)\), with \(x_p(n)\) referring to each \(p\)th underlying source in the mixture, the goal is to estimate all sources \(\hat{x}_p(n)\). In the case of two sources, to simplify the interpretation, and applying the Fourier transformation, we have:

\[
Y(k, l) = |X_1(k, l)e^{i\phi_1(k, l)} + |X_2(k, l)e^{i\phi_2(k, l)}.
\]  

(27)

For perfect recovery of the sources, both spectral amplitude and phase information are required to be estimated from the observed signal. This infers that the superposition of the estimated source spectra should reproduce both spectral amplitude \(|Y(k, l)|\) and spectral phase \(\phi_1(k, l)\) of the observed signal \(y(n)\). The single-channel source separation problem is an ill-conditioned one, as the number of unknowns \((=4)\) is lower than the number of observations \((=2)\). As a work

\(^9\)The audio samples are available at https://www.spsc.tugraz.at/PhaseLab.
around to solve the aforementioned problem, the conventional SCSS methods are often focused on filtering the spectral amplitude of the mixed signal only, while they copy the mixture phase spectrum directly at signal reconstruction stage [142]. The conventional methods are mainly categorized into time-frequency masking (TFM) and non-negative matrix factorization (NMF), both neglecting the phase information consideration for spectral amplitude modification or signal reconstruction. In the following, we provide a brief overview of each method.

The TFM source separation methods rely on estimating a time-frequency mask. The mask is eventually applied on the observed signal to estimate a underlying target or masker signal. The TFM methods are further categorized into binary mask and ratio mask groups, which differ in their selected mask type. The binary masking is introduced as the main goal in computationally auditory scene analysis (CASA). Wang [143] presented details of CASA-based methods, which rely on an estimated ideal binary mask required to separate individual sources from a given mixture. The ideal binary mask has been reported to have the maximum intelligibility achievable by employing a source-driven method (CASA) [144]. A binary mask (that takes values either of 0 or 1) is applied to the observed signal to recover both underlying signals in the mixture. Without loss of generality, the ideal binary mask, for the first source is defined as follows:

$$
\hat{X}_{1}^{\text{IBM}}(k, l) = \begin{cases} 
Y(k, l) & |X_1(k, l)| \geq |X_2(k, l)| \\
0 & \text{Otherwise}
\end{cases}
$$

By swapping the role of the underlying sources, we can also define the second source binary-mask separated signal, which is denoted by $\hat{X}_2^{\text{IBM}}(k, l)$. An ideal ratio mask (IRM) is defined as:

$$
\hat{X}_1^{\text{IRM}}(k, l) = \left( \frac{|\hat{X}_1(k, l)|^2}{|X_1(k, l)|^2 + |\hat{X}_2(k, l)|^2} \right)^{\beta_s} Y(k, l),
$$

where $\beta_s$ is the compression factor (also called as the tuning parameter) used to scale the mask\(^{10}\).

The second categorization of the conventional SCSS is NMF. The goal in NMF is to represent the spectrogram of the observed signal in terms of a weighted sum of spectrograms originating from the different ambient sounds/events of the underlying sources. The weights are estimated in the separation stage while the basis spectrograms (called atoms) are learned in the training stage. Then, each source is represented as [147]:

$$
|\hat{X}_p| = \sum_{m=1}^{M} w_{m,p} \mathbf{b}_{m,p},
$$

where $M$ is the number of atoms, with $m \in [1, M]$ representing the atom index, and $\mathbf{b}_{m,p}$ refers to the $m$th atom for the $p$th source, and $w_{m,p}$ is the non-negative weights quantifying the contribution of its corresponding basis vector.

Neither of the aforementioned conventional SCSS methods (TFM or NMF categories), take the phase information of the sources into account. The phase information, however, as displayed in the block diagram in Figure 4, affects two individual steps in the single-channel source separation problem: i) signal interaction, and ii) signal reconstruction. For signal interaction, the information with respect to the spectral phase difference of the sources plays an important role in the interaction model used to account the interactions between the underlying sources in the observed mixture. In [148], Mowlaee and Martin showed that a phase-sensitive interaction model outperforms the conventional ones, such as log-max or MMSE (in power spectrum), that average out the phase information. Erdogan et al. proposed a phase-sensitive filter model that led to improved single-channel source separation [149] and automatic speech recognition performance [150] compared to phase-insensitive ones (log-max or MMSE).

For signal reconstruction stage, proper phase information about the source leads to certain improvement in the synthesis quality [32, 69]. Like the single-channel speech enhancement described in section 4.1.2, replacing the mixed phase spectrum with a clean spectral phase estimate has a positive impact on the achievable signal reconstruction performance in single-channel source separation. The ideal binary mask is known to produce the upper-bound performance for speech intelligibility in the single-channel source separation literature yet suffering from a poor performance in terms of perceptual quality. Several attempts have been made for improving the perceived quality of the separated sources when a time-frequency mask is applied. For example, Williamson and Wang [151] proposed a

\(^{10}\)For example, for $\beta_s = 0.5$, we get the square-root Wiener filter, which provides the best results as shown in [145, 146].
two-stage approach in which a softmask was applied on the mixed signal. In the second stage, they employed NMF to impose sparsity constraint while reconstructing each source. Their results showed improvement in both perceived quality and speech intelligibility performance. In [152] the same authors presented a review of different signal reconstruction techniques, that could be used to improve the perceptual quality of binary masked speech. Improved perceived quality without loss of speech intelligibility was reported. Boldt et al. [153] applied hidden Markov models to correct errors while estimating the ideal binary mask, and hence, improved the signal reconstruction performance.

Several studies investigated the importance of phase spectrum in improved signal reconstruction in single-channel source separation. Mowlaei et al. proposed a geometric approach to estimate the phase information for signal reconstruction [69]. The method included auxiliary constraint regarding the source phase spectra to resolve the ambiguity in the phase candidates provided from the problem geometry. To demonstrate the effectiveness of phase estimation for signal reconstruction, they compared results with respect to two benchmark methods: ideal binary mask (SCSS upper-bound) and multiple input spectrogram inversion (MISI) [121] (where GL iteration is combined with re-mixing error distribution). Throughout their experiments, it was shown that geometry-based approach improved performance in terms of PESQ [123] and signal-to-distortion ratio [122] measures. More recently, Mayer and Mowlaei [32] proposed a phase estimation method for single-channel source separation, which relied on harmonic phase decomposition, first proposed by Degottex [33]. As phase decomposition relies on an estimated fundamental frequency of each source, a multi-pitch estimator was also proposed [32]. The estimated pitch trajectories of each source was then used to estimate the unwrapped phase. Temporal smoothing filters were applied on the unwrapped phase spectrum to reduce the large variance caused by the interfering source in the mixture phase spectrum. It was demonstrated that the source separation performance of the ideal binary mask as well as NMF is improved in terms of quality and intelligibility when phase recovery is applied. This was a crucial finding as the ideal binary mask is associated with the highest achievable speech intelligibility performance in single-channel source separation. Theses findings altogether highlight the importance of phase information in single-channel source separation since the joint improvement in quality and intelligibility was not achievable using the ideal binary mask or NMF where phase information is not taken into account (the conventional SCSS methods improve either quality or intelligibility at the expense of reducing the other).

The phase information has been also directly taken into account to modify the matrix factorization, which is commonly carried out by employing NMF. In NMF, the mixture spectrogram is approximated as the sum of the spectrogram of the underlying sources, hence the cross term due to the phase difference of the sources is neglected. Parry and Essa [154] investigated the sensitivity of this conventional assumption in spectrogram factorization by considering a probabilistic phase representation. They showed that given the source spectrograms, their probabilistic phase representation contributes to a better fit to the true distribution of the components in the mixture spectrogram, compared to the conventional NMF. Kameoka et al. [155] proposed complex NMF by introducing an interaction model in the complex-spectrum where additivity between the sources holds. This complex interaction model was also used to formulate an iterative approach to extract the magnitude spectra that fit to the complex spectrum and the phase spectra of the underlying signals. King and Atlas [156] proposed to incorporate phase estimations using complex matrix factorization to select the bases for the optimal separation of sources. They demonstrated that incorporating phase information in complex matrix factorization (CMF) improves source separation and automatic speech recognition performance as compared to NMF. Bronson and Depalle in [157] proposed using phase-constrained complex NMF to separate the overlapping time-frequency regions in mixtures composed of harmonic music sources. Ewert et al. [158] proposed a method to refine the dictionaries learned by conventional NMF method by taking into account the interactions of sound sources in the spectrogram on the basis of their phase overlaps and canceling regions. Magron et al. [159] proposed a phase recovery approach in the NMF framework in audio source separation. They presented a comparison between two scenarios: blind separation without prior information and oracle separation with supervised learned model. A comparative study on the GL-based methods in single-channel source separation was made in [117]. In [160], Magron et al. presented a comparative study in which different phase reconstruction techniques including GL-method were used for NMF-based source separation. Their method relied on unwrapping the phase over time and frequencies to meet the temporal and spectral coherence of the signal between partials. The techniques discussed in [158–161] attempt to resolve the linear phase mismatch as signal reconstruction, demonstrated to improve separation performance of pitched musical sources from their mixture.
4.2. Speech Coding

In speech coding the goal is to meet a certain bit-rate constraint while still a transparent speech quality is provided\textsuperscript{11}. The most of speech coders mainly focus on assigning bits to represent the spectral envelope or harmonicity. However, phase has been also shown to be perceptually important [51, 91, 92, 162–164].

Phase discontinuities between consecutive frames may occur due to changes in the fundamental frequency at frame boundaries resulting in perceptually audible artifacts. Transitions from a voiced segment without any connecting silent or unvoiced segments may cause such discontinuities. In hybrid coders the initial phases of the harmonic excitation are utilized from the past excitation vectors. This procedure could be problematic in particular for rapidly-changing speech onsets [165]. Choosing small data blocks in waveform coders, such as code-excited linear prediction (CELP) is a way to get around this issue. However, it causes inaccurate coding of the phase spectrum for a pitch cycle in the speech signal, resulting in linear phase mismatch at signal reconstruction [166]. In particular, any displacement of a fraction of a sample could introduce audible distortion at high frequencies, specially at frames of length shorter than the pitch period.

The perceptual degradation due to the aforementioned displacements in speech coding could be circumvented by considering enough number of bits, hence performing a high resolution coding whereby small changes of phase is taken into account. For example, in mixed-excitation linear prediction (MELP) or CELP speech coders the continuity of the signal is preserved by assigning bits for phase alignment for the encoded frames, which could be important for signal reconstruction due to linear phase mismatch. In order to limit the artifacts made when switching between MELP coder (amplitude-only) and CELP coder (amplitude and phase), the zero phase equalization concept has been proposed [167, 168] to reduce the phase discontinuities by excluding the phase component that is not encoded in the MELP coder. In this approach, the signal phase component which is not coded in MELP is removed. The method was applied for the LPC residual using a filter with finite duration [167], where the coefficients were calculated using a smoothed pitch pulse waveform. The obtained zero-phase-equalized residual signal was eventually used for reconstructing the coded speech using the LPC synthesis filter.

The importance of phase quantization in terms of perceptual artifacts has been studied [163]. It was suggested to account for the fact that the human auditory system is only sensitive to the relative phase changes within one critical band. Therefore, irrelevant phase information of speech could be eliminated in the procedure of quantization and harmonic coding. Subjective test results demonstrate the effectiveness of this approach. More phase components were reported to be perceptually important for low-pitched signals (e.g., male speech) than for high-pitched signals (female speech), concluding that for the sake of a high quality CELP coded speech for male speech, more bits are required to encode the phase information. In [169], the just-noticeable difference (JND) [170] was first measured for the phase of one harmonic versus different fundamental frequencies. A mathematical model of JND was then proposed as a quantization method making it possible to assign bits to the most perceptually important phase information. This procedure led to a quantized speech signal that was perceptually close to the uncoded speech.

In low-bit-rate harmonic speech coders, the phase information is often not transmitted, but is produced at the decoder for speech reconstruction [165]. Examples include phase estimation using minimum-phase assumption [171], where the harmonic phase is predicted based on frequencies at the previous and present frames [50]. Pobloth and Kleijn [162] demonstrated that the human perception capacity towards phase is considerable, therefore, the existing speech coders take into consideration phase distortions, in particular, for low-pitched voices. In [91], Pobloth and Kleijn also studied the relevance of a squared error, occurring between the original and phase-distorted signals. While the squared error correlates well to the perceptual measures for small phase distortions, it does not correlate with increases in perceptual measure for large distortion values.

In other studies [163, 169], Kim showed that phase information below a certain frequency in a harmonic signal is not perceptually relevant. This certain frequency is called the critical phase frequency defined as:

\[ f_{\kappa} = \kappa_c f_0, \]

(30)

where \( \kappa \) is the index of critical phase frequency. The phase modification for those frequencies above \( f_{\kappa} \) is defined as the local phase change, which modifies the phase relationship within a critical band. This changes the timbre of

\textsuperscript{11}A transparent speech quality is referred to when the reconstructed speech at receiver is perceptually close enough to the one at the transmitter.
the signal and is perceptually relevant. In contrast, phase changes for harmonic frequency below the critical phase frequency are in fact glottal phase changes, and only modify the phase relationship between channels, while the relative phase relationship within critical bands is preserved, and is hence, not perceptually important. The parameter $\kappa_c$ is pitch-dependent and using equivalent rectangular bandwidth (ERB) representation, is given by $\kappa_c = \left\lceil Q_{\text{ear}} (1 - B_{\text{min}} f_0) - 0.5 \right\rceil$, \hspace{1cm} (31)

with $Q_{\text{ear}} = 9.26449$ and $B_{\text{min}} = 24.7$ (Hz), following Glasberg and Moore’s recommendation [173]. For example, for a fundamental frequency of $f_0 = 100$ (Hz), a critical phase frequency of $\kappa_c = 7$ is obtained. Further, Kim in [169] concluded that phase information below the critical phase frequency are irrelevant to perceived quality as they result in a lower JND compared to than high frequencies. Finally, in [172], it was shown that encoding phase of harmonics above or equal to the $\kappa_c$th harmonic index suffices to preserve the perceptual quality of coded speech.

In low-bit-rate speech coding, the predictability of lowband harmonics across frames has been taken into account to minimize the number of bits assigned for harmonic phase representation. In particular, in CELP and MELP coders, only the phase information at peaks is transmitted, while at the decoder, the phase information across time is synthesized using interpolation techniques. For example McAualay and Quatieri [86], assumed a minimum phase for the phase contribution by the vocal tract filter and proposed cubic interpolation to provide smooth phase trajectories for speech synthesis. In order to maintain a transparent speech quality, a better phase than minimum phase information at harmonics is required. Phase dispersion has been used to train codebooks for speech coding and complement the minimum phase assumption [92].

Finally, it is important to note that the correct overlap alignment of sinusoidal trajectories regarding their phase coherence has been frequently studied and has been reported of high importance for the sake of high-quality synthesized speech. Phase interpolation techniques are commonly used in speech synthesis using sinusoidal/harmonic models: cubic phase model [86] and quadratic polynomial phase [174]. More recently, Agrimonakkis [99] presented a review of phase smoothing methods used for harmonic speech reconstruction. Agrimonakkis also proposed a novel phase interpolation called quadratic phase splines where the phase-locked pitch-synchronous and the smoothing property of phase were taken into account. The methods showed improved speech quality in the text-to-speech synthesis application.

4.3. Digital Speech Watermarking

In the digital multimedia industry, it is important to develop robust ways to hide and secure multimedia (audio or video) from unauthorized users and intentional manipulation through transmission. To guarantee the proof of ownership for the manufacturers, watermarking is used. A watermark (guest) signal is added to the host signal, enabling the host signal to be identified. In the following, we will present a review on those speech watermarking methods where phase information is taken into account (for a review of multimedia watermarking, please refer to [175]).

A speech watermarking approach, which relied on manipulation of the phase spectrum of unvoiced speech has been proposed while keeping the power spectrum unchanged. The watermark data (guest signal) was embedded into the phase of unvoiced speech segments by replacing the excitation signal of an auto-regressive signal representation. In unvoiced speech, the auditory system does not distinguish the different realization of the Gaussian excitation process as long as the temporal and spectral envelope of speech is preserved [176]. This insensitivity for the fluctuations in unvoiced and whispered speech allows watermark data to be embedded in the phase spectrum of the unvoiced speech, at a rate of up to 540 bits/s [177]. The method was reported to be highly robust against various attacks including nonlinear phase and bandpass filtering.

A speech watermarking method that relies on the harmonic model of speech signal and harmonic phase coding principle was suggested in [52]. The digital watermark data is embedded into the phase at harmonics while preserving the relative phase shift (RPS) as an important structure for high quality speech synthesis (see e.g. [34, 46, 96]). The method used each harmonic as a communication channel, conveying one bit digital data per frame. In theory, all harmonics could be used to hide data, however, the variability of the number of harmonics across frames and speakers, restricts the use of the method. Therefore, only a selection of channels was used to hide data. The method
was reported to be robust to coding algorithms including MP3 and OPUS\textsuperscript{12}. In order to quantify the performance difference between the received and transmitted RPS features, the mean square error criterion of the RPS domain was defined.

Dong et al. [178], proposed a data hiding technique based on the phase modulation of audio signals. Two phase encoding methods were proposed: i) re-assigning the relative phases for selected frequency components in the audio signal, and ii) based on quantization index modulation where variable set of phase quantization steps were employed, to hide data. The capacity of the method in terms of embedded data was reported as 20 kbit of data per minute. A solution to compensate for the phase discontinuity of the audio signal was proposed by enforcing the phase shift to reach zero at the frame boundaries, where phase discontinuity is most pronounced due to linear phase mismatch.

The human auditory system is sensitive to the relative phase difference between frequency components but not their absolute phase values. Liew and Armand [54] exploited this property to propose a speech watermarking method by embedding the watermark data in phase of the randomly chosen frequencies of low amplitudes. The watermark data should be embedded in such a way that the changes in the phase between frequencies remained insignificant and thus, perceptually inaudible. A variable frame strategy was used to improve the robustness of the method against synchronization attacks like time-scaling. An algorithm for covert digital audio watermarking has been proposed [53] that relied on the perceptual insignificance of the long-term multi-band phase modulation. The method achieved a data rate of 20-30 bits/s and was shown to be highly robust to coding methods such as Moving Picture Experts Group Advanced Audio Coding (MPEG AAC) [179].

4.4. Speech Quality Estimation

In this Section, we focus on the use of phase information in predicting the perception of speech quality. This is important in the design for high quality speech communication systems. The perceived quality is often characterized by the naturalness, clarity, or brightness of the speech signal. In contrast, the speech intelligibility refers to the number of correctly recognized words in a speech utterance. For a recent review with regard to speech quality estimation, we refer to Möller et al. [180] (on speech quality), and Kleijn et al. [181] (on speech intelligibility).

Conventional instrumental measures commonly used to evaluate the speech quality are categorized in three groups: 1) SNR-based methods including global SNR (GSNR) [182], frequency-weighted segmental SNR (fSNRseg) [183], segmental SNR (SSNR) [184] and , 2) speech coding-based measures including perceived speech quality estimation (PESQ) [185], log-likelihood ratio (LLR) [186], cepstral distance (CEPS) [187], and Itakura-Saito distance (ISd) [188], 3) source separation based measures including blind source separation evaluation (BSS EVAL) [122] comprised of signal-to-distortion ratio (SDR), signal-to-interference Ratio (SIR) and signal-to-artifact Ratio (SAR), and perceptual evaluation methods for audio source separation (PEASS) [189]. Conventional instrumental measures used to evaluate the speech intelligibility are: speech intelligibility index (SII) [190], SNR loss [191], coherence SII (CSh) [192], normalized covariance measure (NCM) [193], mutual information k-nearest neighbor (MIKNN) [194], speech intelligibility prediction based on mutual information (SIMI) [114], short-time objective intelligibility measure (STOI) [128], DAU model [195], and glimpsing model [196]. Most of these conventional measures either take no phase information into account, or their prediction performance when used for phase-modified speech signals has not been studied.

Although many instrumental measures have been defined and standardized for the performance evaluation of speech processing algorithms, their use is limited to conditions where they have been tested or recommended (see e.g., [110, Ch. 11] or [197] for a detailed list of the conventional objective measures). Because many of these measures rely on a distortion metric as defined between the spectral amplitude of a reference signal (termed desired signal) and the estimated (termed processed signal) signal, it is not clear what the reliability of the conventional measures is when used to predict the performance of a phase-aware speech processing method. Therefore, in this section of the current overview, we present some highlights regarding the recent advances made towards speech quality estimation of enhanced speech when phase-aware signal processing is applied. The importance of phase information in speech quality evaluation could be studied by addressing the following two questions:

1. How reliable are the conventional instrumental measures in predicting the performance of signal enhancement methods when both spectral amplitude and phase are modified?

\textsuperscript{12}OPUS is a lossy audio coding format developed by the Internet Engineering Task Force (IETF) that is particularly suitable for interactive real-time applications over the Internet
2. Is it possible to improve the reliability of the quality/intelligibility prediction by considering new measures (termed phase-aware instrumental metrics) where distortions due to spectral amplitude and phase are both taken into account?

Gaich and Mowlaee suggested a list of phase-aware instrumental measures to predict the speech quality and studied their usefulness for phase-aware speech processing, in terms of the perceived quality [75] and speech intelligibility [76]. The candidate phase-aware measures included in the studies were: group delay deviation (GDD), instantaneous frequency deviation (IFD), unwrapped mean square error\(^{13}\), phase deviation (PD). They also proposed two other measures focused on quantifying the phase estimation error in the unwrapped domain: unwrapped harmonic phase SNR (UnHPSNR) and unwrapped root mean square estimation error (UnRMSE).

Gaich and Mowlaee [75] studied the correlation between the existing instrumental measures for speech quality with subjective listening results for phase-aware enhanced speech signals. They reported that the phase deviation metric shows the highest correlation ($\rho = 0.92$) on average, followed by PESQ ($\rho = 0.9$). Furthermore, the PD metric performed most reliably because it showed a more stable correlation coefficient across noise types. Gaich and Mowlaee repeated their investigation in [76] on speech intelligibility where they reported that coherence speech intelligibility index (CSII) measure for the medium and the low levels followed by STOI led to the highest correlation with the results in the subjective listening tests.

These observations indicate that more studies are needed to develop new instrumental metrics when quantifying the achievable performance of a noise reduction or source separation method that modifies both amplitude and phase spectra. The conventional measures, BSS-EVAL [122] including signal-to-distortion ratio, signal-to-interference ratio, signal-to-artifact ratio, and their follow up measures [189] could be adapted into phase-sensitive versions where both amplitude and phase distortions are taken into account. Finding a reliable predictor for quality and intelligibility evaluation will also contribute to the revision of the cost function used to optimize the signal processing stage that best matches with the subjective listening results.

4.5. Speaker and Speech Recognition

In the following, we consider two important applications in speech technology addressing the advances made towards incorporating phase information in the feature extraction stage.

4.5.1. Automatic Speech Recognition (ASR)

It has been shown that the phase information is important in human speech recognition [104]. Working with phase spectrum, there are several practical issues that must be considered when analyzing phase information, which are commonly computed from the short-time Fourier transform (STFT). Some of these practicalities include dealing with compensation for the inter-frame time step, the effect of the window function, and the lack of a common temporal origin when making comparisons of the phase spectrum over different sequences. Furthermore, it is important to select proper processing parameters such as frame size and window type that are well-suited for phase analysis rather than naively applying parameters that work well for the amplitude spectrum estimation [2, 198].

Phase spectrum has long been used for other applications like pitch and formant extraction [67, 199, 200], and phase information has been found to be useful in epoch extraction and recognition [201]. Delta phase parameters as group delay features are treated as long-term analysis of phase in [202] and by experiments on Japanese version of the AURORA-2 database an improved word recognition accuracy is reported for phase-based features over magnitude-only features in high SNR conditions. Spectral subtraction in its primal form ignores the phase difference between speech and noise frequency components and in order to obtain better word recognition rates, the correlation between speech and noise can be accounted for as in [203]. In [204], a complex spectral subtraction method was used to benefit from phase information in noise-robust automatic speech recognition. The robustness analysis of the group delay-based features against additive or convolutive noise and solutions to circumvent the sensitivity to noise have been studied for speaker recognition [205, 206].

Feeding a raw speech signal to a deep neural network and convolutive neural network is an implicit feature extraction

\[^{13}\text{For small estimation errors it is well resembling the squared-error distortion measure, since } 1 - \cos(\hat{\phi}_s(k,l) - \phi_s(k,l)) = \frac{(\hat{\phi}_s(k,l) - \phi_s(k,l))^2}{2} \text{ for } \hat{\phi}_s(k,l) - \phi_s(k,l) \ll 1.\]
procedure in which amplitude and phase information have both been utilized [207, 208]. Finally, features derived from Hilbert transform has been considered as a way to utilize both amplitude and phase information in a unified way for speech recognition [209].

As phase-aware features, Modified group delay and instantaneous frequency (IF) as frequency and time derivate of spectral phase have been used for ASR [2, 210]. With appropriate windowing, smoothed group delay functions can be computed as shown in [211]. Therefore, by employing glottal closure instances (GCI), Bozkurt and Couvreur proposed group delay of GCI-synchronously windowed speech (GDGCI) for ASR [81]. An alternative way to smooth group delay is to utilize chirp group delay of the zero-phase version (CGDZP) which converts the signal to its zero-version where all zeros occur very close to the unit circle, hence, well resolved formant peaks are provided. Saratxaga et al. [212] used harmonic phase information as relative phase shift (RPS) and showed improved ASR performance compared to classical MFCC parameterization.

4.5.2. Speaker Recognition

Most of the automatic speech and speaker recognition systems are built upon short-term spectro-temporal feature representations, typically calculated from the amplitude spectrum [213]. The amplitude spectrum can be calculated as the magnitude of a complex Fourier transform or other parametric and non-parametric spectrum estimation methods, including linear prediction, multi-tapering and their variants [214–216]. The Mel-frequency cepstral coefficients are among the most popular features derived by applying a perceptually weighted filter-bank to the amplitude spectrum.

In anti-spoofing for speaker recognition systems, phase consistency is being used as a counter measure for synthetic speech detection [217, 218]. Several phase-aware techniques for anti-spoofing methods in speaker recognition were studied in [219]. Wang et al. [220] considered relative phase information to discriminate natural human speech from machine-generated spoofed speech. Xiao et al. [221] studied five types of phase-based features: group delay, modified group delay, instantaneous frequency deviation, baseband phase difference, and pitch-synchronous phase for spoofing speech detection. Liu et al. [222] proposed several spoofing countermeasures based on magnitude and phase information for speaker verification.

The robustness of features derived from group delay against transmission channel and ambient noise has been investigated [59, 78, 205]. The modified group-delay (MGD) function followed by discrete cosine transform (DCT) has been used [64] for a comparative evaluation of speaker recognition on TIMIT and noisy TIMIT corpora. When calculating MGDF, it has been assumed that the speech signal is minimum phase. Acoustic features based on AM-FM representation of the speech signal have been included in different recognition applications. The AM-FM model is generally used to describe a band-pass speech signal by its envelope (instantaneous amplitude) and phase (instantaneous frequency). Hence, the speech signal is first passed through a filterbank [223–227] to decompose it into narrow-band signal that can be approximated locally with a single sinusoid. Features derived from the analytic phase of speech signal have been used for robust speaker recognition [228]. By taking the derivative of instantaneous phase, the instantaneous frequency cepstral coefficients (IFCC) are derived from smoothed sub-band instantaneous frequencies. Similarly, amplitude weighted instantaneous frequency (AWIF) features [73] have been built upon characterization based on speech signal pyknograms (density plot of the frequencies present in the input signal) [229]. The application of instantaneous frequency (and its deviation) along with the delta-phase spectrum have been used for feature extraction in order to account for the large variability of phase caused by the starting point of analysis window [22, 198]. All-pass modeling of LP spectrum is considered for speaker recognition [230]. However, the residual phase feature alone would not characterize the complete phase information in the speech signal. A parametric model on the time-domain phase was proposed in [231], where all-pass cepstral coefficients (APCC) were calculated from time-domain phase after normalizing out the amplitude spectrum. A common way to utilize phase-derived features is to directly combine the features derived from amplitude and phase individually [232]. The amplitude and phase information has been combined on the feature- or score-levels [233–236]. Features derived from Hilbert transform has been considered as a way to utilize both amplitude and phase information in a unified way for speaker recognition [233, 237].

4.6. Speech Synthesis

As in other areas of speech technology where the target listener is human (i.e., speech enhancement), text to speech synthesis application also benefits from phase-aware signal processing. Currently, there are three main approaches for
speech synthesis: a) unit selection, b) statistical parametric speech synthesis, and c) hybrid approaches. Among these three approaches, the approach based on unit selection is in less demand in signal processing. The hybrid approach requires some signal processing especially for feature extraction from the speech signal. During synthesis, hybrid approaches may not use these features to generate the final synthetic speech signal since the approach is then similar to the concatenation of units, selected using the same mechanism as in the unit selection approach. However, there are hybrid systems that use generated units by the acoustic model features to synthesize parts of the final synthetic speech. Finally, the statistical parametric speech synthesis systems are currently in heavy demand in signal processing. This is both in terms of speech analysis (speech modeling) and during synthesis, where the parameters of the speech model (given a text) are generated by maximizing the likelihood of a probabilistic acoustic model learned during the training of the system. Although unit selection based approaches provide a high quality of synthetic speech, they suffer from not being flexible in terms of modifying emotions of the generated speech and creating a customized voice without requiring too many new recordings. On the other hand, statistical parametric speech synthesis systems are quite flexible but the synthetic speech quality they generate is currently not as good as the one in unit selection. Recent advancements in statistical parametric speech synthesis, however, using deep learning for the acoustic model, show that the quality of synthetic speech using that approach can be improved and be competitive to the unit selection approach [238]. Hybrid systems exist between these two approaches both in terms of their resulting speech quality and flexibility. We show below that, independent of the approach taken for speech synthesis, phase information is important either for increasing quality or enabling flexibility.

Although unit selection does not use many speech models, it has been shown that pre-processing units in the speech database can facilitate the concatenation of units during the synthesis. Units are usually phone or even half-phone [239]. Time alignment of these units during their concatenation is important to remove linear phase mismatches during synthesis, which are perceived as annoying clicks. Cross-correlation can be used during synthesis to alleviate the linear phase problem, but at the expense of an increased computational cost. Another solution is to mark the glottal closure instants (GCIs) in the units and then align the units based on these time instants. Although GCIs detection is prone to errors especially for emotional data and, for speakers with high pitch, the databases used in this approach are of large size, such an approach may still result in poor linear phase alignment. In [240], the audible clicks due to the linear phase mismatches at the frame boundaries were studied and a common reference point, referred to as the center of gravity, for all the speech frames (usually two pitch periods long) was defined. It was shown that the center of gravity of each frame could be simply approximated by the phase of the first harmonic component of the signal [240]. A simpler model relying on linear phase subtraction from all harmonic phases was shown to work very well when pre-processing the frames in the speech database, synthesizing speech with linear phase mismatches, without the need to estimate GCIs or compute cross-correlation functions during synthesis. As a proof-of-concept example, Figure 6 shows speech frames (two pitch periods long) from various instances in the database before and after alignment using the linear phase correction mechanism, described above. The left column of the figure shows the different position of the analysis window before linear phase correction, while the right column shows it after linear phase correction. For further details we refer to [240].

To increase the flexibility of unit selection-based text-to-speech synthesis systems, techniques in voice conversion could be of advantage [241]. However, in that case, speech modeling is required. To have a successfully perceived converted speaker, both in the sense of similarity to the target speaker as well as of quality of the converted signal, approaches using phase information as mentioned in the previous sections for the relationship between amplitude and phase, and speaker identification should be considered.

Hybrid systems require more signal processing than unit selection-based systems, especially during training. Spectral magnitude based information is often used in that case (i.e., cepstral based representations) and phase information is not used. During synthesis, spectral magnitude information is used to select units from a speech database, in a similar way as in the unit selection systems. However, there are instances when speech signal should be synthesized using the generated spectral magnitude and fundamental frequency information. The latter is used to create the linear phase term (as shown in (12)) while the vocal tract based phase given in (12) is computed by the magnitude spectrum. This results in a phase that consists of a minimum phase term plus a linear phase term with the dispersion phase part missing\textsuperscript{14}, resulting in a synthesized signal with a buzzy perceived quality. In terms of signal processing, this quality is due to a highly unnatural correlation between the generated speech samples. To overcome this problem, the band

\textsuperscript{14}following the harmonic model with phase decomposition presented in Eq. (12)
aperiodicity (BAP) information can be used [242]. This is mainly an SNR measurement in certain frequency bands. BAP is used to insert noise in some frequency bands. This has the effect of reducing the correlation and therefore reducing the buzziness effect. In essence, this is an ad-hoc solution for not being able to model the dispersion phase in (12) accordingly. In order to increase the flexibility of the hybrid systems, the same issues as for the unit selection based systems, mentioned above, need to be addressed, and phase information again will play an important role for improving current results.

Finally, the statistical parametric-based speech synthesis systems face the same problems as the hybrid ones. However, in this case, during synthesis, all the generated parameters are used for synthesizing a speech signal. Therefore, the problems of buzziness mentioned above for the Hybrid systems, are more frequent and even more pronounced than for the hybrid ones. These observations have encouraged many researchers in speech synthesis to look for new ways to model phase information. It has been widely accepted that better phase models will boost the quality of synthetic speech in particular in the current statistical parametric speech synthesis systems. Towards that goal, Maia et al. [243] proposed a complex cepstrum approach trying to model amplitude and phase information together. In more recent work on complex cepstra [244], Maia et al. studied complex cepstrum-based speech factorization for acoustic modeling in statistical parametric synthesizers. Throughout the investigation of various complex cepstrum factorization approaches, it has been found that the factorization of minimum phase and all-pass filters provided the best results in terms of quality of the synthetic speech. More recently, Agiomyrgiannakis [99] presented an AM-FM sinusoidal based approach for statistical parametric speech synthesis, where manipulation of phase spectrum was shown to be critical to boost the quality of the synthetic speech.

All the above observations showed that, for speech synthesis, phase-aware processing is crucial to improve the quality of synthetic speech and highly flexible speech synthesis systems. However, optimal modeling of phase information in speech synthesis is still an active research topic.

5. Conclusions and Future Work

In this paper, we presented an overview of the earlier and more recent advances made towards incorporating phase information in the processing of speech signals. We exemplified the phase-aware signal processing contributions made by researchers in diverse fields of speech processing applications including: speech analysis, speaker and speech recognition, speech enhancement, source separation, speech coding, watermarking, and speech synthesis. Throughout this review, it has been concluded that phase-awareness has had positive impact on the achievable performance as conventionally reported by amplitude-only signal processing methods during the last decades.

With the increasing interests of the researchers in the community towards phase-aware signal processing in speech
communication, it is deemed that synergy will raise among the researchers actively working on phase-aware signal processing in different applications of speech communication. The attempt in this work was targeted to unify the existing scattered evidence in the literature with regard to phase-aware signal processing, where researchers individually incorporated phase-awareness in different applications. In the future, works follow up to the current special issue series, could benefit from an increased, improved overall understanding regarding the existing proposals to tackle the controversial topic of phase processing. From a practical viewpoint, several examples regarding the positive impact of phase-aware signal processing in different speech applications delivers the big picture regarding how the current state of knowledge could be used as the starting reference to continue working towards new proposals in the emerging field of phase-aware speech signal processing, with a high potential of pushing the limits of the conventional state-of-the-art methods.

Acknowledgment

The work of Pejman Mowlaee was supported by the Austrian Science Fund (project number P28070-N33). The work of Rahim Saeidi is supported by the Academy of Finland (project number 256961 and 284671).

References


