Master Thesis

Phase Estimation using Time-frequency Constraints

conducted at the
Signal Processing and Speech Communications Laboratory
Graz University of Technology, Austria

by
Michael Pirolt, 1073101

Supervisor:
Dr. Pejman Mowlaee

Assessor/Examiner:
Dr. Pejman Mowlaee

Graz, January 17, 2017

Acknowledgements

First of all, I would like to thank my supervisor Dr. Pejman Mowlaee for the support throughout the last year. His relentless dedication towards research and his helpful comments kept me on the right track with my work.

Furthermore, my gratitude goes to my colleagues for the time spent together, which helped me to stay focused and motivated.

Last, but definitely not least, I would like to thank my friends, my family and especially my girlfriend Johanna, for supporting and encouraging me in any possible way. I am grateful that you were always there for me.

Abstract

Single-channel speech enhancement is an essential part in different speech based applications such as automatic speech recognition, mobile telephony and hearing aids. Throughout the years, many different speech enhancement methods have been developed, most of them are formulated in the Short-Time Fourier Transform (STFT) domain. The majority of the conventional STFT-based speech enhancement methods aim to enhance the amplitude only, while the noisy spectral phase is left unprocessed. In this thesis a novel phase enhancement method which exploits the relation between the spectral phase at harmonics in speech, is presented. After discussing the fundamentals of phase processing, prominent existing phase enhancement methods from the past are explained. Among others, a novel phase representation, the Phase Quasi Invariant (PQI), is introduced. Based on this phase representation, two enhancement methods which exclusively modify the spectral phase, are introduced. In experiments, the effectiveness of these novel enhancement methods are demonstrated by comparing them to phase enhancement benchmarks. The performance evaluation is conducted by means of objective measures for speech quality, intelligibility and phase estimation error. All experiments and algorithms included in this thesis were implemented in MATLAB.

Kurzfassung

Einkanalige Sprachsignalverbesserung ist ein wichtiger Bestandteil automatischer Spracherkennung, Mobiltelefonie und Hörgeräten. Über die Jahre wurde eine Vielzahl von Sprachverbesserungsalgorithmen entwickelt, deren Mehrheit auf der Kurzzeit-Fourier-Transformation basiert. Die meisten konventionellen STFT-basierten Sprachverbesserungsmethoden bearbeiten nur die spektrale Amplitude, während die Signalphase unbearbeitet bleibt. In der vorliegenden Arbeit wird eine neue Phasenverbesserungsmethode beschrieben, welche auf den Phasenverhältnissen zwischen harmonischen Schwingungen in Sprachsignalen basiert. Zunächst werden die Grundlagen der Phasenverarbeitung besprochen und verschiedene effektive Phasenverbesserungsmethoden beschrieben. Neben anderen Phasenrepresentationen, wird der Phase Quasi Invariant (PQI) vorgestellt. Auf Basis dieser Phasenrepräsentationen werden zwei Sprachverbesserungsmethoden, welche ausschließlich die spektrale Phase verbessern, eingeführt. In Experimenten wird die Effektivität dieser neuen Phasenverbesserungsmethoden demonstriert, indem sie mit Referenzmethoden verglichen werden. Zur Evaluierung werden objektive Maße für Sprachqualität und Sprachverständnis herangezogen. Alle in dieser Arbeit enthaltenen Experimente und Algorithmen wurden in MATLAB implementiert.

Statutory Declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly marked all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present Master's thesis.

date	(signature)

Contents

1 Introduction

Speech is the most natural way of human communication. As advances in signal processing technologies proceed, more and more listening devices find their way into our daily lives. Different speech communication applications such as digital telephony, acoustic human-machine communication and digital hearing aids are expected to function as accurately as possible to guarantee a reliable speech communication experience. Therefore, robust performance must be ensured for noise conditions of everyday life, such as driving in a car, walking along the street, being in a restaurant or factory, just to name a few. Additionally, the performance of any speech processing device gets aggravated through distortions of the communication channel, caused by acoustic echoes or room reverberations. As a result, the desired clean speech signal is often only accessible as a corrupted, noisy version [1].

Designing a speech algorithm that deals with all these problems is a challenging as well as rewarding task. A conventional single channel speech communication chain, consisting of several blocks, each targeting one processing step, is depicted in Figure 1.1.

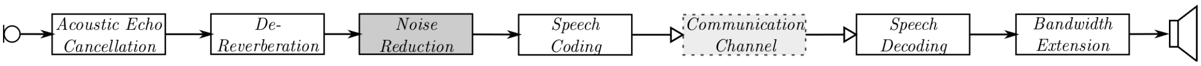

Figure 1.1: Block diagram of a conventional speech communication system, from transmitter (microphone) to receiver end (loudspeaker), consisting of the blocks: beamforming, acoustic echo cancellation, de-reverberation, noise reduction, speech coder and artificial band width extension

For years, the most popular tool to analyse and modify the signals at the individual stages has been the short-time Fourier Transform (STFT), where the signal is often represented by sum of amplitudes, frequencies and phases.

The majority of the literature has been dedicated towards modification of the spectral amplitude only, while the importance of spectral phase has been a controversial topic in the speech processing community. While early studies report the insignificance of the phase spectrum in terms of perception [2, 3], more recent publications highlight the importance of the spectral phase for different speech processing applications such as: speech coding [4], speech recognition [5,6], source separation [7–9] and speaker recognition [10,11].
In particular, the field of noise reduction has received increasing attention by researchers lately. Some examples are model-based short-time Fourier transform phase improvement [12], maximum a posteriori harmonic (MAP) phase estimation [13], as well as temporal smoothing of the unwrapped harmonic phase [14]. Apart from improved signal reconstruction, spectral phase information can also be used to derive improved spectral amplitude estimators, see [15, 16]. The advances of phase-aware processing are limited by the accuracy of the estimated phase. Therefore, it is a challenging research topic to find novel approaches that help to achieve more robust and accurate estimators of the clean spectral phase given the noisy speech observation. In this thesis a novel single channel phase estimation technique, which exploits the relation between the harmonic phases of a speech signal, is discussed.

The thesis is structured as follows. Chapter 2 gives an overview about the fundamentals of STFT-based speech enhancement. In Chapter 3, some of the most relevant phase enhancement methods are presented and explained in detail. Chapter 4 discusses different phase representations other than the instantaneous phase, which are helpful to reveal harmonic structures in phase. In Chapter 5, a phase enhancement method based on exploiting the relation between the phase spectrum of harmonics is presented. Chapter 6 shows the experiment setup and presents the results for the proposed phase enhancement method. Chapter 7 concludes the work and gives a future outlook.

2

Fundamentals of Phase in Speech Enhancement

2.1 Speech Production Process

Before talking about phase enhancement algorithms, it is important to gain some insight about speech signals in general. Human speech is the result of a complex interaction performed by various physiological components. A contraction of the aspiratory muscles forces the air enclosed in the lungs to exit, which results in an air flow through the trachea up to the glottis. The glottis, which denotes the opening between the vocal cords and the larynx, is considered as the excitation-source signal, as it converts the airflow into acoustic oscillation. The frequency of this oscillation can range from 50 Hz to 250 Hz for male speakers and from 120 Hz to 500 Hz for female speakers and is known as the fundamental frequency f_0 [17].
Additionally, the vocal chords produce a larger amount of harmonics by colliding with themselves. Harmonics are normally integer multiples of the fundamental frequency. The next major part in the speech production chain is the vocal tract consisting of the oral and the nasal cavity. The vocal tract acts as an acoustic resonator that spectrally shapes the source excitation signal. Throughout this process, different harmonics are emphasized, depending on the shape of the vocal tract. Finally, speech is emitted through lips and nostrils, where it is also dynamically shaped [18]. A schematic of the involved human speech organes can be observed in Figure 2.1.

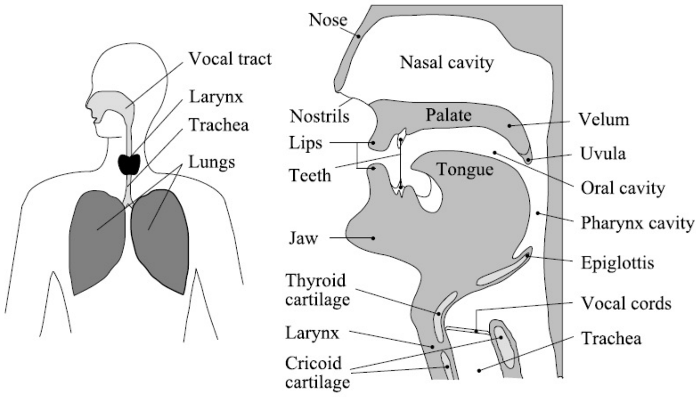

Figure 2.1: Human speech production organs [1]

Not all speech sounds are produced upon a periodic oscillation of the vocal cords. Unvoiced sounds, like /f/, /s/ or /sh/ are produced with relaxed vocal cords, resulting in an excitation signal which can be described as white Gaussian noise. Therefore, no harmonic structure or fundamental frequency is present. Plosive sounds, like /p/ or /t/ are created after building up pressure in the oral cavity. In a burst, the build-up air is released causing a sound which is perceived as a transient burst of noise [17].

2.2 Fundamentals of the Fourier Transform

Speech signals can be recorded in many different ways, depending on the associated purpose. Speech enhancement methods are either based on multi-channel methods, where the signal is captured by multiple microphones, or single-channel methods, where the signal is captured by one single microphone. This thesis focuses on single-channel applications. In order to develop an efficient speech enhancement system a method is used to access certain properties and characteristics of the speech captured, which are not available in time domain [19]. This calls for a different signal representation. To that extent, an analysis-modification-synthesis (AMS) framework depicted in Figure 2.2 is used. With the help of a signal transform, the input signal is transformed into another domain, where certain modification can be performed, followed by a re-synthesis step.

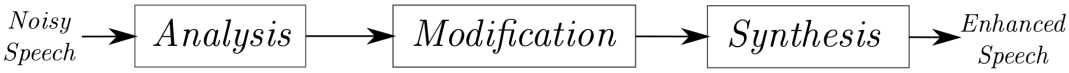

Figure 2.2: Block diagram of a basic AMS model

Probably the most established signal transform for the purpose of speech enhancement is the Fourier transform. The Fourier transform is a spectral transform, which analyses a signal in terms of its spectral components [18]. With the help of the Fourier transform, a continuous time signal $x_a(t)$ is related to its frequency domain representation $X_a(j\omega)$:

$$X_a(j\omega) = \int_{t=-\infty}^{\infty} x_a(t)\mathrm{e}^{-j\omega t}dt, \tag{2.1}$$

where t denote the continuous time index and $\omega = 2\pi f$ the frequency in radiants, respectively. The Fourier domain representation is continuous in both time and frequency. In signal processing, though, we deal with a digital signal $x(n)$ with discrete time index n. Given the analog continuous time signal $x_a(t)$, the corresponding discrete time signal $x(n)$ is obtained by sampling with the sampling period T_s [17]:

$$\begin{aligned} x(n) &= x_a(t)\big|_{t=nT_s} \qquad\qquad -\infty < n < \infty \\ &= x_a(nT_s). \end{aligned} \tag{2.2}$$

From the sampling period T_s the sampling frequency can be derived as:

$$f_s = \frac{1}{T_s}. \tag{2.3}$$

In order to process discrete time-signals, the discrete-time Fourier transform (DTFT) is introduced:

$$X(\mathrm{e}^{j\Omega}) = \sum_{n=-\infty}^{\infty} x(n)\mathrm{e}^{-j\Omega n}, \tag{2.4}$$

where $\Omega = 2\pi f T_s$ denotes the continuous normalized angular frequency. In order to calculate the DTFT of a signal properly, an input signal with infinite length would be needed. In practice, however, the input signal $x(n)$ is finite in duration, consisting of N samples. The finite amount of samples leads to uniformly spaced frequencies in the Fourier domain, yielding a new transform referred to as the discrete Fourier transform (DFT) [17].

The DFT of the signal $x(n)$ is given by:

$$X(k) = \sum_{n=0}^{N-1} x(n) \mathrm{e}^{-j\frac{2\pi kn}{N}} \qquad 0 \leq k \leq N-1, \tag{2.5}$$

where N denotes the length of the input signal and k the frequency index. The DFT represents the signal spectrum at equally spaced points on the normalized frequency axis $\Delta\Omega = \frac{2\pi}{N}$. Taking the sampling frequency into account, this means that the frequency components of the DFT are spaced apart by $\Delta f = \frac{f_s}{N}$, Δf is also referred to as the frequency resolution [1]. The DFT is discrete in time and frequency.

Generally, speech signals are highly non-stationary signals. On the closer look, however, certain speech segments are considered to be quasi-stationary. Therefore, it is common to process speech signals frame-wise. Commonly, frame durations with a length of 20 ms - 40 ms are used [18]. After the signal segmentation, a window function is applied, where the length of the window determines the length of the signal frame. This helps to deal with the problem of spectral leakage. To some extent, time consecutive values of $X(k)$ contain redundancy, which is caused by redundancy some redundancy in consecutive values of $x(n)$. Therefore, it is common to introduce a frame shift Z, which means that $X(k)$ is computed only every Z-th sample. Taking this into account, the short-time Fourier transform (STFT) can be defined as:

$$X(k,l) = \sum_{n=0}^{N-1} x(n+Zl) w(n) \mathrm{e}^{-j\frac{2\pi kn}{N}}, \tag{2.6}$$

where l denotes the frame index and $w(n)$ the window function with $w(n) \neq 0$ for $0 \leq n \leq N-1$.

As any Fourier transform, the short-term Fourier transform of a time domain signal is a complex valued function of frequency. This means it can either be expressed in terms of its real and imaginary parts

$$X(k,l) = Re\left\{X(k,l)\right\} + jIm\left\{X(k,l)\right\}, \tag{2.7}$$

or in terms of its magnitude and phase spectra using polar form

$$X(k,l) = |X(k,l)| \mathrm{e}^{j\angle X(k,l)}, \tag{2.8}$$

where $|X(k,l)|$ denotes the magnitude spectrum and $\angle X(k,l)$ denotes the phase spectrum. The STFT domain allows perfect reconstruction and is considered to be efficient in terms of computational complexity, therefore the STFT is a widely used transform in speech communication systems. To study other properties and the synthesis of the Fourier transform in detail, the reader is referred to [1], as it is out of the scope of this thesis.

2.3 Conventional Speech Enhancement Methods

In single channel speech enhancement it is often assumed that a clean signal is deteriorated by additive noise yielding a noisy observation

$$y(n) = x(n) + d(n), \tag{2.9}$$

where $y(n)$, $x(n)$ and $d(n)$ denote the noisy speech, the clean speech and the noise signal, respectively. It is common to consider the signals as realizations of stochastic processes which are assumed to be statistically independent. Therefore, the problem definition can be formulated as follows:
Given a realization of the noisy speech signal $y(n)$, find an estimate of the clean speech $x(n)$.

Throughout the years, a variety of speech enhancement methods have been developed. They differ from each other in terms of the domain they are formulated in (e.g. frequency, time, modulation) and in terms of assumptions they assert about the statistics of the signal realizations. Basically, they can be divided into the following categories:

- Spectral subtraction algorithms [20]
- Statistical-model based methods [21] and Wiener filtering methods [22]
- Subspace Algorithms [23]

Besides other modulation based signal domains, the STFT has been the predominant choice for speech enhancement methods. While only few methods tend to modify the real and imaginary STFT components directly [24], the majority of the methods tend to modify the signal in terms of their magnitude and phase spectra. Usually they follow an AMS model similar to the one depicted in Figure 2.3.

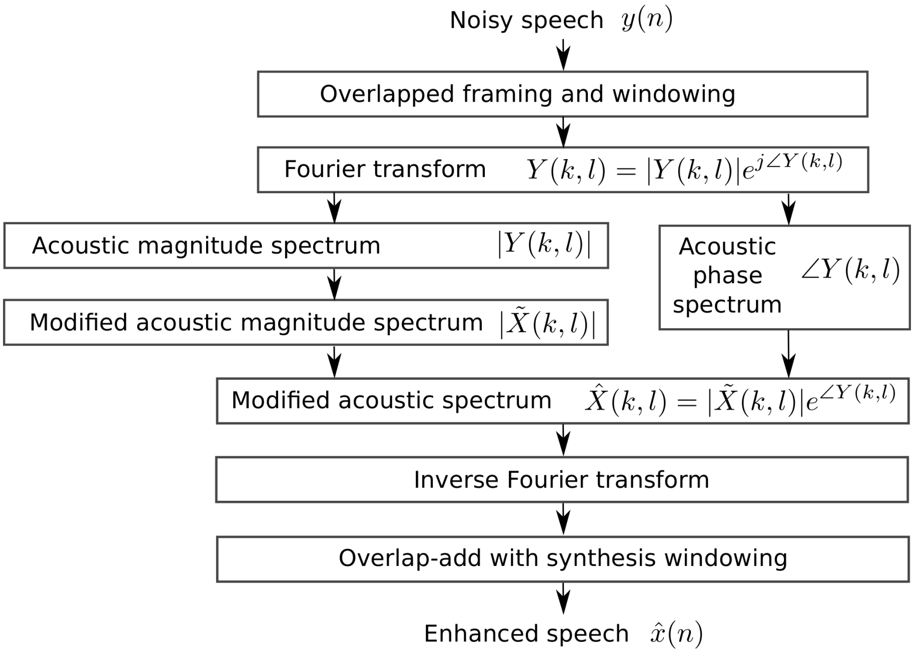
Noisy speech $y(n)$
Overlapped framing and windowing
Fourier transform $Y(k,l) = |Y(k,l)|e^{j\angle Y(k,l)}$
Acoustic magnitude spectrum $|Y(k,l)|$
Acoustic phase spectrum $\angle Y(k,l)$
Modified acoustic magnitude spectrum $|\tilde{X}(k,l)|$
Modified acoustic spectrum $\hat{X}(k,l) = |\tilde{X}(k,l)|e^{\angle Y(k,l)}$
Inverse Fourier transform
Overlap-add with synthesis windowing
Enhanced speech $\hat{x}(n)$

Figure 2.3: Block diagram of an STFT-based AMS framework for enhancement of the spectral magnitude [18]

This means most speech enhancement algorithms, e.g. [20, 21, 25–27], only aim to enhance the spectral amplitude, whereby the estimation methods of $|\tilde{X}(k,l)|$ differ from one another. But the spectral phase, which is clearly distorted as well, is likely to be left unprocessed and used for signal synthesis.

The main reason for this is the intuitive nature of the spectral magnitude. A lot of signal information, such as spectral energy distribution or pitch information, is easily accessible by observing the spectral amplitude. This can be observed in the Figures 2.4 and 2.5, where the magnitude and phase spectra of two windowed speech segments of a female speaker are depicted. Figure 2.4 shows a voiced speech segment, Figure 2.5 shows an unvoiced segment.

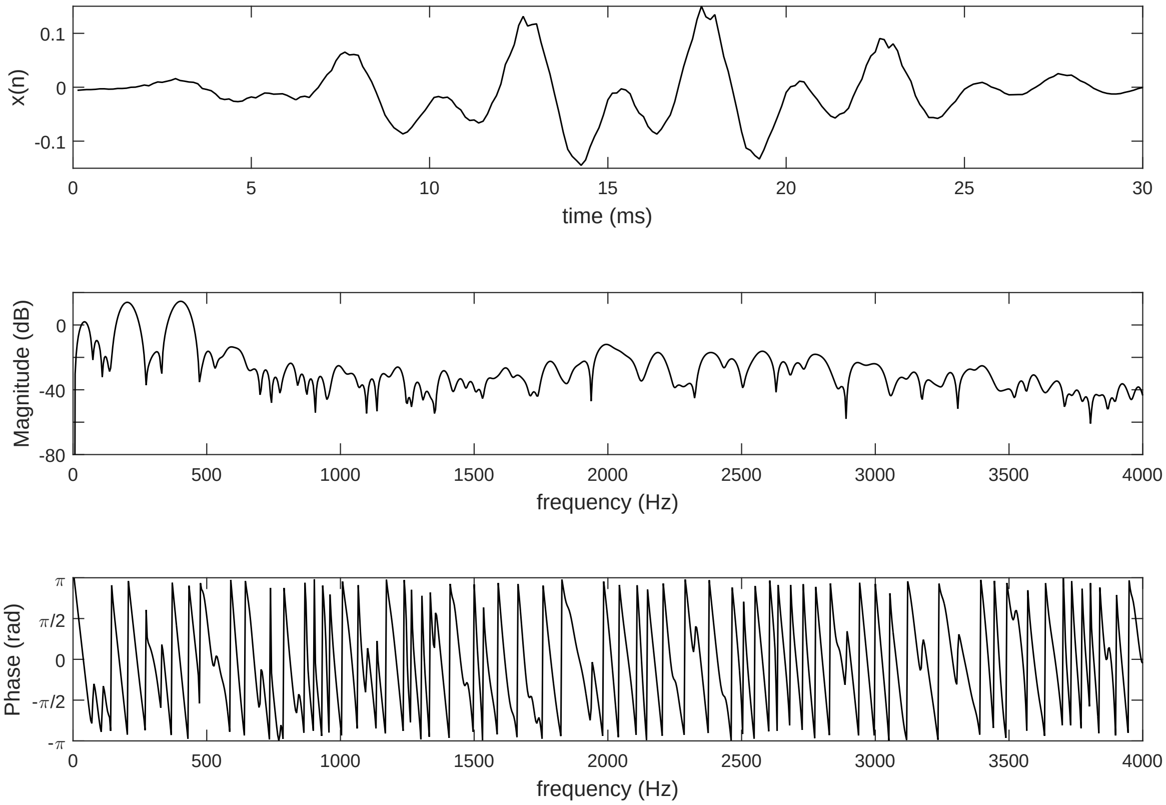

Figure 2.4: Time and frequency representation of a Hamming windowed voiced speech segment with 30ms length. Shown are time signal (top), magnitude spectrum in db (middle) and phase spectrum (bottom)

For voiced segments, the evenly spaced peaks in the magnitude spectrum correspond to pitch harmonics, while peaks in higher amplitude regions often correspond to formants. In unvoiced segments, where no periodic excitation signal is present, the pitch harmonics disappear. Compared to the magnitude spectrum, the phase spectrum in the bottom of Figure 2.4 and Figure 2.5 does not seem to show any structure. Independent of the excitation signal, it seems to change randomly across frequency. Therefore, it was believed to contain little useful information [2]. On the closer look, however, this noise-like behaviour can be attributed to the cyclic wrapping of the spectral phase. Chapter 4 addresses this problem, and different methods to uncover useful structures from the instantaneous phase are presented.

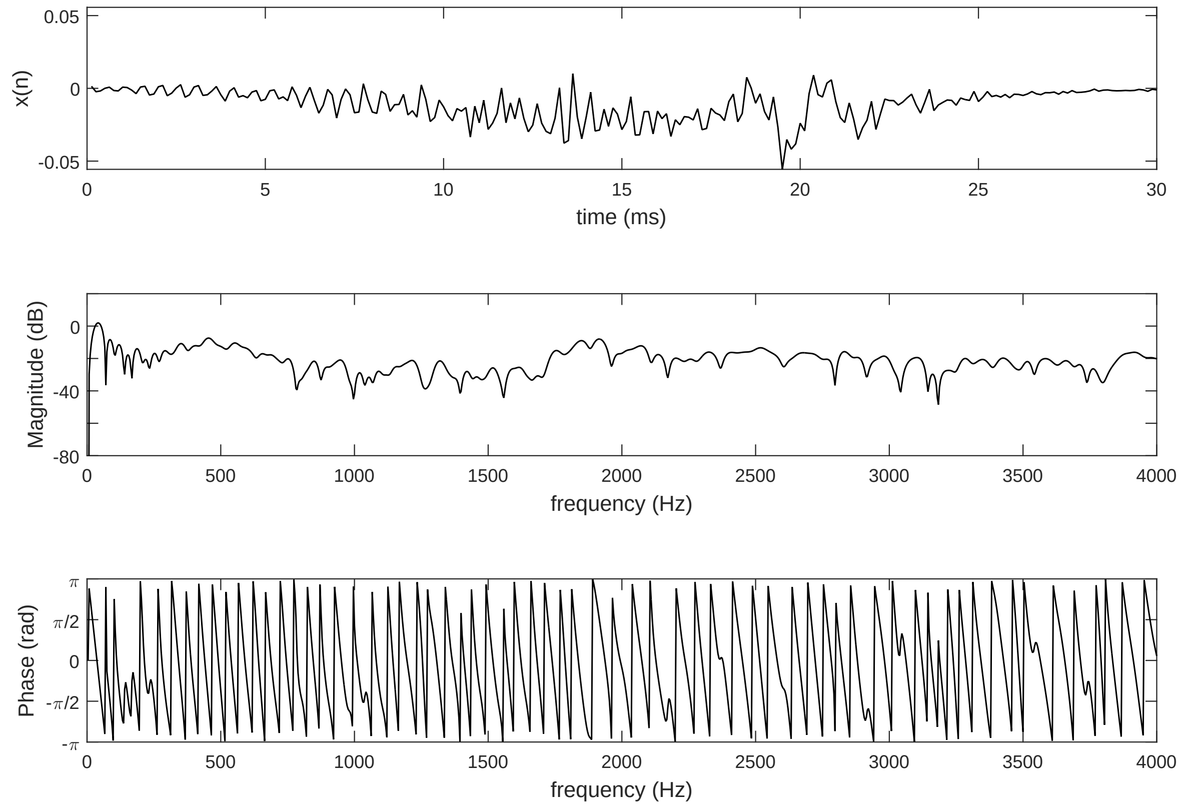

Figure 2.5: Time and frequency representation of a Hamming windowed unvoiced speech segment with 30ms length. Shown are time signal (top), magnitude spectrum in db (middle) and phase spectrum (bottom)

Another reason why the phase spectrum has often been discarded in speech enhancement frameworks was the early work of Wang and Lim [3] and Vary [28]. Both observed how the perceived quality of a clean speech signal changes when the spectral phase is distorted. Both works reported that the possible achievable gain through phase enhancement is small compared to the one gained through amplitude enhancement and therefore can be neglected for high SNRs.

After decades it was Paliwal et al. in [29] who addressed this topic again. They conducted a couple of experiments that disputed with the earlier observations of Wang and Lim. Paliwal reported that estimating a better phase spectrum, while a challenging task, could be worthwhile, as accurately estimated phase spectra have the potential to significantly enhance the speech quality. For a detailed overview on why the phase has been neglected in speech enhancement for so long the reader is referred to [30].

3 Phase Estimation Methods

This Chapter presents some of the phase estimation algorithms that have been introduced over the years. Most of them are based on a harmonic signal model. In the following, each method and its central idea will be discussed.

3.1 Short-Time Fourier Transform Phase Improvement

One method to improve the noisy spectral phase was introduced in [31] by Krawczyk and Gerkmann. It is based on exploiting the correlation between neighbouring phase values across time and frequency and the spectral amplitudes in voiced speech. The algorithm can therefore be performed along frequency or along time, a combination of both is also possible. The method was labeled as STFT phase improvement and is further referred to as STFTPI.

The algorithm is based on a harmonic signal model, where a voiced speech sound is modelled as a weighted sum of sinusoids at integer multiples of the fundamental frequency. It is given as:

$$x(n) = \sum_{h=1}^{H} 2A(h) \cdot \cos\Big(\omega(h) \cdot n + \psi(h)\Big), \tag{3.1}$$

with the real-valued amplitude $2A(h)$, the harmonic phase $\psi(h)$ and the normalized angular frequency, which is given as:

$$\omega(h) = 2\pi \frac{f_h}{f_s} = 2\pi \frac{hf_0}{f_s} = 2\pi \cdot hf_0 T_s. \tag{3.2}$$

The sum of harmonics from (3.1) is transformed into STFT domain using the Equation (2.6), resulting in $X(k,l)$. Then $X(k,l)$ can again be transformed into its baseband by using

$$X_B(k,l) = X(k,l)e^{-j\omega_k Zl}, \tag{3.3}$$

with center frequency $\omega_k = \frac{2\pi k}{N}$ and baseband STFT representation $X_B(k,l)$. By using Equations (3.1) and (3.3), we obtain the baseband STFT representation for the voiced speech model $x(n)$, which denotes as:

$$X_B(k,l) = \sum_{n=0}^{N-1} w(n) \sum_{h=1}^{H_l} A(h,l) \left(\mathrm{e}^{j\Big((\omega(h,l)-\omega_k)(n+Zl)+\psi(h,l)\Big)} + \mathrm{e}^{-j\Big((\omega(h,l)-\omega_k)(n+Zl)+\psi(h,l)\Big)} \right). \tag{3.4}$$

Following Equation (3.4) each frequency bin k is depending on every harmonic. When the length of a signal segment, N, and therefore also the window length, is high enough to get a good frequency resolution, it is safe to assume that each STFT bin k is dominated only by the closest complex component. The harmonic component is given as:

$$\omega_h^k = \arg\min_{\omega(h,l)} (|\omega_k - \omega(h,l)|). \tag{3.5}$$

Note that $\omega_k N/(2\pi)$ is an integer, while $\omega_h^k N/(2\pi) \in \mathbb{R}$. This means that the harmonic frequency is not necessarily identical to the center frequency of the DFT. The assumption from above reduces Equation (3.4) and leads to a formulation of the baseband STFT amplitude and phase:

$$\begin{aligned}
X_B(k,l) &\approx A(h,l) \sum_{n=0}^{N-1} w(n) \mathrm{e}^{j((\omega_h^k - \omega_k)(n+Zl)+\psi(h,l))} \\
&\approx A(h,l) \mathrm{e}^{j(\omega_h^k-\omega_k)Zl} \mathrm{e}^{j\psi(h,l)} \underbrace{\sum_{n=0}^{N-1} w(n) \mathrm{e}^{j(\omega_h^k-\omega_k)n}}_{|W\left(k-\omega_h^k \frac{N}{2\pi}\right)| \mathrm{e}^{j\psi_W\left(k-\omega_h^k \frac{N}{2\pi}\right)}} \\
&\approx \underbrace{A(h,l) \left| W\left(k - \omega_h^k \frac{N}{2\pi}\right)\right|}_{|X_B(k,l)|} \underbrace{\mathrm{e}^{j\left(\psi(h,l)+(\omega_h^k-\omega_k)Zl+\psi_W\left(k-\omega_h^k \frac{N}{2\pi}\right)\right)}}_{\psi_{X_B}(k,l)},
\end{aligned} \tag{3.6}$$

where $|W\left(k - \omega_h^k \frac{N}{2\pi}\right)| \mathrm{e}^{j\psi_W\left(k-\omega_h^k \frac{N}{2\pi}\right)}$ denotes the contribution of the analysis window, modulated by the frequency of the dominant harmonic.

3.1.1 Phase Reconstruction across Time

Based on Equation (3.6), a formula for a recursive segment-to-segment computation of the baseband STFT-phase $\psi_{X_B}(k,l)$ is derived. Additionally, we assume that the fundamental frequency f_0 is changing slowly from one frame to another, which means that:

$$\omega_h^k(l-1) \approx \omega_h^k(l). \tag{3.7}$$

The simplified version of the baseband STFT phase difference is given as:

$$\Delta\psi_{X_B}(k,l) = \psi_{X_B}(k,l) - \psi_{X_B}(k,l-1) = \left(\omega_h^k - \omega_k\right) Z. \tag{3.8}$$

Which can be reformulated to obtain the recursive formula for phase reconstruction across time:

$$\psi_{X_B}(k,l) = \psi_{X_B}(k,l-1) + \left(\omega_h^k - \omega_k\right) Z. \tag{3.9}$$

Equation (3.9) suggests that the baseband phase change depends only on the difference between the closest harmonic ω_h^k, the STFT center-frequency ω_k and the frame shift Z. Given an initial estimation of the STFT baseband phase $\psi_{X_B}(k,l_0)$, the reconstruction is conducted for all harmonics across the frequency bands.

3.1.2 Phase Reconstruction across Frequency

The phase reconstruction along frequency follows the same assumptions as the phase reconstruction across time. Given a phase estimate of the STFT-band k', which is evaluated following

Equation (3.9), Equation (3.6) can be rewritten as:

$$X_B(k',l) \approx A(h,l)\mathrm{e}^{j(\omega_h^{k'}Zl+\psi(h,l))} \underbrace{\mathrm{e}^{-j\omega_{k'}Zl}W(k'-\omega_h^{k'}\frac{N}{2\pi})}_{f(k')}, \tag{3.10}$$

where only $f(k')$ depends on STFT frequency bin k'. Following this, the phase of the frequency bin $k'+\Delta k'$ can be estimated based on the reference of $\psi_{X_B}(k',l)$:

$$\psi_{X_B}(k'+\Delta k',l) = \psi_{X_B}(k',l) - \Delta k'\frac{2\pi}{N}Zl + \psi_W\left(k'+\Delta k' - \omega_h^{k'}\frac{N}{2\pi}\right) - \psi_W\left(k' - \omega_h^{k'}\frac{N}{2\pi}\right). \tag{3.11}$$

With Equation (3.9) and (3.11) the baseband phase can be estimated in every time-frequency point of a voiced speech signal. At unvoiced frames the noisy phase is assigned. Therefore this method relies on an accurate estimation of the fundamental frequency as well as on a voice activity detection. The method showed improvement in instrumental measures of the speech quality.

3.2 Temporal Smoothing of the Unwrapped Phase

The next phase enhancement method is basically centered around decomposing the instantaneous, noisy phase and was introduced in [14]. The method proposes a phase model which decomposes the instantaneous phase into two major components: the linear phase and the unwrapped phase. After the removal of the linear phase part, a temporal smoothing filter is applied onto the unwrapped phase. Therefore, this method is referred to as temporal smoothing of the unwrapped phase (TSUP). Figure 3.1 illustrates the processing steps of the TSUP estimator.

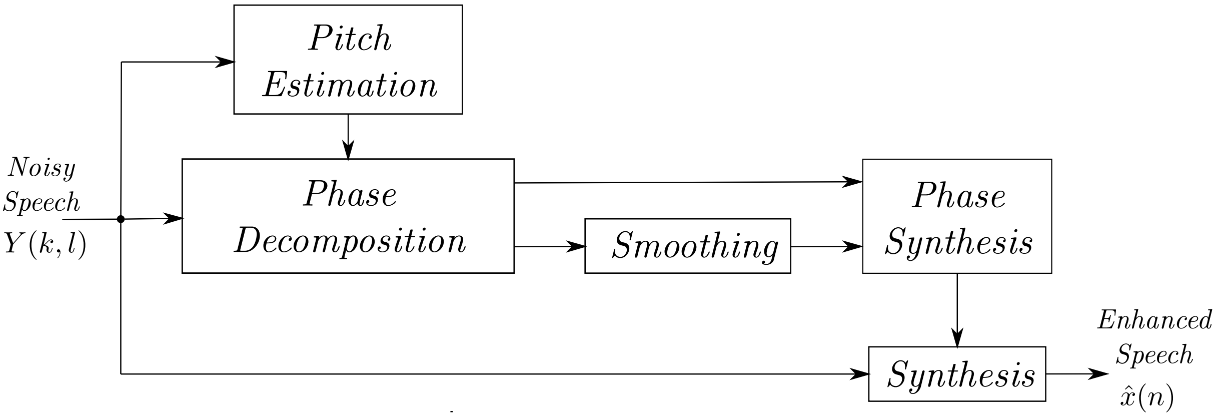

Figure 3.1: Block diagram of the TSUP estimator. $\psi_{lin}(h,l)$ denotes the linear phase, $\Psi(h,l)$ and $\hat{\Psi}(h,l)$ the unwrapped phase before and after the enhancement step, respectively. The enhanced instantaneous phase, which is used to synthesize the enhanced speech $\hat{x}(n)$, is denoted by $\hat{\psi}(h,l)$.

For TSUP phase estimation the speech signal $x(n)$ is sliced into windowed frames according to:

$$x_w(n',l) = x(n' + t(l))w(n'), \tag{3.12}$$

where $t(l)$ and $w(n)$ denote the time instance at frame l and the analysis window, respectively and n' denotes the STFT time index, which is defined by length of the analysis window N_l, $n' \in [-\frac{(N_l-1)}{2}, \frac{(N_l-1)}{2}]$. In [32], it was shown that a pitch-synchronous signal segmentation is advantageous for the processing of the phase. Therefore, the time instances of the frames $t(l)$ are evaluated according to the fundamental frequency $f_0(l)$:

$$t(l) = t(l-1) + \frac{1}{4 \cdot f_0(l-1)}. \tag{3.13}$$

Every frame $x_w(n',l)$ of a voiced signal region can be presented as the sum of harmonics consisting of amplitude $a(h,l)$ and phase $\psi(h,l)$:

$$x_w(n',l) \approx \sum_{h=1}^{H_l} a(h,l)\cos(h\omega_0(l)n' + \psi(h,l))w(n'), \tag{3.14}$$

with harmonic index $h \in [1, H_l]$ and $\omega_0(l) = 2\pi f_0(l)/f_s$. In this framework, a blackman window was used, as it was documented to be the best performing window type for phase estimation [33].

3.2.1 Phase Decomposition

The instantaneous phase $\psi(h,l)$ in Equation (3.14) can be represented by decomposing it into its basic components using a model introduced in [32]:

$$\psi(h,l) = \underbrace{h\sum_{l'=0}^{l}\omega_0(l')(t(l') - t(l'-1))}_{\text{linear phase } \psi_{lin}(h,l)} + \underbrace{\overbrace{\angle V(h,l)}^{\text{minimum phase}} + \overbrace{\psi_d(h,l)}^{\text{dispersion phase}}}_{\text{unwrapped phase } \Psi(h,l)}. \tag{3.15}$$

The first part, which denotes the linear phase $\psi_{lin}(h,l)$, captures the influence of the harmonic frequency $h\omega_0$ and the time instant. The linear phase part can be considered as a factor that maps the phase to its frequency and time instance, which explains the cyclic wrapping of the instantaneous phase. The second term of Equation (3.15), which is called minimum phase term, denotes the phase response of the vocal tract $V(l)$ sampled at harmonic h. The minimum phase part can be estimated from the magnitude of the signal using Hilbert transform [34].

The last term in Equation (3.15) is referred to as dispersion phase $\psi_d(h,l)$. It contains the stochastic characteristics of the phase, which are not captured by the linear phase and the minimum phase. In speech coding it is also referred to as source shape [35], as it characterizes the pulse shape. The combination of minimum phase and dispersion phase is referred to as unwrapped phase $\Psi(h,l)$. The unwrapped phase term is what is left after removing the linear and also deterministic term of the phase. $\Psi(h,l)$ can be considered as a non-deterministic random variable. The TSUP estimator only decomposes the instantaneous phase into unwrapped and linear phase and tries to eliminate the stochastic changes introduced by additive noise by temporal smoothing.

3.2.2 Temporal Smoothing of the Unwrapped Phase

To obtain the unwrapped phase, the linear phase has to be removed as follows:

$$\Psi(h,l) = \psi(h,l) - \psi_{lin}(h,l). \tag{3.16}$$

Since the linear phase heavily relies on the fundamental frequency, the quality of the estimation of the unwrapped phase $\Psi(h,l)$ as well depends heavily on the quality of the fundamental frequency estimate. The TSUP estimator is based on the observation that at voiced speech segments, the unwrapped phase evolves smoothly across time. Noise in the signal results in higher fluctuations of the unwrapped phase, therefore $\Psi(h,l)$ is smoothed across time by taking the short term circular mean value:

$$\hat{\Psi}(h,l) = \angle \sum_{l'=l-\frac{W}{2}}^{l+\frac{W}{2}} \mathrm{e}^{j\Psi(h,l')}, \tag{3.17}$$

where W denotes the number of frames that lie within a time span of 20 ms and $\hat{\Psi}(h,l)$ denotes the enhanced unwrapped phase, respectively. The time span of the temporal smoothing is chosen according to the quasi-stationarity character of speech [18].

3.2.3 Signal Reconstruction

The so-obtained enhanced unwrapped phase is then re-combined with the linear phase to result in the enhanced instantaneous phase at harmonics:

$$\hat{\psi}(h,l) = \hat{\Psi}(h,l) + \psi_{lin}(h,l). \tag{3.18}$$

Before synthesis, $\hat{\psi}(h,l)$ is transformed to the STFT domain, which is done by modifying all frequency bins within the main-lobe of the analysis window around harmonic h.

$$\hat{\phi}(\lfloor h\omega_0 K \rfloor + i, l) = \hat{\psi}(h,l), \qquad \forall i \in [-N_p/2, N_p/2], \tag{3.19}$$

where N_p denotes the length of the analysis window. The enhanced STFT components are obtained by merging the enhanced phase with the noisy spectral amplitude:

$$\hat{X}(k,l) = |Y(k,l)|\mathrm{e}^{j\hat{\phi}(k,l)}. \tag{3.20}$$

The enhanced time domain signal is obtained by overlapping and adding of the inverse Fourier transform of $\hat{X}(k,l)$.

3.3 Maximum A Posteriori Phase Estimator

Besides other estimation methods like Maximum Likelihood (ML), the Maximum A Posteriori (MAP) estimate can be used for point estimates of an unobserved quantity of empirical data. The MAP estimate takes into account a specific prior distribution over the quantity which has to be estimated. In [13] the authors proposed a MAP estimator to estimate the phase at harmonics given the noisy speech.

3.3.1 Derivation of the MAP Phase Estimator

The phase estimator is derived starting with one sinusoid in noise:

$$\bar{y}(n) = A\cos(h\omega_0 n + \psi) + d(n), \tag{3.21}$$

The sinusoid is characterized by the sinusoidal triple parameters amplitude A, frequency ω_0, phase ψ and additive noise $d(n)$, which is considered to be a zero-mean Gaussian process with variance σ^2. When defining the observation vector $\bar{\mathbf{y}} = \{\bar{y}(n)\}_{n=0}^{N-1}$, the MAP estimate $\hat{\psi}_{\text{MAP}}$ of the harmonic phase ψ is obtained by solving the equation:

$$\hat{\psi}_{\text{MAP}} = \arg\max_{\psi} \frac{p(\bar{\mathbf{y}}|\psi)p(\psi)}{p(\bar{\mathbf{y}})} = \arg\max_{\psi}\; p(\bar{\mathbf{y}}|\psi)p(\psi). \tag{3.22}$$

If white Gaussian noise is assumed for ν, the likelihood $p(\bar{\mathbf{y}}|\psi)$ can be formulated as:

$$p(\bar{\mathbf{y}}|\psi) = (2\pi\sigma^2)^{-\frac{N}{2}}\,\mathrm{e}^{-\frac{1}{\sigma^2}\sum_{n=0}^{N-1}(\bar{y}(n)-A\cos(h\omega_0 n+\psi))^2}. \tag{3.23}$$

When it comes to the prior probability of the phase $p(\psi)$, other estimation methods tend to use a uniform phase distribution, resulting in the noisy phase as the optimal phase-estimate. The novelty of this MAP estimator lies in incorporating the Von Mises distribution. The Von Mises distribution is the maximum entropy distribution for a given circular mean value μ_c and a concentration parameter κ. With the assumption that ψ follows a Von Mises distribution, $p(\psi)$ can be written as:

$$\psi \sim \mathcal{VM}(\mu_c, \text{¸}\kappa) \quad ; \quad p(\psi) = \frac{\mathrm{e}^{\kappa\cos(\psi-\mu_c)}}{2\pi I_0(\kappa)}, \tag{3.24}$$

where I_0 denotes as the modified Bessel function of the first kind for order zero.

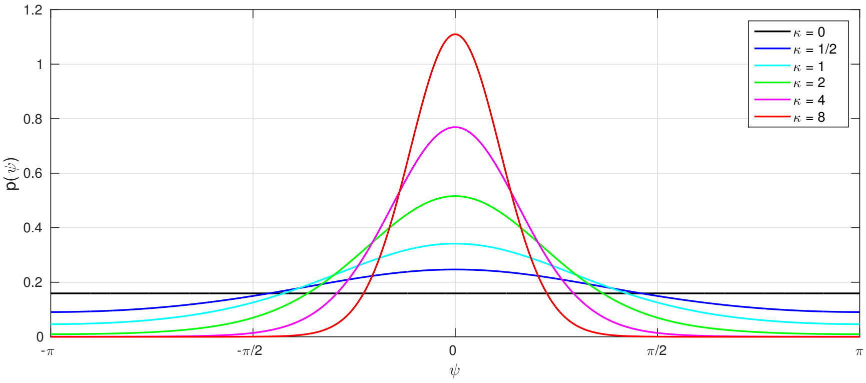

Figure 3.2: Probability density function of a Von Mises distributed variable ψ, with mean value $\mu_c = 0$ and different concentration parameters $\kappa = \left\{0, \frac{1}{2}, 1, 2, 4, 8\right\}$.

Figure 3.2 depicts the probability density function (PDF) of the Von Mises distribution for different concentration parameters κ. In the first extreme case of $\kappa = 0$, the PDF results in a uniform distribution, in the second extreme case of $\kappa = \infty$ the PDF of the Von Mises distribution turns into a delta Dirac. With the assumptions from Equation (3.23) and (3.24) and after discarding all the constants, Equation (3.22) can be rewritten as:

$$\hat{\psi}_{\mathrm{MAP}} = \arg\max_{\psi} \quad -\frac{1}{\sigma^2}\sum_{n=0}^{N-1}\left(\bar{y}(n) - A\cos(h\omega_0 n + \psi)\right)^2 + \kappa\cos(\psi - \mu_c). \tag{3.25}$$

The MAP solution for the phase ψ of a single sinusoid in noise is obtained by setting the derivative of the right-hand term of Equation (3.25) to zero:

$$\hat{\psi}_{\mathrm{MAP}} = \tan^{-1}\left(\frac{-\frac{2A}{\sigma^2}\sum_{n=0}^{N-1}\bar{y}(n)\sin(h\omega_0 n) + \kappa\sin(\mu_c)}{\frac{2A}{\sigma^2}\sum_{n=0}^{N-1}\bar{y}(n)\cos(h\omega_0 n) + \kappa\cos(\mu_c)}\right), \tag{3.26}$$

Equation (3.26) shows, that the MAP estimate is a function of the Von Mises parameter μ_c and κ, the length of the input signal N and the local signal-to-noise ratio (SNR), which is denoted by $\frac{2A}{\sigma^2}$. At high SNRs, which means that $A \gg \sigma^2$, the estimate rather relies on the noisy phase than on the mean value μ_c. This is motivated by [36], where it was shown that the ML estimate of the clean phase is the noisy DFT phase sampled at harmonics. Therefore, the noisy phase is considered as a good estimate for high SNR scenarios. In the case of a low harmonic SNR, the estimator relies only on the mean value μ_c.

3.3.2 Extension to Speech Signal

In Section 3.3.1 the MAP estimate was derived for a single sinusoid in noise. To extend this derivation to a speech signal, the segmented noisy signal $y(n,l)$ is considered as a harmonic signal given as:

$$y(n,l) = \sum_{h=1}^{H_l}\sum_{n=0}^{N-1} A(h,l)\cos(h\omega_0(l)n + \psi(h,l)) + \nu(n,l). \tag{3.27}$$

Under the assumption of non-interacting sinusoidal components, the MAP estimate from Equation (3.26) is then applied at each STFT frame l and each harmonic h separately. Therefore, we get:

$$\hat{\psi}_{\mathrm{MAP}}(h,l) = \tan^{-1}\left(\frac{-\frac{2A(h,l)}{\sigma^2(h,l)}\sum_{n=0}^{N-1}\bar{y}(n,l)\sin(h\omega_0 n) + \kappa(h,l)\sin(\mu_c(h,l))}{\frac{2A(h,l)}{\sigma^2(h,l)}\sum_{n=0}^{N-1}\bar{y}(n,l)\cos(h\omega_0 n) + \kappa(h,l)\cos(\mu_c(h,l))}\right). \tag{3.28}$$

3.3.3 Von Mises Distribution Parameter Estimation

Given the DFT of the windowed noisy input signal, which is defined as:

$$Y(k,l) = DFT\left\{y(h,l)w(n)\right\}, \tag{3.29}$$

the spectral phase $\phi(k,l)$ is defined as $\phi(k,l) = \angle Y(k,l)$. In order to estimate the parameters used for MAP estimation, the spectral phase $\psi(h,l)$ of each harmonic is estimated by a linear interpolation along frequency of $\phi(k,l)$. Following the phase decomposition from Equation (3.16), the deterministic linear phase is removed to obtain the unwrapped phase $\Psi(h,l)$. Then, a Von Mises distribution is fitted on $\Psi(h,l)$ to characterize its statistical behavior. The Von

Mises parameter are obtained starting with:

$$z(h,l) = \frac{1}{W} \sum_{l'=l-\frac{W}{2}}^{l+\frac{W}{2}} \mathrm{e}^{j\Psi(h,l')}, \tag{3.30}$$

$$\mu(h,l) = \angle z(h,l), \tag{3.31}$$

Where W denotes all the frames that lie within the time span of w_{filt} of frame l. The exact parameter setup used in the experiment section will be given in Section 6.1.3. The circular mean value of the harmonic phase μ_c is then calculated by adding back the linear phase part:

$$\mu_c(h,l) = \mu(h,l) + \psi_{\text{lin}}(h,l). \tag{3.32}$$

The circular variance of $\Psi(h,l)$ is given by:

$$\sigma_c^2(h,l) = 1 - |z(h,l)|. \tag{3.33}$$

To obtain the concentration parameter κ, the following relation, which was proposed in [37], has to be inverted:

$$\sigma_c^2(h,l) = 1 - \frac{I_1(\kappa(h,l))}{I_0(\kappa(h,l))}. \tag{3.34}$$

3.3.4 Signal Reconstruction

The enhanced harmonic phase $\hat{\psi}_{\text{MAP}}(h,l)$ is transformed into STFT domain by utilizing Equation (3.19). The resulting enhanced phase spectrum $\hat{\phi}_{\text{MAP}}(k,l)$ is finally used to obtain the phase-enhanced time-domain signal by applying the inverse DFT on:

$$\hat{X}(k,l) = |Y(k,l)|\mathrm{e}^{j\hat{\phi}_{\text{MAP}}(k,l)}, \tag{3.35}$$

followed by an overlap-and-add routine.

4 Harmonic Phase Representations

In Figures 2.4 and 2.5 it was shown that the instantaneous phase extracted from STFT does not show any useful patterns or details. This Chapter presents some phase-based features which help to get further insight and uncovers the hidden structure of the clean spectral phase. The concept used in these representation is presented as well as closely connected references [30]. All of the representations rely on a harmonic signal model:

$$x(n,l) = \sum_{h=1}^{H_l} A(h,l) \cos\Big(h\omega_0(l)n + \psi(h,l)\Big) \cdot w(n), \tag{4.1}$$

and model the instantaneous harmonic phase $\psi(h,l)$. Other STFT based methods, such as Group Delay or Instantaneous Frequency are not covered in this thesis.

4.1 Relative Phase Shift (RPS)

The Relative Phase Shift (RPS), introduced in [38], proposes a useful representation of the harmonic phase. Given the signal model in (4.1), the RPS representation is based on the difference between the instantaneous phase of an arbitrary harmonic h and the instantaneous phase of the fundamental frequency, which is defined as:

$$RPS(h,l) = \Big(\psi(h,l) - h\psi(1,l)\Big)\Big|_{-\pi}^{+\pi}, \tag{4.2}$$

where $\psi(1,l)$ corresponds to the phase of the fundamental frequency f_0 and $(\alpha)\Big|_{-\pi}^{+\pi}$ denotes the process of wrapping angle α to the interval $[-\pi, +\pi]$.

If Equation (4.2) is rewritten with the phase decomposition principle from Section 3.2.1, it can be observed that the RPS representation is independent of the linear phase:

$$\begin{aligned} RPS(h,l) &= \left(\underbrace{h\sum_{l}\omega_0(l)\Delta t + \Psi(h,l)}_{\psi(h,l)} - h\Big(\underbrace{\sum_{l}\omega_0(l) + \Psi(1,l)}_{\psi(1,l)}\Big)\right)\Bigg|_{-\pi}^{+\pi} \\ &= \Big(\Psi(h,l) - h\Psi(1,l)\Big)\Big|_{-\pi}^{+\pi}. \end{aligned} \tag{4.3}$$

Therefore, the RPS can also be considered as the phase shift of the unwrapped phase between a harmonic h and the first harmonic.

Now it is assumed that the signal waveform is locally stable, which means the unwrapped phase term is also constant. Only the linear phase part is changing according to its dependencies on time shift and frequency. As the RPS circumvents this wrapping problem, the RPS patterns are known to be smooth across time for voiced segments [39].

This property can be observed in Figure 4.1, where the waveform, the instantaneous phase of the second harmonic $\psi(2,l)$ and the RPS of the second harmonic $RPS(2,l)$ are displayed for a voiced segment of a female Speaker[1], taken from GRID corpus.

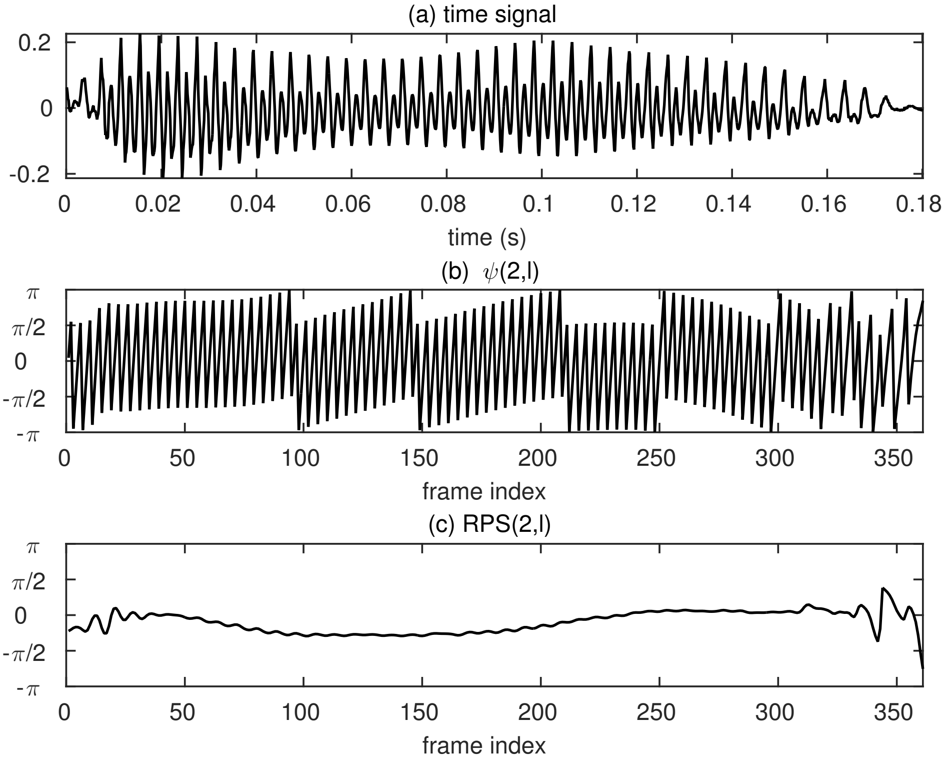

Figure 4.1: The Figure shows (a) time signal, (b) instantaneous phase of the second harmonic $\psi(2,l)$ across frames and (c) RPS of the second harmonic $RPS(2,l)$ across frames for a voiced segment of speech. The utterance was taken from speaker 4 of GRID corpus uttering the sound /u/.

In Figure 4.1 the fluctuations of the instantaneous phase can be observed, the majority of this

[1] Speaker 4

is attributed to the linear phase term. The RPS patterns, however, evolve slowly across time for the whole speech segment. This encourages for further modeling and manipulation.

In the following, a whole utterance of a female speaker taken from GRID corpus is analyzed in terms of its RPS. The speaker performs the sentence 'bin blue at l 4 soon'. The time signal of the utterance is displayed in Figure 4.2. For more information about the speech corpus used, the reader is referred to Section 6.1.1.

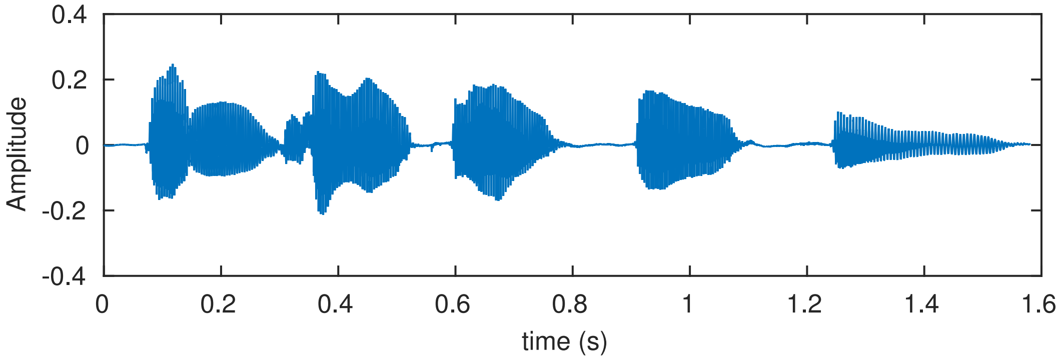

Figure 4.2: Waveform of the signal 'bin blue at l 4 soon' spoken by speaker 4 of GRID corpus.

To analyse RPS over time, the phasegram is presented in the following. The phasegram is the counterpart to the spectrogram. It illustrates the evolution over time of the phase information. Figure 4.3 shows the RPS in terms of a phasegram evaluated from the utterance of GRID corpus, displayed in Figure 4.2, for different noise and SNR scenarios.

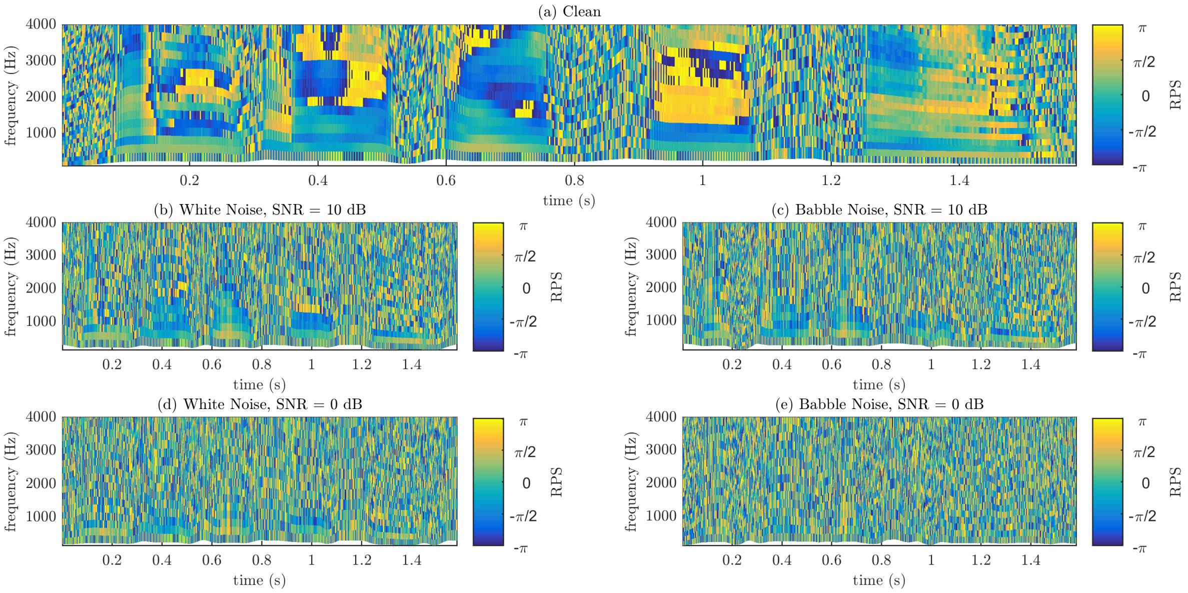

Figure 4.3: RPS phasegram of an utterance of Speaker 4 form GRID corpus for different scenarios: (a) clean, (b) 10 dB white noise, (c) 10 dB babble noise, (d) 0 dB white noise and (e) 0 dB babble noise.

If the RPS for the case of h = 1 is calculated following Equation (4.2), the output would be zero for all frames. Therefore the instantaneous phase of the fundamental frequency is inserted, following

$$RPS(1,l) = \psi(1,l), \tag{4.4}$$

as it is of high interest for re-synthesis of the other harmonic phase components, in Equation (4.5). To re-synthesize the harmonic phase given the RPS patterns,

$$\psi(h,l) = \Big(RPS(h,l) + h\psi(1,l)\Big)\Big|_{-\pi}^{+\pi} = \Big(RPS(h,l) + hRPS(1,l)\Big)\Big|_{-\pi}^{+\pi}. \tag{4.5}$$

can be applied. To display the RPS in terms of a phasegram, the RPS values in dependency of time and frequency $RPS(f,t)$ have to be evaluated. Since the framework for RPS processing calls for a harmonic signal model, the harmonic RPS values are mapped according to the corresponding frequency bin k using:

$$RPS(k,l) = RPS(\bar{h},l) \qquad , \qquad \bar{h} = \arg\min_{h} \left\{ \left| h \cdot \frac{2\pi f_0(l)}{fs} - \frac{2\pi k}{N} \right| \right\} \tag{4.6}$$

and according to the corresponding time instance $t(l)$ using:

$$RPS(k,t(l)) = RPS(k,l) \qquad , \qquad t(l) = l\Delta t \tag{4.7}$$

The phasegram in Figure 4.3 shows the progression of $RPS(k,t)$. The phasegram for the clean signal shows clearly stable (or rather slow varying) RPS patterns for certain segments of the signal. When comparing to the waveform of the clean signal in Figure 4.2, it can be observed that these signal segments refer to the voiced parts of the speech signal. This corresponds with the previously mentioned smoothness of the representation. When adding noise to the signal, it can be seen that these patterns get lost. While some structure can still be detected at low harmonics for an SNR of 10dB, this residual structure is destroyed almost throughout the whole signal for the low SNR scenario.

In [40], an alternative version of the RPS has been presented. It is defined as:

$$\tilde{\psi}(h,l) = \psi(h,l) - \angle V(h,l), \tag{4.8}$$

where $\tilde{\psi}(h,l)$ denotes the instantaneous phase after removing the minimum-phase term $\angle V(k,l)$, which corresponds to removing influence of the amplitude envelope. The alternative version $\widetilde{RPS}(h,l)$ is then defined as:

$$\widetilde{RPS}(h,l) = \Big(\tilde{\psi}(h,l) - h\tilde{\psi}(1,l)\Big)\Big|_{-\pi}^{+\pi} = \Big(\psi_d(h,l) - h\psi_d(1,l)\Big)\Big|_{-\pi}^{+\pi}. \tag{4.9}$$

To see the impact of the removal of the minimum phase term, the reader is referred to [40].

4.2 Phase Distortion (PD)

The concept of Phase Distortion (PD), also introduced in [40], is another way to extract meaningful characteristics from the instantaneous phase. The PD is defined as phase difference between two components. When using a harmonic model, the PD is equal to a finite difference and therefore similar to the group delay. Following [41] the PD can be written in the following form:

$$\begin{aligned} PD(h,l) &= \underset{h}{\Delta}\widetilde{RPS}(h,l) = \Big(\widetilde{RPS}(h+1,l) - \widetilde{RPS}(h,l)\Big)\Big|_{-\pi}^{+\pi} \\ &= \left(\Big(\tilde{\psi}(h+1,l) - (h+1)\tilde{\psi}(1,l)\Big) - \Big(\tilde{\psi}(h,l) - h\tilde{\psi}(1,l)\Big)\right)\Bigg|_{-\pi}^{+\pi} \\ &= \Big(\tilde{\psi}(h+1,l) - \tilde{\psi}(h,l) - \tilde{\psi}(1,l)\Big)\Big|_{-\pi}^{+\pi}. \end{aligned} \tag{4.10}$$

To avoid an approximation of the vocal tract filter response and the associated cepstrum estimation, a modified version of PD is defined, where the vocal tract filter response is not removed:

$$\begin{aligned} \widetilde{PD}(h,l) &= \Big(\psi(h+1,l) - \psi(h,l) - \psi(1,l)\Big)\Big|_{-\pi}^{+\pi} \\ &= \left(\Big(\Psi(h+1,l) + (h+1)\omega_0 n\Big) - \Big(\Psi(h,l) + h\omega_0 n\Big) - \Big(\Psi(1,l) + \omega_0 n\Big)\right)\Bigg|_{-\pi}^{+\pi} \\ &= \Big(\Psi(h+1,l) - \Psi(h,l) - \Psi(1,l)\Big)\Big|_{-\pi}^{+\pi} \\ &= \Big(RPS(h+1,l) - RPS(h,l)\Big)\Big|_{-\pi}^{+\pi} \\ &= \underset{h}{\Delta} RPS(h,l). \end{aligned} \tag{4.11}$$

It can be seen that the phase distortion is equivalent to a RPS difference across harmonics. As a consequence, the synthesis equation for the instantaneous phase is given as:

$$\begin{aligned} \psi(h,l) &= h \cdot \psi(1,l) + \sum_{\bar{h}=1}^{h-1} \widetilde{PD}(\bar{h},l) \\ &= \Big(h \cdot \tilde{\psi}(1,l) + \sum_{\bar{h}=1}^{h-1} PD(\bar{h},l)\Big) + \angle V(h,l) \end{aligned} \tag{4.12}$$

Figures 4.4 displays the behavior of the modified PD for different SNR scenarios by means of a phasegram, calculated using the Equations (4.6), (4.7) and (4.9).

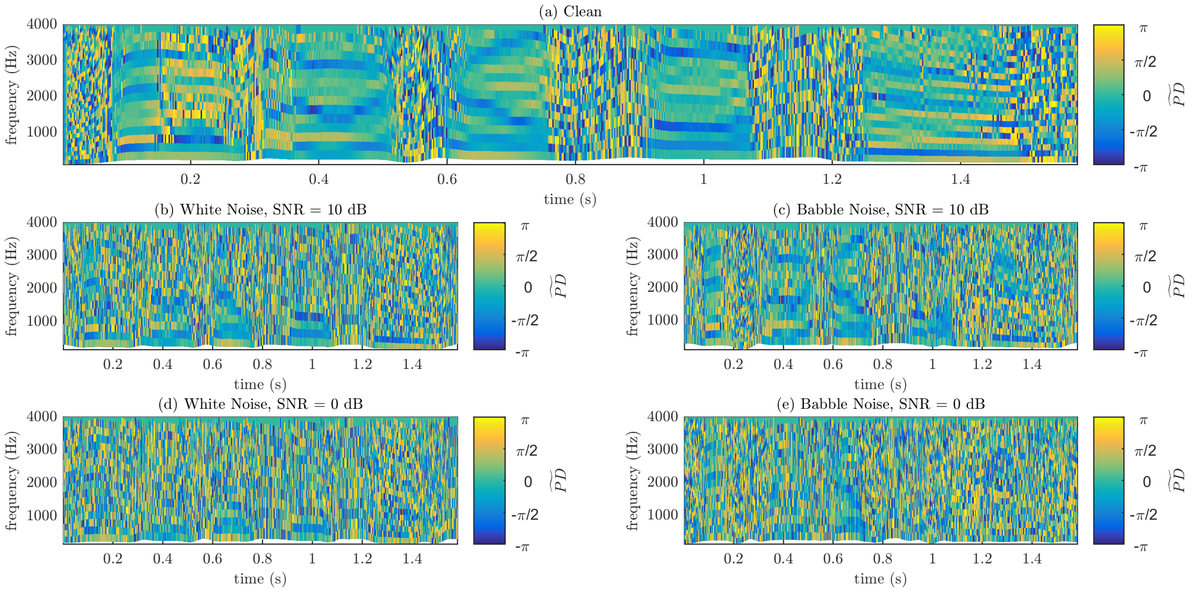

Figure 4.4: PD phasegram of an utterance of Speaker 4 form GRID corpus for different scenarios: (a) clean, (b) 10 dB white noise, (c) 10 dB babble noise, (d) 0 dB white noise and (e) 0 dB babble noise.

4.3 Phase Quasi-Invariant (PQI)

Another phase representation that can be implemented in a harmonic signal model, is called Phase Quasi-Invariant (PQI). The method, which was originally formulated for radar and sonar systems, was first proposed in [42] for the analysis of phase relations in speech. It has been successfully applied for rotary machines condition monitoring [43]. Similar to the RPS, the PQI representation is based on the phase difference measures between harmonics.

The definition of PQI for two arbitrary harmonics $\bar{h}$ and h, with $\bar{h}, h \in$ [1,H], is based on the following relation:

$$\Delta\Psi(\bar{h}, h, l) = \left(\psi(\bar{h}, l) - \frac{\psi(h, l) \cdot \bar{h}}{h}\right). \tag{4.13}$$

Following the phase decomposition from Section 3.2.1, this can be rewritten as:

$$\begin{aligned}\Delta\Psi(\bar{h}, h, l) &= \left(\left(\Psi(\bar{h}, l) + \bar{h}\omega_0 n\right) - \frac{\bar{h} \cdot \left(\Psi(h, l) + h\omega_0 n\right)}{h}\right) \\ &= \left(\Psi(\bar{h}, l) - \frac{\Psi(h, l) \cdot \bar{h}}{h}\right).\end{aligned} \tag{4.14}$$

Equation (4.14) shows that the linear phase part and therefore the dependency on fundamental frequency is discarded. Therefore, a behaviour similar to Figure 4.1 can be assumed.

In the next step, the phase difference is wrapped according to its unambiguous definition range, which is given as $\left[\frac{-\pi\cdot\bar{h}}{h};\frac{+\pi\cdot\bar{h}}{h}\right]$. Therefore, the final relation of PQI is given as:

$$\Delta\Psi(\bar{h},h,l)=\Delta\Psi_{\bar{h}}(h,l)=\frac{h}{\bar{h}}\left(\Psi(\bar{h},l)-\frac{\Psi(h,l)\cdot\bar{h}}{h}\right)\Bigg|_{\frac{-\pi\cdot\bar{h}}{h}}^{\frac{+\pi\cdot\bar{h}}{h}}. \tag{4.15}$$

Note that the PQI can be evaluated for arbitrary parameters $\bar{h}$ and h. For simplicity reasons $\bar{h}$ is further related to as the *reference harmonic*.

In Figure 4.5 the PQI for $\bar{h}=1$ is displayed for the same scenarios as in Sections 4.1 and 4.2.

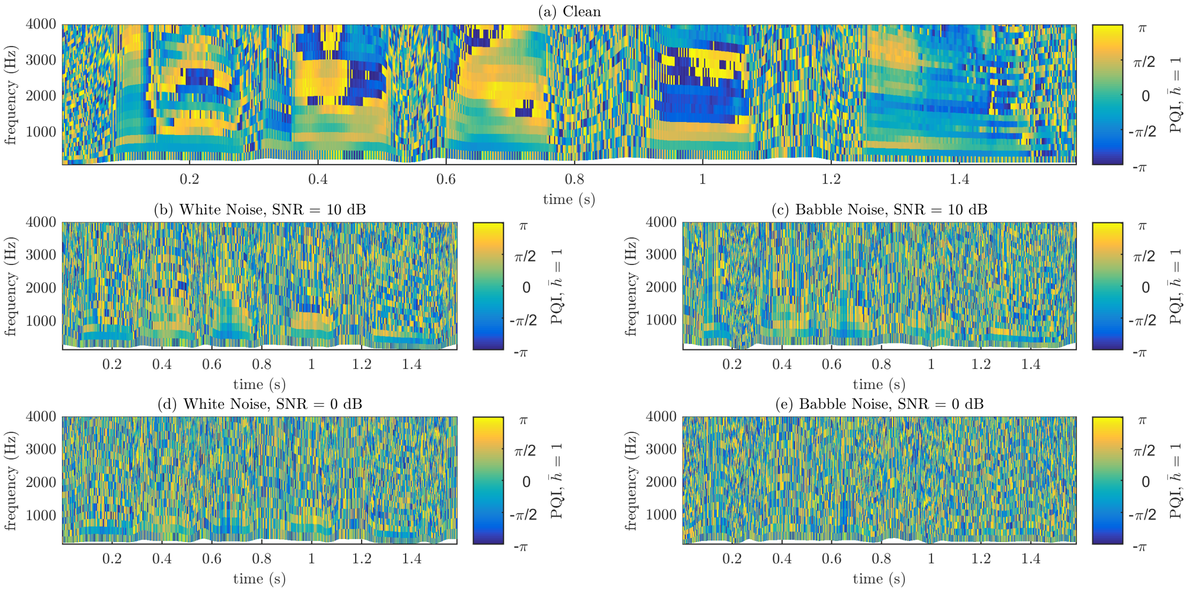

Figure 4.5: PQI phasegram for $\bar{h}=1$ of an utterance of Speaker 4 form GRID corpus for different scenarios: (a) clean, (b) 10 dB white noise, (c) 10 dB babble noise, (d) 0 dB white noise and (e) 0 dB babble noise.

Again, the calculation of PQI following Equation (4.15) for $h=\bar{h}$ results in zero, therefore $\Delta\Psi_{\bar{h}}(h=\bar{h},l)$ is replaced with the reference phase $\psi(\bar{h},l)$, as it is of high interest for signal reconstruction.

4.3.1 PQI Representation for Higher Reference Harmonics

In the following, the reference harmonic $\bar{h}$ is increased. Figures 4.6 and 4.7 show the PQI phasegram of the same speaker and noise scenarios as earlier, but for $\bar{h}=[2,3]$.

By observing Figure 4.6 (a) and Figure 4.7 (a), slow varying PQI patterns can be determined in voiced segments only for the harmonics which are integer multiples of $\bar{h}$. Again the additive noise tends to deteriorate the existing structure of the clean patterns. This deterioration gets stronger with decreasing SNR. Harmonics at non integer multiples of $\bar{h}$ do not show any structure on first sight, independent of the voicing state of the segment or SNR scenario. The properties mentioned previously also continue for higher choices of $\bar{h}$.

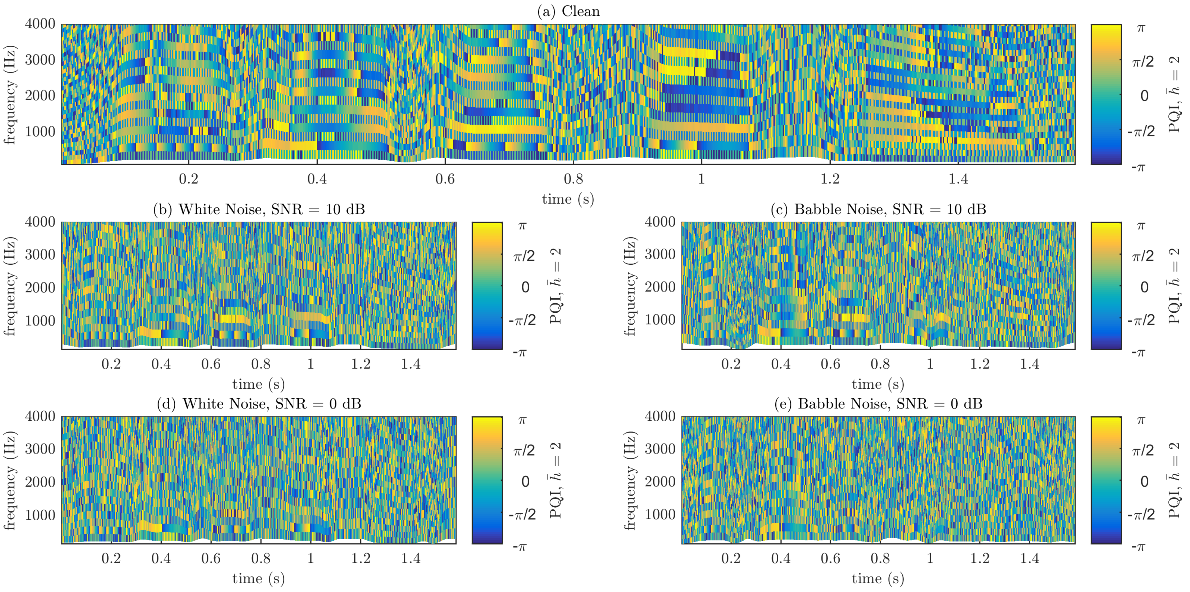

Figure 4.6: PQI phasegram for $\bar{h} = 2$ of an utterance of Speaker 4 form GRID corpus for different scenarios: (a) clean, (b) 10 dB white noise, (c) 10 dB babble noise, (d) 0 dB white noise and (e) 0 dB babble noise.

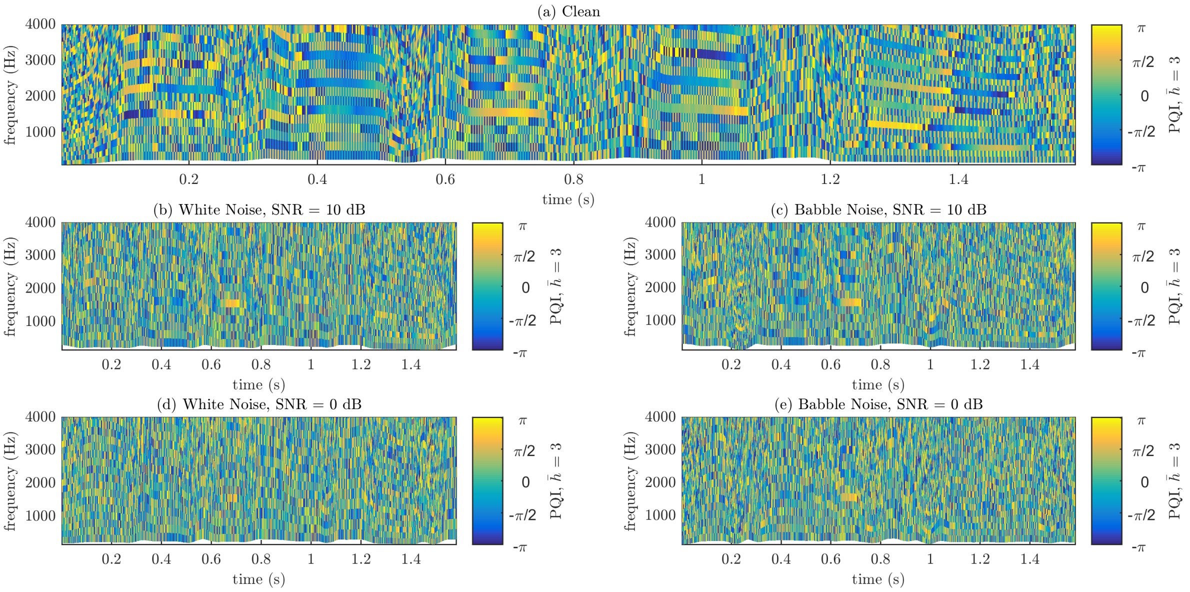

Figure 4.7: PQI phasegram for $\bar{h} = 3$ of an utterance of Speaker 4 form GRID corpus for different scenarios: (a) clean, (b) 10 dB white noise, (c) 10 dB babble noise, (d) 0 dB white noise and (e) 0 dB babble noise.

To access the PQI structure of these harmonics an additional operation is needed. In the following, the modified PQI is presented, which makes it possible to access the hidden structure for harmonics at non integer multiples of the reference harmonic:

$$\Delta\Psi^{mod}_{\bar{h}}(h,l) = \left(\Delta\Psi_{\bar{h}}(h,l)\Big|^{\frac{+\pi}{\bar{h}}}_{\frac{-\pi}{\bar{h}}}\right)\bar{h} \qquad , \qquad \forall h \notin [\bar{h}, 2\bar{h}, 3\bar{h}, \dots] \tag{4.16}$$

In Figures 4.8 and 4.9 the phasegram of the modified PQI $\Delta\Psi^{mod}_{\bar{h}}(h,l)$ for the reference harmonics $\bar{h} = [2, 3]$ can be observed for the same speaker and SNR scenarios as in the Figures 4.3 - 4.7.

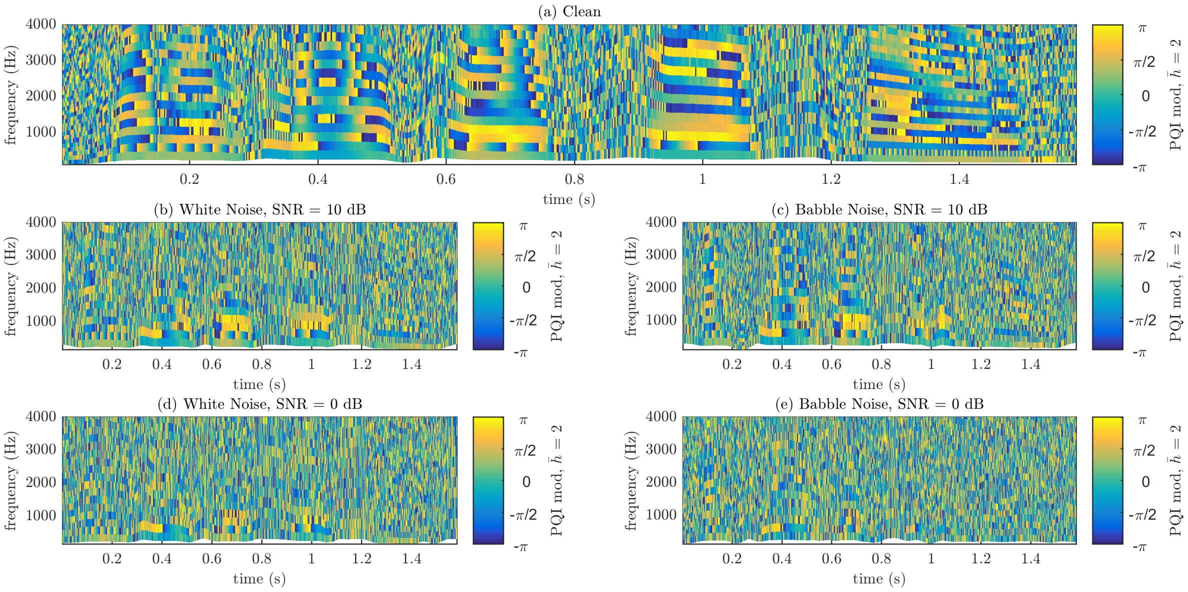

Figure 4.8: modified PQI phasegram for $\bar{h} = 3$ of an utterance of Speaker 4 form GRID corpus for different scenarios: (a) clean, (b) 10 dB white noise, (c) 10 dB babble noise, (d) 0 dB white noise and (e) 0 dB babble noise.

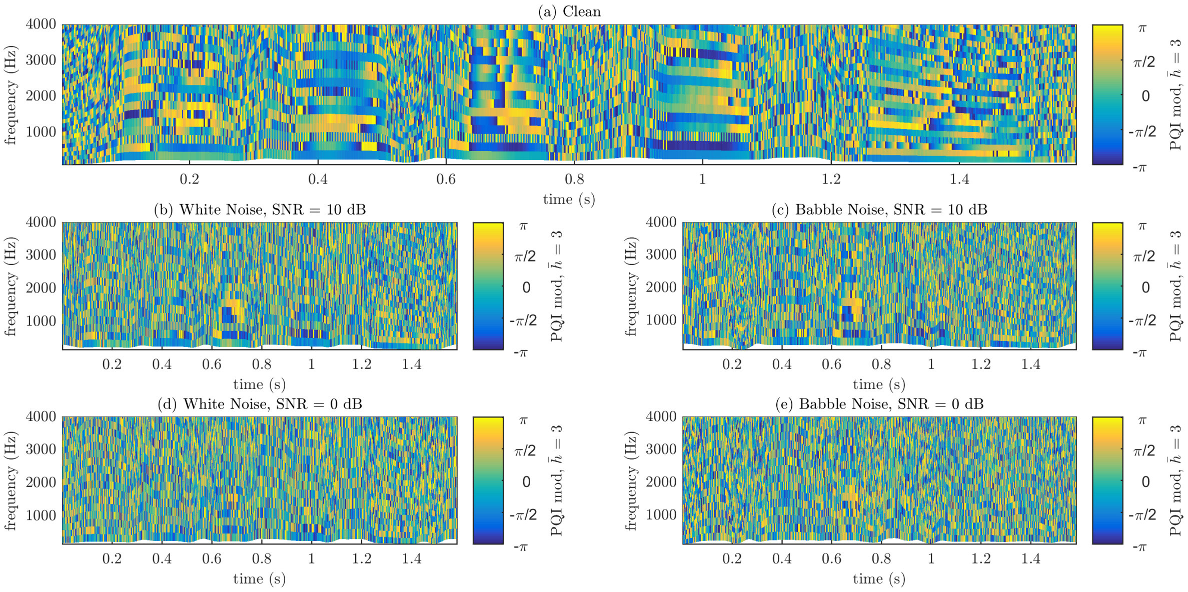

Figure 4.9: modified PQI phasegram for $\bar{h} = 3$ of an utterance of Speaker 4 form GRID corpus for different scenarios: (a) clean, (b) 10 dB white noise, (c) 10 dB babble noise, (d) 0 dB white noise and (e) 0 dB babble noise.

It can be observed that the additional wrapping in Equation (4.16) reveals similar smooth patterns for non integer harmonics of the reference phase, the final multiplication with $\bar{h}$ ensures a range of $[-\pi, +\pi]$.

4.3.2 Re-synthesis of the Instantaneous Phase

In Equation (4.15) the original PQI is wrapped and bound to its unambiguous definition range. With the step of additional wrapping in Equation (4.16), information of the phase signal is lost, which makes an accurate re-synthesis of the instantaneous phase from the modified PQI impossible. Therefore it is necessary to save the residual signal for the reconstruction step:

$$\Theta(h,l) = \Delta\Psi_{\bar{h}}^{mod}(h,l) - \Delta\Psi_{\bar{h}}(h,l) \tag{4.17}$$

Rewriting Equation (4.14) gives the re-synthesis equations for the instantaneous phase or unwrapped phase at harmonics:

$$\psi(h,l) = \frac{h \cdot \psi(\bar{h},l)}{\bar{h}} - \Delta\Psi_{\bar{h}}(h,l) = \frac{h \cdot \psi(\bar{h},l)}{\bar{h}} - \left(\frac{\Delta\Psi_{\bar{h}}^{mod}(h,l)}{\bar{h}} - \Theta(h,l)\right) \tag{4.18}$$

$$\Psi(h,l) = \frac{h \cdot \Psi(\bar{h},l)}{\bar{h}} - \Delta\Psi_{\bar{h}}(h,l) = \frac{h \cdot \Psi(\bar{h},l)}{\bar{h}} - \left(\frac{\Delta\Psi_{\bar{h}}^{mod}(h,l)}{\bar{h}} - \Theta(h,l)\right), \tag{4.19}$$

respectively.

4.4 Comparison of Phase Representation Methods

By comparing the phasegrams in Section 4.1 to the ones in Section 4.3, similarities in the structure of the patterns can be perceived. The relation between RPS and PQI can be formulated as:

$$\begin{aligned} RPS(h,l) &= \Big(\psi(h,l) - h \cdot \psi(1,l)\Big)\Big|_{-\pi}^{+\pi} \\ &= \Big(\Psi(h,l) - h \cdot \Psi(1,l)\Big)\Big|_{-\pi}^{+\pi} \\ &= \frac{-h}{1} \cdot \left(\Psi(1,l) - \frac{\Psi(h,l)}{h}\right)\Bigg|_{\frac{-\pi\cdot 1}{h}}^{\frac{+\pi\cdot 1}{h}} . \\ &= -\Delta\Psi_1(h,l) \end{aligned} \tag{4.20}$$

Equation (4.20) shows that $RPS(h,l)$ can be represented by the negative PQI with reference harmonic $\bar{h} = 1$. The major benefit of the PQI representation is that it is not limited to the fundamental frequency phase, as the reference harmonic $\bar{h}$ is free to choose. Therefore the RPS could be considered as a special case of the PQI. The RPS was successfully used in the field of speech synthesis [44] and speaker recognition [45], hence it can be assumed that implementation of the PQI could lead to similar results. In Chapter 5, a phase enhancement method relying on PQI constraints is presented.

When comparing the RPS and PQI phasegrams, the smoothness of both phase representations for voiced segments is visible, especially for clean signals. When noise is added, less smooth patterns arise even at high SNRs. This can be attributed to a variety of reasons. For the first part, an erroneous estimation of the fundamental frequency leads to errors in the unwrapping process and the harmonic framework. These miscalculations therefore lead to bigger errors at high harmonics, as the error in fundamental frequency is multiplied with the harmonic index. It is also known that higher order harmonics often show lower amplitudes, which means that they are masked by noise more easily. When harmonics are masked by noise, they do not appear in the spectrogram. We can observe the same loss of harmonic structure in the phasegrams.

From Equation (4.11) we can conclude the relation between PD, RPS and PQI:

$$\begin{aligned}\widetilde{PD}(h,l) &= \underset{h}{\Delta} RPS(h,l) \\ &= -\underset{h}{\Delta}(\Delta\Psi_1(h,l))\end{aligned} \tag{4.21}$$

The PD can be considered as the difference between the RPS or the PQI with reference harmonic $\bar{h} = 1$ across harmonics. In Figure 4.3 and 4.4 it can be observed that this difference leads to a further smoothing across time for voiced phonemes, resulting in a lower variance.

5 Speech Enhancement using Phase Representations

This Chapter shows how the phase representations derived in Chapter 4 can be used for the purpose of spectral phase enhancement of a noisy speech signal.

5.1 PQI Enhancement with Fixed Reference Harmonic

This Section presents the phase estimation method proposed in [46]. The paper, which was written in the course of this work, is included in the Appendix A. The idea is motivated by the TSUP phase estimator presented in Section 3.2, where it was shown that the temporal smoothing of the non-deterministic part of the instantaneous phase leads to an enhancement in terms of the speech quality and intelligibility. From Equation 4.14 it is known that the PQI eliminates the deterministic part of the phase. When the speech signal is distorted with noise, the PQI patterns therefore capture the noise contribution on the spectral phase.

The PQI phase enhancement method addresses this problem by applying smoothing filters in PQI domain. This smoothing can be performed via time, frequency or a combination of both. The smoothing process aims to restore the initial phase relations between the harmonic components, which defines the shape of the waveform. The enhanced instantaneous harmonic phase is obtained by re-synthesis with the smoothed PQI patterns. A block diagram of this method is depicted in Figure 5.1.

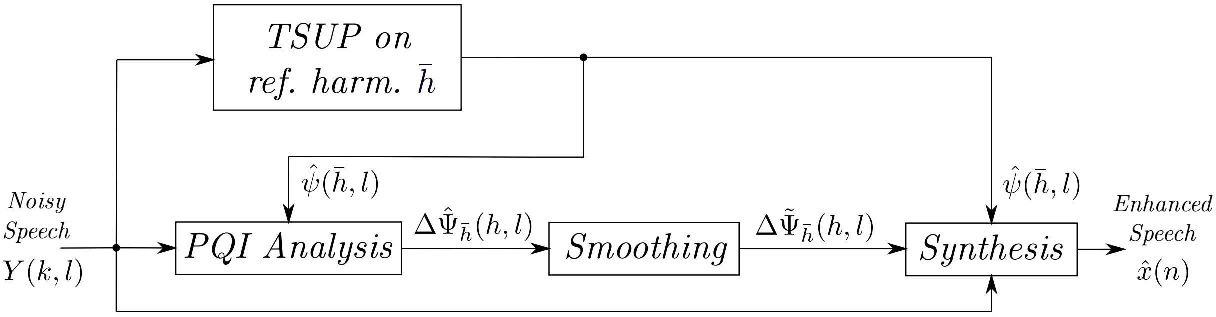

Figure 5.1: Block diagram of the PQI phase enhancement with fixed reference harmonic. $\hat{\psi}(\bar{h},l)$ denotes the enhanced reference phase obtained by TSUP [14], $\hat{\Psi}_{\bar{h}}(h,l)$ the PQI evaluated upon the enhanced reference phase and $\tilde{\Psi}_{\bar{h}}(h,l)$ the smoothed PQI evaluated upon the enhanced reference phase, respectively.

Starting with the problem definition from Equation 2.9, the noisy signal $y(n)$ is represented by an assembly of frames $y(n,l)$. Similar as in Equation (4.1), this signal is modeled based on a harmonic signal model in noise:

$$y(n,l) = \left(\sum_{h=1}^{H_l} a(h,l) \cos \Big(h\omega_0(l)n + \psi(h,l) \Big) + d(n,l) \right) \cdot w(n). \tag{5.1}$$

The time instances of each frame are given by [32]:

$$t(l) = t(l-1) + \frac{1}{4 \cdot f_0(l-1)}. \tag{5.2}$$

5.1.1 Enhancement of the Reference Phase

In the next step the reference harmonic $\bar{h}$ has to be selected. The position and the quality of the reference phase $\psi(\bar{h},l)$ is significantly important for this enhancement method. If the reference phase contains a high amount of noise, the PQI analysis step of the other harmonics is distorted to a certain extent. Therefore, it is recommended to pre-enhance the reference phase before the PQI analysis step. Hence, solely the reference phase is enhanced by removing the linear phase part followed by a temporal smoothing filter from Section 3.2.

Given the instantaneous phase at harmonics $\psi(h,l)$ and the enhanced reference harmonic phase $\hat{\psi}(\bar{h},l)$ the corresponding pre-enhanced PQI values can be calculated following:

$$\Delta\hat{\Psi}_{\bar{h}}(h,l) = \frac{h}{\bar{h}} \left(\hat{\psi}(\bar{h},l) - \frac{\psi(h,l) \cdot \bar{h}}{h} \right) \Bigg|_{\frac{2\pi \cdot \bar{h}}{h}}. \tag{5.3}$$

5.1.2 Smoothing Filter

The results from Equation (5.3) can then be smoothed across time, across harmonics or across both, to obtain an enhanced PQI denoted as $\Delta\tilde{\Psi}_{\bar{h}}(h,l)$.

Smoothing Across Time

The temporal smoothing is performed by mean averaging:

$$\Delta\tilde{\Psi}_{\bar{h}}(h,l) = \angle \frac{1}{|\mathcal{W}|} \sum_{\tilde{l} \in \mathcal{W}} e^{j\Delta\hat{\Psi}_{\bar{h}}(h,\tilde{l})}, \tag{5.4}$$

where $\mathcal{W}$ denotes all frames that lie within a range of t_{filt} around frame l. The parameter setup which was used for the experiments is explained in Section 6.1.3 After the smoothing process, the enhanced phase is used for signal synthesis.

Smoothing Across Frequency

The smoothing across harmonics is performed by:

$$\Delta\tilde{\Psi}_{\bar{h}}(h,l) = \angle\frac{1}{|\mathcal{H}|}\sum_{\tilde{h}\in\mathcal{H}} e^{j\Delta\hat{\Psi}_{\bar{h}}(\tilde{h},l)}, \tag{5.5}$$

where $\mathcal{H}$ denotes the harmonics that lie within a certain range h_{filt} of h.

$$\mathcal{H} = \left[h - \left\lfloor\frac{h_{filt}}{2}\right\rfloor, h + \left\lfloor\frac{h_{filt}}{2}\right\rfloor\right], \qquad \forall\mathcal{H} \in H_l. \tag{5.6}$$

5.1.3 Signal Synthesis

The signal synthesis is based on the framework of [13]. The enhanced harmonic phase is transformed to the STFT domain by modifying the frequency bins within the main-lobe width of the analysis window. Let $Y(k,l)$ denotes the DFT of the noisy signal with k as the corresponding frequency bin and N as the DFT length with $k \in [0, N-1]$. Then $|Y(k,l)|$ denotes the noisy spectral amplitude and $\vartheta(k,l) = \angle Y(k,l)$ denotes the noisy STFT phase. The enhanced STFT phase $\hat{\vartheta}(k,l)$ is then calculated by:

$$\hat{\vartheta}(\lfloor h\omega_0(l)N\rfloor + i, l) = \left(\frac{h\cdot\hat{\psi}(\bar{h},l)}{\bar{h}} - \Delta\tilde{\Psi}_{\bar{h}}(h,l)\right), \qquad \forall i \in [-N_p(l)/2, N_p(l)/2]. \tag{5.7}$$

where $N_p(l)$ denotes the minimum value of either the main-lobe width of the analysis window N_w or the frequencies close to the neighbouring harmonic $N_p(l) = \min(N_w, \omega_0(l)N/(2\pi))$. Further we obtain the phase enhanced signal in the STFT domain by:

$$\hat{X}(k,l) = |Y(k,l)|e^{j\hat{\vartheta}(k,l)}. \tag{5.8}$$

By applying the inverse DFT on $\hat{X}(k,l)$, the enhanced framed time signal $\hat{x}(n,l)$ is obtained. The time signal frames are then overlapped and added to construct the time signal $\hat{x}(n')$, with discrete time index n'.

5.2 PQI Enhancement with Multiple Reference Harmonic

The PQI enhancement method from Section 5.1 is always bound to analyze and synthesize the PQI patterns based upon one single reference harmonic. In this Section, a method, which relies on more than one reference Harmonic at a time, is presented. The method is still based on the same principle of eliminating the deterministic part of the phase. This is performed in PQI domain with the help of smoothing across time or frequency and was also developed in the course of this work.

Figure 5.2 shows an overview of the different parts involved in the phase estimation process.

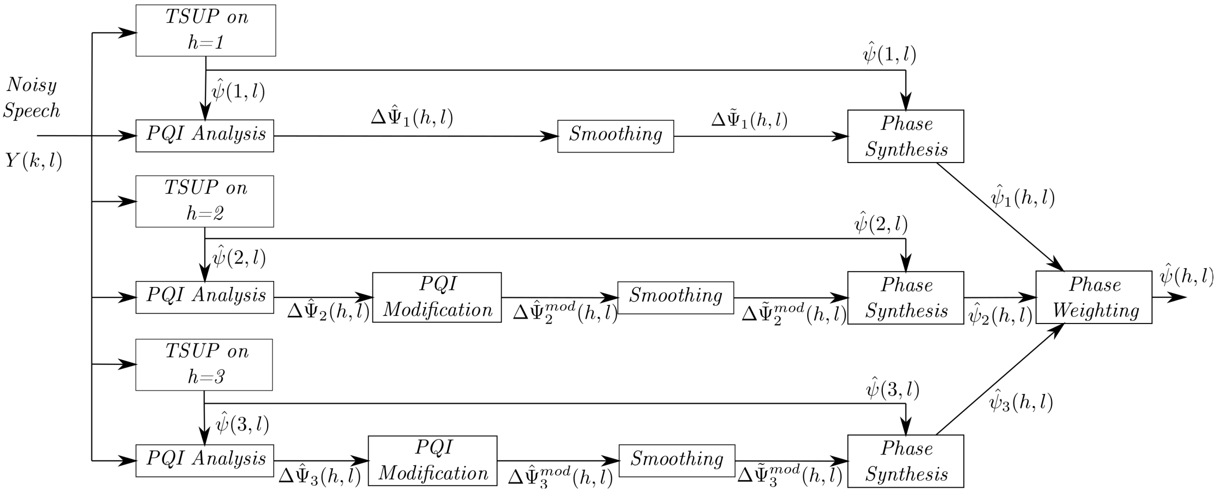

Figure 5.2: Block diagram of the PQI phase enhancement with multiple reference harmonics. $\hat{\psi}([1,2,3],l)$ denote the enhanced reference phases obtained by TSUP [14], $\hat{\Psi}_{[1,2,3]}(h,l)$ the PQI evaluated upon the enhanced reference phases, $\hat{\Psi}^{mod}_{[2,3]}(h,l)$ the modified PQI values, $\tilde{\Psi}_{[1,2,3]}(h,l)$ the smoothed modified PQI evaluated upon reference harmonic 2 and 3, $\tilde{\Psi}_1(h,l)$ the smoothed PQI evaluated upon reference harmonic 1 and $\hat{\psi}_{[1,2,3]}(h,l)$ the enhanced instantaneous phases, respectively.

The enhancement process in Figure 5.2 takes the first three reference harmonics, $\bar{h} = [1,2,3]$, into account. The number of harmonics was chosen due to their relative importance in voiced speech. It has to be noted, that an even higher amount of reference harmonics could be chosen as well.

The phase estimator is based on the same pitch-synchronous harmonic model as the one in Section 5.1. Before PQI evaluation, the three reference harmonics are pre-enhanced by TSUP individually. After the PQI evaluation step, the 3 independent PQI transforms $\psi_1(\hat{h},l)$, $\psi_2(\hat{h},l)$ and $\psi_3(\hat{h},l)$, which were evaluated upon the enhanced reference phase according to Equation (5.3), are obtained.

The PQI patterns are then modified by wrapping the non-integer multiple harmonics of $\bar{h}$ according to Equation (4.16), which makes all harmonics available for modification. This routine does not have to be performed for $\hat{\Psi}_1(h,l)$, as all harmonics are an integer multiple of 1. The smoothing step is again performed either across frequency, time or a combination of both by following Section 5.1.2. The enhanced PQI patterns are re-synthesized with the enhanced reference phase by utilizing Equation (4.18).

5.2.1 Phase Weighting

At this stage, the method consists of 3 independent stages of enhancement methods proposed in Section 5.1. The resulting enhanced instantaneous phases are denoted by $\hat{\psi}_1(h,l)$, $\hat{\psi}_2(h,l)$, $\hat{\psi}_3(h,l)$ and obtained by

$$\hat{\psi}_b(h,l) = \left(\frac{h \cdot \hat{\psi}(b,l)}{b} - \Delta\tilde{\Psi}_b^{mod}(h,l) \right). \tag{5.9}$$

These phase estimates are then merged to receive the final phase estimate $\hat{\psi}(h,l)$. For this step, two weighting methods are presented in the following.

Binary Weighting

It is known that the quality of the reference phase is of high importance for PQI processing. As a sort of quality criterion, the first weighting method takes the local signal-to-noise ratio $\mathrm{SNR}_{\bar{h}}(l)$ into account. $\mathrm{SNR}_{\bar{h}}(l)$ is evaluated for the three references, $\bar{h} = [1,2,3]$, at frame l. It is believed that the phase estimate evaluated upon the reference harmonic with the highest SNR leads to the best overall estimate. Therefore the reference harmonic with the highest local SNR is selected:

$$\tilde{h}(l) = \arg\max_{\bar{h}} \ \{\mathrm{SNR}_{\bar{h}}(l)\}, \quad \bar{h} \in \{1,2,3\}. \tag{5.10}$$

The final phase at frame l is obtained by;

$$\hat{\psi}(h,l) = \hat{\psi}_{\tilde{h}}(h,l). \tag{5.11}$$

Soft Weighting

The second weighting method also takes the local SNR into account. Instead of choosing the best reference harmonic at each frame, the phase estimates are merged according to their weight. The weights are defined as the normalized local SNR:

$$w_{\bar{h}} = \frac{\mathrm{SNR}_{\bar{h}}(l)}{\sum_{h'=1}^{3} \mathrm{SNR}_{h'}(l)}, \quad \bar{h} \in \{1,2,3\}. \tag{5.12}$$

The weights are normalized to keep the in the interval of [0,1]. The final phase $\hat{\psi}(h,l)$ is then given by:

$$\begin{aligned} \hat{\psi}(h,l) &= \arctan\left(\frac{w_1 \cdot \sin\left(\hat{\psi}_1(h,l)\right) + w_2 \cdot \sin\left(\hat{\psi}_2(h,l)\right) + w_3 \cdot \sin\left(\hat{\psi}_3(h,l)\right)}{w_1 \cdot \cos\left(\hat{\psi}_1(h,l)\right) + w_2 \cdot \cos\left(\hat{\psi}_2(h,l)\right) + w_3 \cdot \cos\left(\hat{\psi}_3(h,l)\right)} \right) \\ &= \arctan\left(\frac{\sum_{b=1}^{3} w_b \cdot \sin\left(\hat{\psi}_b(h,l)\right)}{\sum_{b=1}^{3} w_b \cdot \cos\left(\hat{\psi}_b(h,l)\right)} \right) \end{aligned} \tag{5.13}$$

This means the higher the normalized weight of a certain reference harmonic, the higher it influences the resulting phase estimate. It has to be noted that the binary weighting method can be interpreted as selecting the highest weight and setting it to 1, while the remaining weights are set to 0. The soft weighting method weights the estimated phases individually according to their local SNR and calculated the weighted circular mean value over them, resulting in the final phase estimate $\hat{\psi}(h, l)$. The enhanced phase $\hat{\psi}(h, l)$ is then used for signal synthesis, which is performed following Section 5.1.3.

6

Experiments and Results

6.1 Experimental Setup

In this section the setup for the experiments is discussed. A detailed summary of the speech and noise databases, evaluation criteria, phase enhancement methods and the corresponding parameter setup is given in the following.

6.1.1 Speech and Noise Databases

GRID corpus

The speech files used in the experiments were taken from GRID [47], a large audio sentence corpus with 1000 sentences spoken by 34 talkers (18 male, 16 female). The recorded sentences all consist of a command, a color, a preposition, a letter of the alphabet, a digit and an adverb. Resulting in short sentences of less than 3 seconds where the words have no meaningful relation, such as: "bin blue at l 4 soon". Out of the whole corpus, 50 utterances spoken by 20 speakers (10 female and 10 male) were randomly selected. The database is available at a sampling rate of $f_s = 25$ kHz. For this experiment all speech files were down-sampled to $f_s = 8$ kHz to simulate telephony speech.

NOISEX-92 corpus

The noise files used in the experiment section were taken from NOISEX-92 [48]. Two different noise types were chosen, one stationary and one non-stationary. The spectrogram and the long term power spectral density of the stationary white noise is visible in the left column of Figure 6.1. It can be observed that the power spectral density is distributed uniformly across the spectrum. The spectrogram and the long term power spectral density of the non-stationary babble noise can be observed in the right column of Figure 6.1. The babble noise is a record of people talking and should simulate everyday-life scenarios. Therefore a weak harmonic structure can be observed in the spectrum, the power spectral density has its maximum between 100 and 600 Hz.

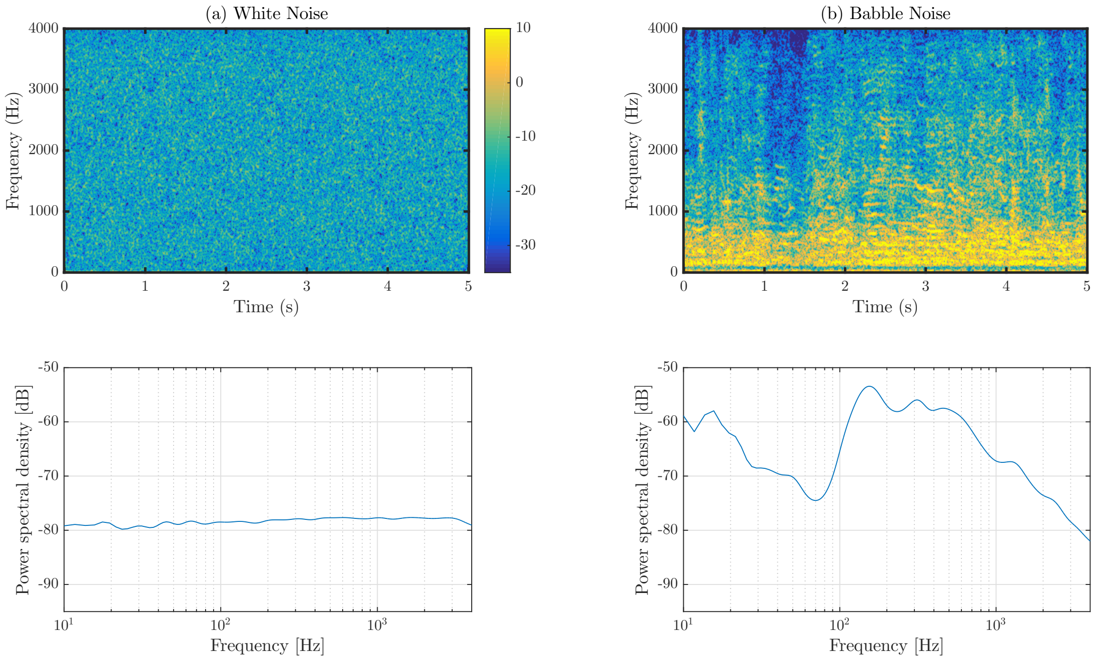

Figure 6.1: (top): Spectrogram (bottom): long term power spectral density of the used noise files: (a) white noise and (b) babble noise

The noise corrupted speech data used in the experiments is generated based on the additive noise model:

$$y(n) = x(n) + \lambda \cdot d(n), \tag{6.1}$$

where λ denotes the noise weighting factor, which is derived from the SNR. Given the energy of the clean speech signal and the noise signal with length N is is defined as:

$$E_x = \sum_{n=0}^{N-1} x(n)^2 \tag{6.2}$$

$$E_d = \sum_{n=0}^{N-1} d(n)^2. \tag{6.3}$$

The noise weighting factor λ is calculated following [17], which results in:

$$\lambda = \sqrt{\frac{E_x}{E_d \cdot 10^{\frac{SNR}{10}}}}. \tag{6.4}$$

Each utterance was corrupted with noise at an SNR of 0 dB, 5 dB and 10 dB.

6.1.2 Evaluation Criteria

In the experiments three objective evaluation criteria were selected:

- Perceptual Evaluation of Speech Quality (PESQ) from [49], as a measure of speech quality.
- Short-Time Objective Intelligibility (STOI) from [50], as a measure of speech intelligibility.
- Unwrapped Root Mean Square Estimation Error (UnRMSE) from [51], as a measure of phase estimation error.

The evaluation measures were calculated at each SNR and for each utterance, then they were averaged over all speaker and presented as a mean score.

6.1.3 Phase Enhancement Methods

In this Section the methods and parameter setups, which were used in the performance evaluation, are discussed. All the presented PQI enhancement methods rely on the signal model and synthesis model discussed in Section 5.1. The window used in the analysis and synthesis stage for the PQI enhancement methods was a Blackman window. The fundamental frequency was obtained by PEFAC [52].

Fixed Reference Phase Methods

Section 6.2.1 displays a comparison of the evaluation performance of PQI based phase enhancement method with fixed reference harmonics, which was discussed in Section 5.1, for different parameter settings.

FS $\bar{h} = [1, 2, 3]$

The PQI enhancement method from Section 5.1, where the smoothing filter was applied across frequency only. The harmonic smoothing parameter h_{filt} was set to 5. This method was evaluated for the first three reference harmonics $\bar{h} = [1, 2, 3]$.

TS $\bar{h} = [1, 2, 3]$

This abbreviation denotes the enhancement method with fixed reference harmonics from Section 5.1, where the smoothing filter was applied across time only. The temporal smoothing parameter t_{filt} was set to 40ms. The results for the first three reference harmonics $\bar{h} = [1, 2, 3]$ are displayed in the results section.

TFS $\bar{h} = [1, 2, 3]$

Again, the PQI enhancement method from Section 5.1. This time the smoothing filter was applied across time and then across frequency. The harmonic smoothing parameter h_{filt} was set to 5 and the temporal smoothing parameter t_{filt} was set to 40ms This method was evaluated individually for the first three reference harmonics $\bar{h} = [1, 2, 3]$.

Multiple Reference Phase Methods

Section 6.2.3 displays the evaluation performance of the PQI based phase enhancement method with multiple reference harmonics, discussed in Section 5.2, for different parameter settings.

FS binary mask

Abbreviation for the method introduced in Section 5.2. The smoothing operation was performed across frequency only. The parameter h_{filt} was again set to 5. The enhanced phase was calculated with binary weighting.

FS soft mask

The same method as above, but this time the enhanced phase was calculated with soft weighting.

TS binary mask

The enhancement method from Section 5.2, where the smoothing filter was applied across time with a length of $t_{\text{filt}} = 60$ms. The final enhanced phase was obtained by binary weighting.

TS soft mask

The same parameter setup, but the final enhanced phase was obtained by soft weighting.

TFS binary mask

Again the method from Section 5.2. The smoothing operation was performed across first across time and then across harmonics. The parameters were set to $t_{\text{filt}} = 60$ms and $h_{\text{filt}} = 5$, followed by binary weighting.

TFS soft mask

Finally, the time and frequency smoothing methods combined with soft weighting.

Benchmark Methods

In Section 6.2.5 the best performing PQI based phase enhancement methods are compared to the following benchmark methods:

Maximum a posteriori (MAP)

The phase enhancement method proposed in [13] and explained in detail in Section 3.3. As an analysis and synthesis window, a Blackman window was used. The parameter w_{filt} was set to 20ms, as it was proposed by the authors. PEFAC was used as s noise-robust f_0 estimator.

STFT phase improvement (STFTPI)

The phase estimation method from Section 3.1, which was introduced in [31]. The phase reconstruction was performed along time and frequency. In the analysis and synthesis step a Hanning window was used with an overlap of 87.5%, as it was suggested by the authors. Again with PEFAC as f_0 estimator.

Temporal smoothing of unwrapped phase (TSUP)

The final benchmark method is the TSUP estimator which was explained in detail in Section 3.2 and proposed in [14]. PEFAC was used to estimate f_0.

6.2 Results

In this Section the results for the previously discussed phase enhancement methods are presented.

6.2.1 Fixed Reference Harmonic - White Noise

This Section deals with the comparison of the PQI enhancement methods with fixed reference harmonics with additive white noise. The mean score of PESQ, STOI and UnRMSE of these methods are displayed in Figure 6.2.

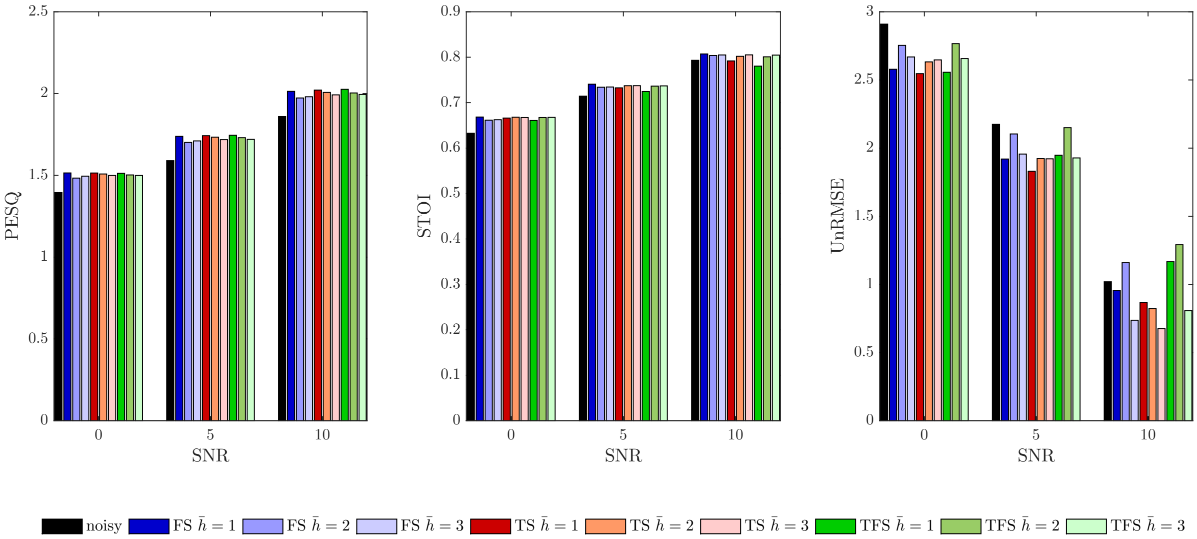

Figure 6.2: PESQ, STOI and UnRMSE for PQI based phase enhancement methods with fixed reference harmonics, evaluated for 50 Speakers of GRID corpus with additive white noise for SNR =[0,5,10], in comparison to the noisy signal.

For a better visualization the results are also displayed in terms of their delta improvement compared to the noisy signal in Figure 6.3, the corresponding values are listed in in Table 6.1. It can be observed that the majority of the proposed methods jointly improve the speech quality and the speech intelligibility for most SNR scenarios. This is an important finding as many speech enhancement methods are reported to degrade speech intelligibility, or are not capable of improving the quality and intelligibility jointly.

From Figure 6.3 it is apparent that in terms of the delta PESQ improvement it is best to rely on the lowest reference harmonic, as all the methods for $\bar{h} = 1$ are superior. In terms of intelligibility and phase estimation error, however, this is not the case. Especially for temporal smoothing the higher reference harmonics result in superior STOI and UnRMSE improvement. Further it also can be observed that the time and frequency smoothed methods, which are displayed in different shades of green, introduce a small phase estimation error at high SNRs (i.e 10 dB).

When it comes to smoothing across harmonics the *FS* $\bar{h} = 1$ method is clearly superior to higher reference harmonics. Overall the *FS* $\bar{h} = 1$ method is arguably the best performing method in terms of the applied evaluation criterion throughout all SNR scenarios, followed by the *TS* $\bar{h} = 2$ methods.

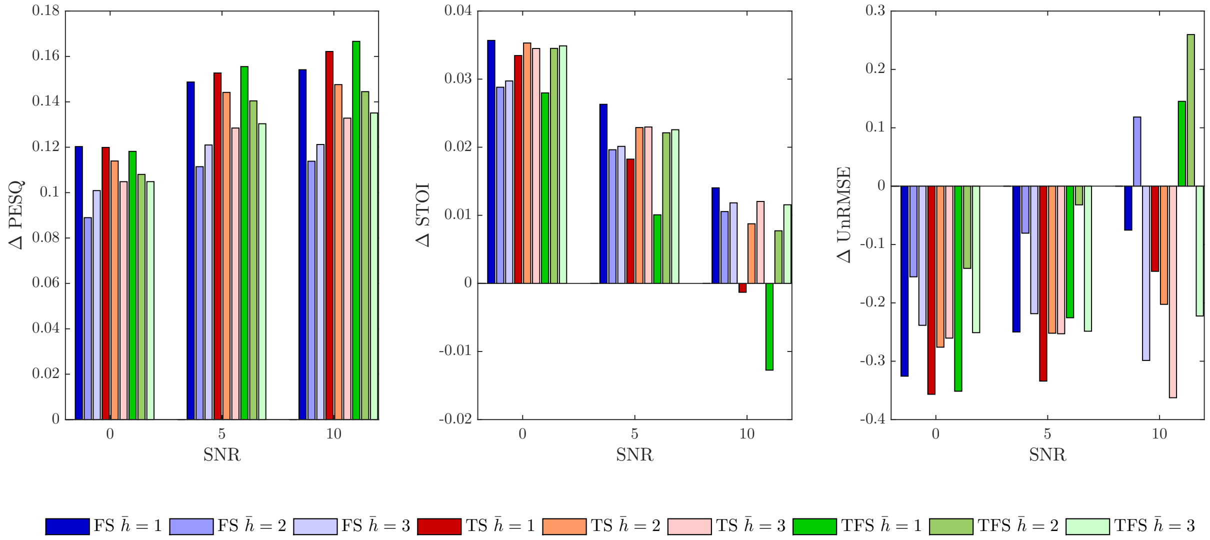

Figure 6.3: ΔPESQ, ΔSTOI and ΔUnRMSE for PQI based phase enhancement methods with fixed reference harmonics, evaluated for 50 Speakers of GRID corpus with additive white noise for SNR =[0,5,10], in comparison to the noisy signal.

	ΔPESQ			ΔSTOI			ΔUnRMSE		
SNR level (dB)	0	5	10	0	5	10	0	5	10
FS $\bar{h} = 1$	**0.12**	0.149	0.154	**0.036**	**0.026**	**0.014**	-0.331	-0.254	-0.063
FS $\bar{h} = 2$	0.089	0.111	0.114	0.029	0.02	0.011	-0.156	-0.07	0.14
FS $\bar{h} = 3$	0.101	0.121	0.121	0.03	0.02	0.012	-0.241	-0.218	-0.282
TS $\bar{h} = 1$	**0.12**	0.153	0.162	0.033	0.018	-0.001	**-0.363**	**-0.344**	-0.151
TS $\bar{h} = 2$	0.114	0.144	0.148	0.035	0.023	0.009	-0.277	-0.252	-0.196
TS $\bar{h} = 3$	0.105	0.128	0.133	0.035	0.023	0.012	-0.262	-0.253	**-0.343**
TFS $\bar{h} = 1$	0.118	**0.156**	**0.167**	0.028	0.01	-0.013	-0.353	-0.226	0.147
TFS $\bar{h} = 2$	0.108	0.14	0.144	0.035	0.022	0.008	-0.143	-0.024	0.272
TFS $\bar{h} = 3$	0.105	0.13	0.135	0.035	0.023	0.012	-0.253	-0.246	-0.212

Table 6.1: Delta scores for PQI based phase enhancement methods with fixed reference harmonics, evaluated for 50 Speakers of GRID corpus with additive white noise for SNR =[0,5,10]. Delta (Left) ΔPESQ, (Middle) ΔSTOI, (Right) ΔUnRMSE

6.2.2 Fixed Reference Harmonic - Babble Noise

This Section deals with the comparison of the PQI enhancement methods with fixed reference harmonics with additive babble noise. The mean score of PESQ, STOI and UnRMSE of these methods are displayed in Figure 6.4.

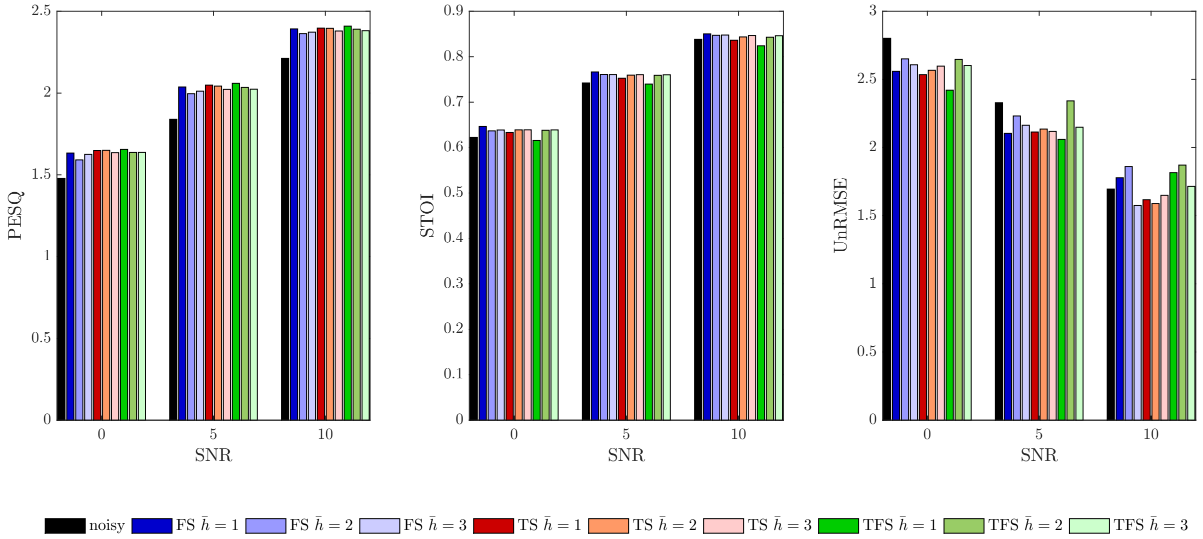

Figure 6.4: PESQ, STOI and UnRMSE for PQI based phase enhancement methods with fixed reference harmonics, evaluated for 50 Speakers of GRID corpus with additive babble noise for SNR =[0,5,10], in comparison to the noisy signal.

Figure 6.5 displays the delta results compared to the noisy signal, the corresponding values are listed in Table 6.2. Figure 6.5 confirms the observations made above. In terms of PESQ it is beneficial to rely on low reference harmonics, in terms of STOI this is not the case. The only exception for this is *FS* $\bar{h}$ *= 1*, which has the overall highest STOI improvement of all the evaluated methods, especially at lower SNRs. For the methods which rely on temporal smoothing (displayed by the green and red shaded bars), it can be observed that a low reference harmonic leads to a decreased performance in terms of STOI, despite the higher PESQ improvement. This also sounds buzzy in a subjective perception.

Based on its relative high intelligibility improvement in babble noise, the *FS* $\bar{h}$ *= 1* method is the best performing algorithm throughout this experiment. Based on this results it can be observed that it is beneficial for frequency smooth in PQI domain to rely on low reference harmonics and for temporal smoothing, higher references are beneficial in terms of STOI and perceived perception.

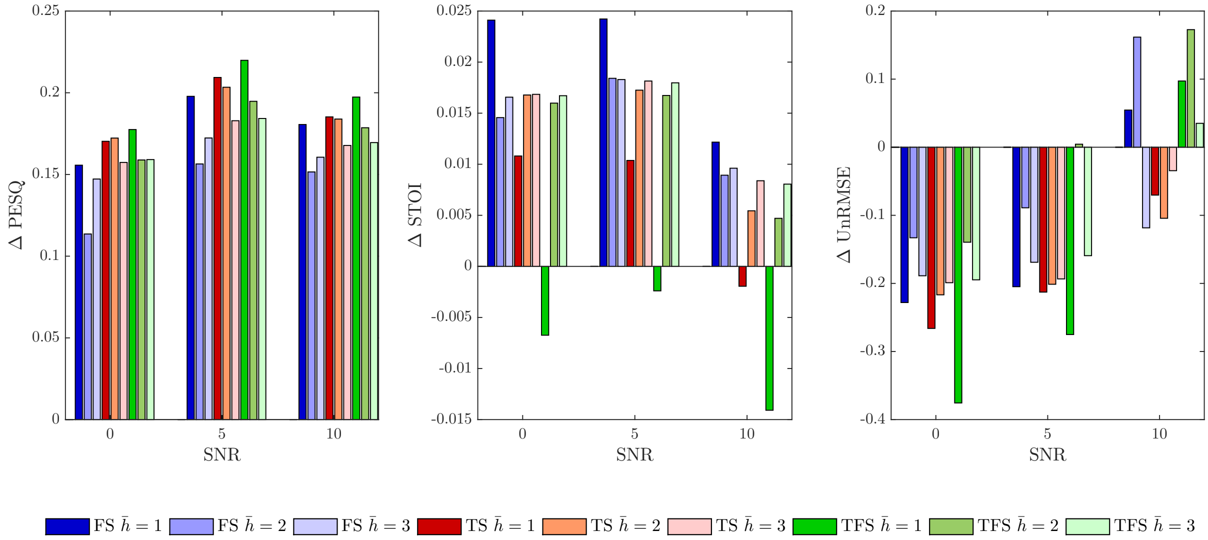

Figure 6.5: ΔPESQ, ΔSTOI and ΔUnRMSE for PQI based phase enhancement methods with fixed reference harmonics, evaluated for 50 Speakers of GRID corpus with additive babble noise for SNR =[0,5,10], in comparison to the noisy signal.

	ΔPESQ			ΔSTOI			ΔUnRMSE		
SNR level (dB)	0	5	10	0	5	10	0	5	10
FS $\bar{h} = 1$	0.156	0.198	0.181	**0.024**	**0.024**	**0.012**	-0.242	-0.225	0.082
FS $\bar{h} = 2$	0.114	0.156	0.152	0.015	0.018	0.009	-0.15	-0.097	0.164
FS $\bar{h} = 3$	0.147	0.172	0.161	0.017	0.018	0.01	-0.194	-0.165	**-0.121**
TS $\bar{h} = 1$	0.17	0.209	0.185	0.011	0.01	-0.002	-0.267	-0.214	-0.079
TS $\bar{h} = 2$	0.172	0.203	0.184	0.017	0.017	0.005	-0.235	-0.193	-0.108
TS $\bar{h} = 3$	0.157	0.183	0.168	0.017	0.018	0.008	-0.204	-0.21	-0.045
TFS $\bar{h} = 1$	**0.177**	**0.22**	**0.197**	-0.007	-0.002	-0.014	**-0.38**	**-0.269**	0.119
TFS $\bar{h} = 2$	0.159	0.195	0.179	0.016	0.017	0.005	-0.155	0.014	0.176
TFS $\bar{h} = 3$	0.159	0.184	0.17	0.017	0.018	0.008	-0.2	-0.179	0.02

Table 6.2: Delta scores for PQI based phase enhancement methods with fixed reference harmonics, evaluated for 50 Speakers of GRID corpus with additive babble noise for SNR =[0,5,10]. Delta (Left) ΔPESQ, (Middle) ΔSTOI, (Right) ΔUnRMSE

6.2.3 Multiple Reference Harmonics - White Noise

In this section the comparison betweens PQI based phase enhancement methods with multiple reference harmonics for additive white noise is discussed. The mean score of PESQ, STOI and UnRMSE for 50 Speakers of GRID corpus are displayed in Figure 6.6. The delta improvement compared to the noisy signal is displayed in Figure 6.7.

In the white noise scenario all methods improve PESQ and STOI throughout all the evaluated SNRs. Most of the enhancement methods also reduce the phase estimation error. The only exceptions here are the time and frequency smoothed methods, which are displayed in cyan, for high SNRs. When comparing the filtering methods, a similar behavior as in Figure 6.2 can be observed.

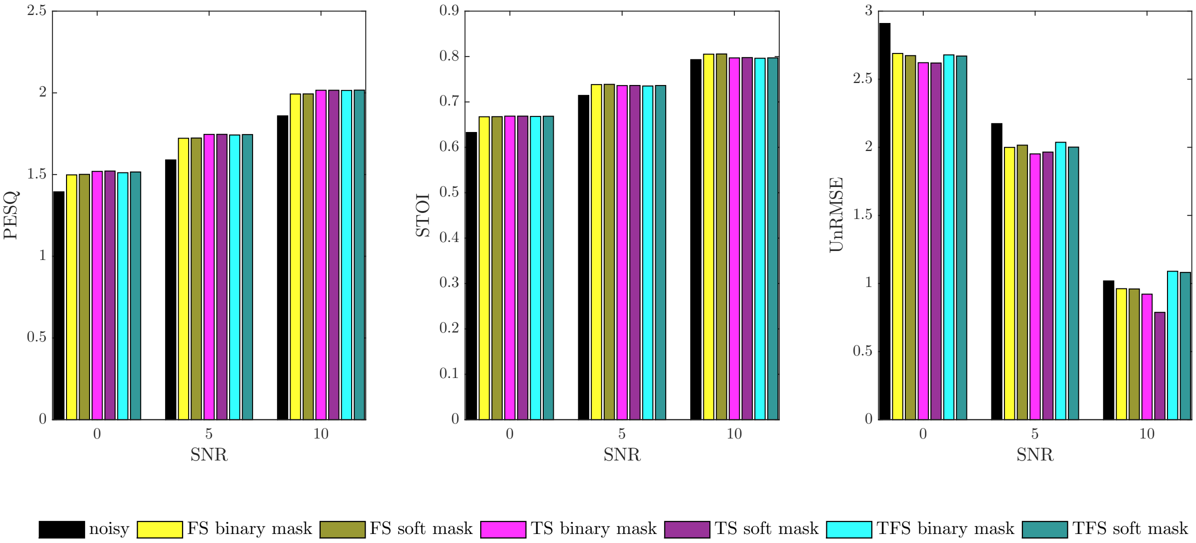

Figure 6.6: PESQ, STOI and UnRMSE for PQI based phase enhancement methods with multiple reference harmonics, evaluated for 50 Speakers of GRID corpus with additive white noise for SNR =[0,5,10], in comparison to the noisy signal.

The filtering methods which include temporal smoothing (which are displayed by the magenta and cyan colored bars in the Figure 6.6 and 6.7) result in a higher PESQ increase than the frequency smoothed only methods. In terms of STOI, the frequency smoothed only methods result in a higher STOI increase at high SNRs.

Overall, it can be observed that for white noise the soft weighted methods perform better than the binary weighted methods throughout almost every scenario. The *TS soft mask* method can be considered as the best performing method due to its relative low spectral phase estimation error at high SNRs, followd by *FS soft mask*. When comparing to the methods with fixed reference harmonics, it can be observed that the level of improvement in term of PESQ, STOI and UnRMSE is basically the same. But the joint improvement seems more consistent and shows little outliers for the multiple reference harmonic methods.

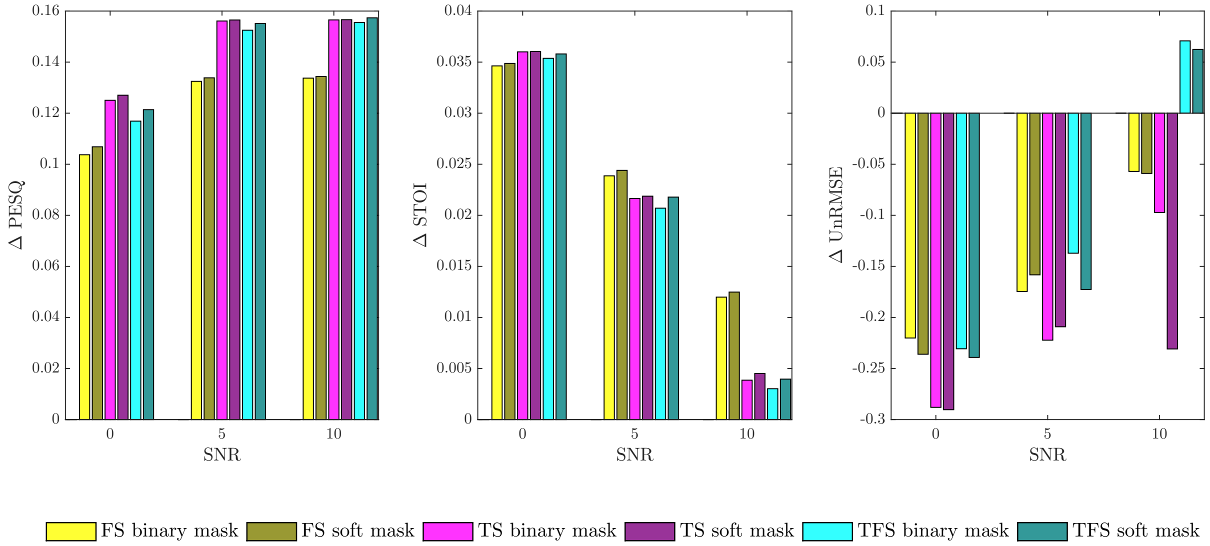

Figure 6.7: ΔPESQ, ΔSTOI and ΔUnRMSE for PQI based phase enhancement methods with multiple reference harmonics, evaluated for 50 Speakers of GRID corpus with additive white noise for SNR =[0,5,10], in comparison to the noisy signal.

	ΔPESQ			ΔSTOI			ΔUnRMSE		
SNR level (dB)	0	5	10	0	5	10	0	5	10
FS binary mask	0.104	0.132	0.134	0.035	**0.024**	**0.012**	-0.22	-0.175	-0.057
FS soft mask	0.107	0.134	0.134	0.035	**0.024**	**0.012**	-0.236	-0.158	-0.059
TS binary mask	0.125	0.156	**0.157**	**0.036**	0.022	0.004	-0.288	**-0.222**	-0.097
TS soft mask	**0.127**	**0.157**	**0.157**	**0.036**	0.022	0.005	**-0.29**	-0.209	**-0.231**
TFS binary mask	0.117	0.152	0.156	0.035	0.021	0.003	-0.231	-0.137	0.071
TFS soft mask	0.121	0.155	**0.157**	**0.036**	0.022	0.004	-0.239	-0.173	0.062

Table 6.3: Delta scores for PQI based phase enhancement methods with multiple reference harmonics, evaluated for 50 Speakers of GRID corpus with additive white noise for SNR =[0,5,10]. Delta (Left) ΔPESQ, (Middle) ΔSTOI, (Right) ΔUnRMSE

6.2.4 Multiple Reference Harmonics - Babble Noise

In this section the comparison between the PQI based phase enhancement methods with multiple reference harmonics for additive babble noise is discussed. The mean score of PESQ, STOI and UnRMSE for 50 Speakers of GRID corpus methods are displayed in Figure 6.8.

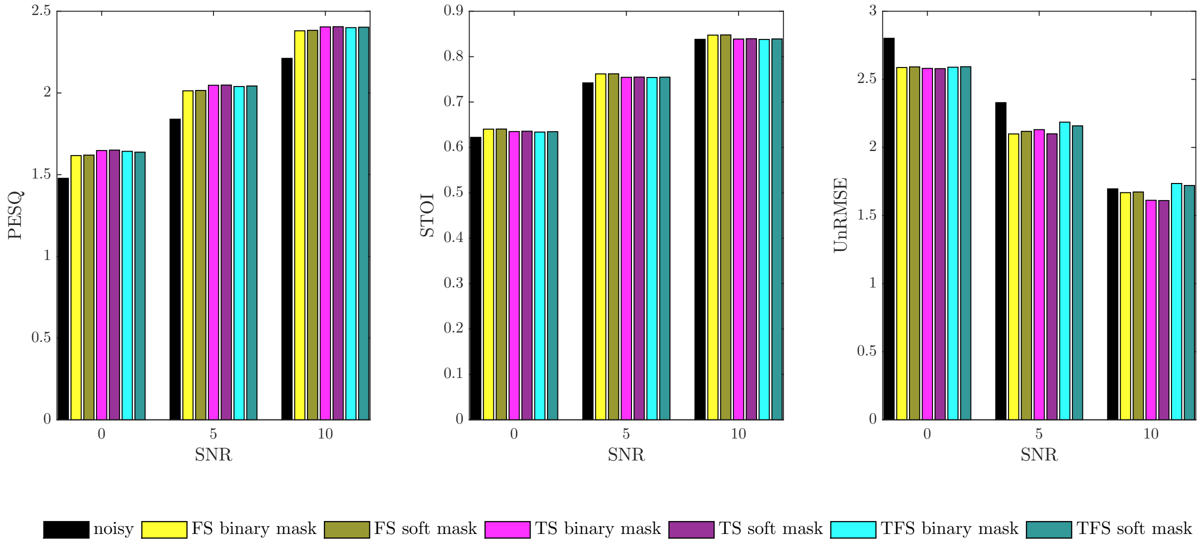

Figure 6.8: PESQ, STOI and UnRMSE for PQI based phase enhancement methods with multiple reference harmonics, evaluated for 50 Speakers of GRID corpus with additive babble noise for SNR =[0,5,10], in comparison to the noisy signal.

The delta improvement compared to the noisy signal is displayed in Figure 6.9, Table 6.4 displays the corresponding values.

Compared to the white noise, the overall STOI improvement decreases and the overall PESQ improvement increases. Again the improvement of the soft weighted methods is higher than the improvement of the binary weighted methods, throughout almost all scenarios. Temporal smoothing methods show an increased PESQ improvement, but the frequency smoothing methods show a much higher STOI improvement. Therefore the *FS soft mask* method can be considered as the best performing algorithm in babble noise.

From all Figures above it can be observed that the methods which apply the smoothing across time and frequency (displayed by the cyan bars) introduce a higher phase estimation error at high SNRs and a lower PESQ and STOI improvement as well. Therefore it is arguably better to perform the PQI smoothing filter either across time or frequency only.

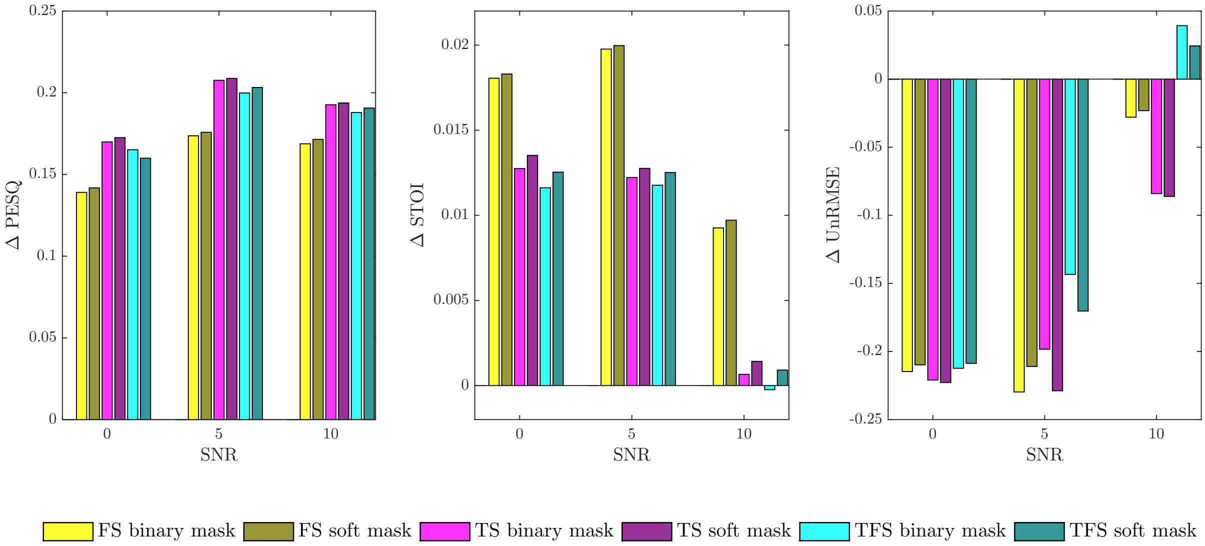

Figure 6.9: ΔPESQ, ΔSTOI and ΔUnRMSE for PQI based phase enhancement methods with multiple reference harmonics, evaluated for 50 Speakers of GRID corpus with additive babble noise for SNR =[0,5,10], in comparison to the noisy signal.

	ΔPESQ			ΔSTOI			ΔUnRMSE		
SNR level (dB)	0	5	10	0	5	10	0	5	10
FS binary mask	0.139	0.174	0.169	**0.018**	**0.02**	0.009	-0.215	**-0.23**	-0.028
FS soft mask	0.142	0.176	0.171	**0.018**	**0.02**	**0.01**	-0.21	-0.211	-0.023
TS binary mask	0.17	0.208	0.193	0.013	0.012	0.001	-0.221	-0.198	-0.084
TS soft mask	**0.173**	**0.209**	**0.194**	0.014	0.013	0.001	**-0.223**	-0.229	**-0.086**
TFS binary mask	0.165	0.2	0.188	0.012	0.012	0	-0.212	-0.143	0.039
TFS soft mask	0.16	0.203	0.191	0.013	0.013	0.001	-0.209	-0.17	0.024

Table 6.4: Delta scores for PQI based phase enhancement methods with multiple reference harmonics, evaluated for 50 Speakers of GRID corpus with additive babble noise for SNR =[0,5,10]. Delta (Left) ΔPESQ, (Middle) ΔSTOI, (Right) ΔUnRMSE

6.2.5 Comparison to Benchmark Methods - White Noise

In this Section 4 of the best performing PQI based phase enhancement methods were selected and compared to the benchmark methods from Section 6.1.3. The PQI based phase enhancement methods, which are *FS* $\bar{h} = 1$, *TS* $\bar{h} = 2$, *FS soft mask* and *TS soft mask*, were selected according to the discussion in the Sections 6.2.1-6.2.4. All of the proposed methods rely on a harmonic signal model, which means they depend heavily on the quality of the frequency estimation. Therefore each method was additionally evaluated for the oracle f_0 case, where the f_0 was estimated by PEFAC applied on the clean signal.

Figure 6.10 shows the comparison of these methods to the benchmarks for white noise scenario, Table 6.5 shows the corresponding values.

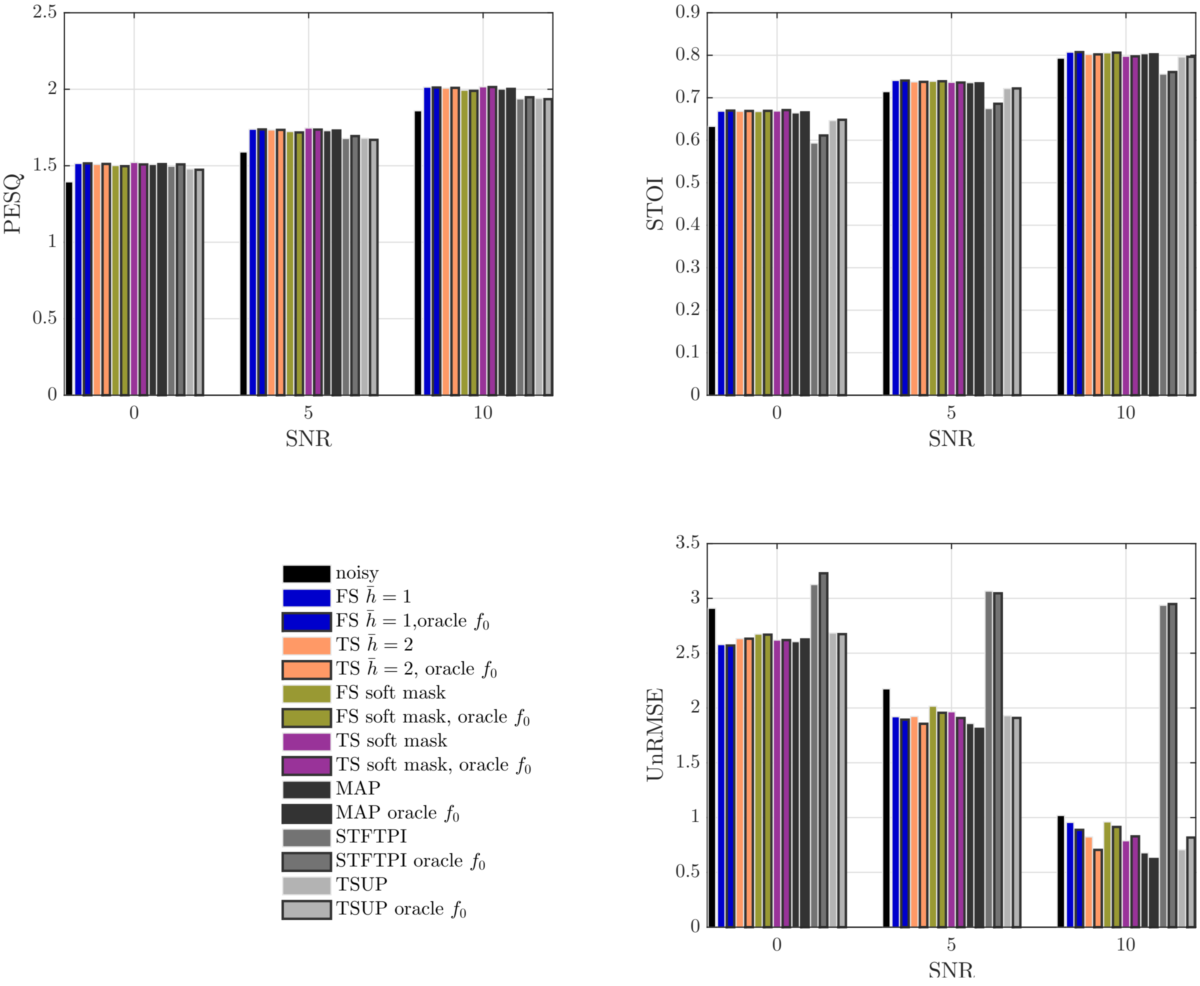

Figure 6.10: PESQ, STOI and UnRMSE for PQI based phase enhancement methods: FS $\bar{h} = 1$*, TS* $\bar{h} = 2$*, FS soft mask, TS soft mask and the benchmark methods: MAP, STFTPI, TSUP, with and without oracle* f_0*. The results were evaluated for 50 Speakers of GRID corpus with additive white noise for SNR =[0,5,10] and displayed in comparison to the noisy signal.*

Figure 6.11 shows the delta improvement of all methods with respect to the noisy signal for white noise.

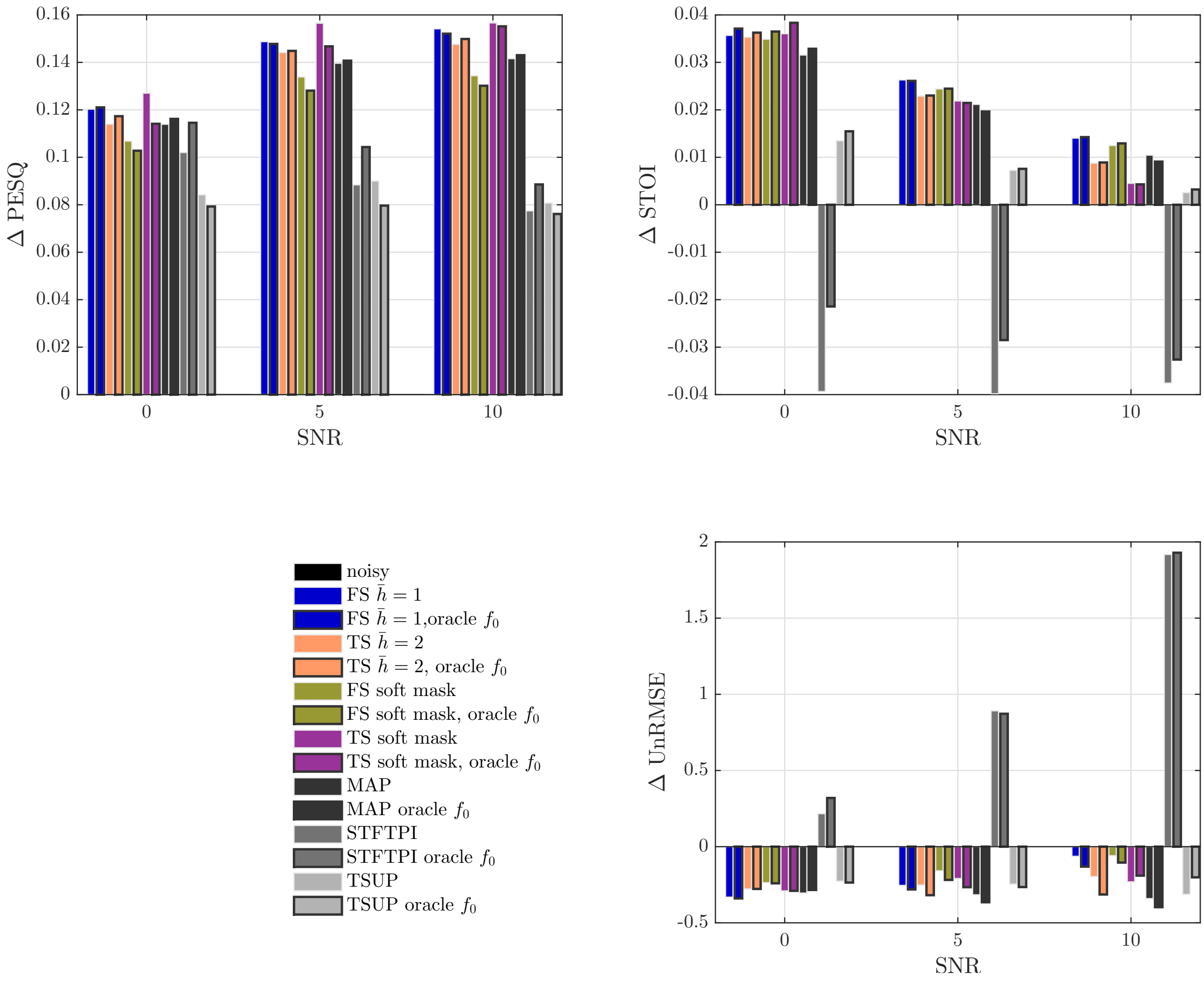

Figure 6.11: ΔPESQ, ΔSTOI and ΔUnRMSE for PQI based phase enhancement methods: FS $\bar{h} = 1$,TS $\bar{h} = 2$,FS soft mask,TS soft mask and the benchmark methods: MAP, STFTPI, TSUP, with and without oracle f_0. The results were evaluated for 50 Speakers of GRID corpus with additive white noise for SNR =[0,5,10] and displayed in comparison to the noisy signal.

In Figure 6.11 it can be observed that the MAP phase estimator is the best performing benchmark method, as it outperforms the other benchmark methods in terms of all evaluation criteria and SNRs. Overall in terms of PESQ the *FS* $\bar{h} = 1$, *TS* $\bar{h} = 2$ and *TS soft mask* methods are better than all benchmark methods, while the *TS soft mask* achieves the highest PESQ.

In terms of STOI *FS* $\bar{h} = 1$ performs best, especially for high SNR scenarios, even tough the delta improvements compared to the noisy signal are rather small for all of examined methods. For STFTPI it can be observed that the jointly improvement of PESQ, STOI and UnRMSE is not possible, as the STOI is degraded and the phase estimation error is increased due to the buzzing sound quality at high frequencies.

In terms of UnRMSE, the PQI based methods decrease in their performance the higher the SNR gets. On average, the MAP benchmark performs best, as it introduces a lower phase estimation error variance at high SNRs.

Overall the f_0 oracle scenario does not change the perceptual evaluation of the different methods by a lot for white noise scenario. For TSUP, *FS soft mask* and *TS soft mask* the PESQ score are sometimes even higher for the case of noisy f_0.

6.2.6 Comparison to Benchmark Methods - Babble Noise

The PQI based phase enhancement methods *FS* $\bar{h} = 1$,*TS* $\bar{h} = 2$,*FS soft mask* and *TS soft mask* are now compared to the benchmark methods in a babble noise scenario. Figure 6.12 shows the PESQ, STOI and UnRMSE of the previously mentioned phase estimators.

	ΔPESQ			ΔSTOI			ΔUnRMSE		
SNR level (dB)	0	5	10	0	5	10	0	5	10
FS $\bar{h} = 1$	0.12	0.149	0.154	0.036	**0.026**	**0.014**	-0.331	-0.254	-0.063
oracle f_0, *FS* $\bar{h} = 1$	0.121	0.148	0.152	0.037	**0.026**	**0.014**	**-0.34**	-0.281	-0.131
TS $\bar{h} = 2$	0.114	0.144	0.148	0.035	0.023	0.009	-0.277	-0.252	-0.196
oracle f_0, *TS* $\bar{h} = 2$	0.117	0.145	0.15	0.036	0.023	0.009	-0.277	-0.318	-0.314
FS soft mask	0.107	0.134	0.134	0.035	0.024	0.012	-0.236	-0.158	-0.059
oracle f_0, *FS soft mask*	0.103	0.128	0.13	0.036	0.024	0.013	-0.24	-0.218	-0.104
TS soft mask	**0.127**	**0.157**	**0.157**	0.036	0.022	0.005	-0.29	-0.209	-0.231
oracle f_0, *TS soft mask*	0.114	0.147	0.155	**0.038**	0.021	0.004	-0.29	-0.266	-0.19
MAP	0.114	0.14	0.142	0.032	0.021	0.01	-0.304	-0.316	-0.341
oracle f_0, *MAP*	0.116	0.141	0.143	0.033	0.02	0.009	-0.284	**-0.365**	**-0.397**
STFTPI	0.102	0.088	0.077	-0.039	-0.04	-0.038	0.216	0.892	1.918
oracle f_0, *STFTPI*	0.115	0.104	0.089	-0.021	-0.029	-0.033	0.319	0.872	1.929
TSUP	0.084	0.09	0.081	0.013	0.007	0.003	-0.227	-0.245	-0.313
oracle f_0, *TSUP*	0.079	0.08	0.076	0.015	0.008	0.003	-0.236	-0.265	-0.201

Table 6.5: Delta (Left) ΔPESQ, (Middle) ΔSTOI, (Right) ΔUnRMSE for PQI based phase enhancement methods: FS $\bar{h} = 1$,*TS* $\bar{h} = 2$,*FS soft mask,TS soft mask and the benchmark methods: MAP, STFTPI, TSUP, with and without oracle* f_0*. The results were evaluated for 50 Speakers of GRID corpus with additive white noise for SNR =[0,5,10] and displayed in comparison to the noisy signal.*

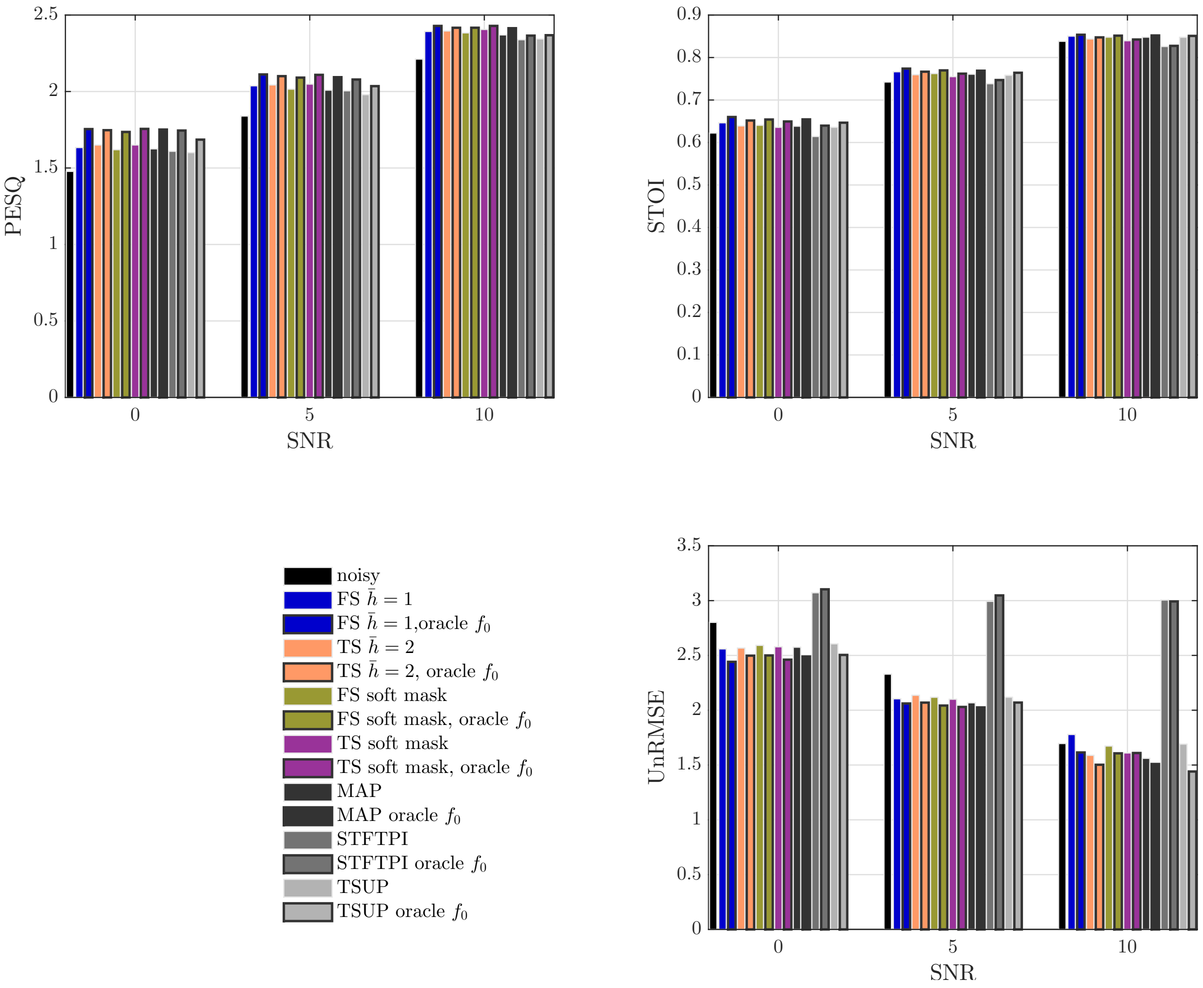

Figure 6.12: PESQ, STOI and UnRMSE for PQI based phase enhancement methods: FS $\bar{h} = 1$,TS $\bar{h} = 2$,FS soft mask,TS soft mask and the benchmark methods: MAP, STFTPI, TSUP, with and without oracle f_0. The results were evaluated for 50 Speakers of GRID corpus with additive babble noise for SNR =[0,5,10] and displayed in comparison to the noisy signal.

For a better visualization, the evaluation scores are displayed in Figure 6.13 in terms of their delta improvement compared to the noisy signal, Table 6.6 shows the corresponding values. In the babble noise scenario the importance of the f_0 estimation gets visible. The oracle f_0 enhances the perceived quality, intelligibility and phase estimation error of all methods by a large amount.

FS $\bar{h} = 1$,*TS* $\bar{h} = 2$ and *TS soft mask* show a higher PESQ improvement than the benchmark methods, while *FS soft mask* is the best performing. In terms of STOI *FS* $\bar{h} = 1$ shows the best performance, followed by the MAP estimator. The STFTPI does decrease the intelligibility, when evaluated upon the noisy f_0. For low SNRs the method evaluated upon the clean f_0 also shows a joint improvement. Again it can be observed that the performance of the PQI methods deceases with higher SNRs. This is visible best for the UnRMSE scores.

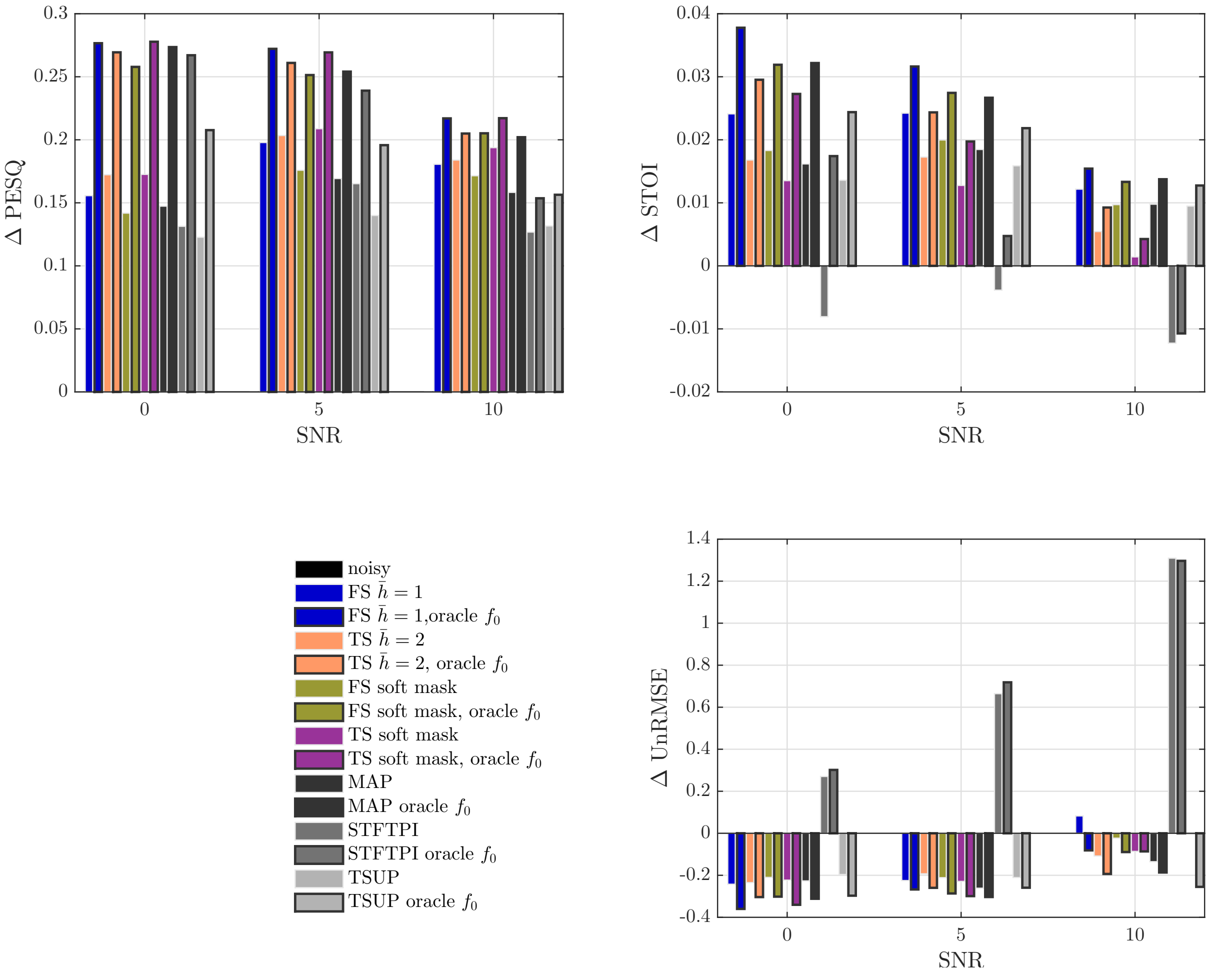

Figure 6.13: ΔPESQ, ΔSTOI and ΔUnRMSE for PQI based phase enhancement methods: FS $\bar{h} = 1$,TS $\bar{h} = 2$,FS soft mask,TS soft mask and the benchmark methods: MAP, STFTPI, TSUP, with and without oracle f_0. The results were evaluated for 50 Speakers of GRID corpus with additive babble noise for SNR =[0,5,10] and displayed in comparison to the noisy signal.

For low SNR scenarios, the *FS* $\bar{h} = 1$ achieves the lowest phase estimation error, but for SNR=10dB a higher phase estimation error is introduced. The MAP estimator again performs best in terms of UnRMSE for SNR of 10dB.

In summary, *FS* $\bar{h} = 1$ can be considered as the best performing phase enhancement method in terms of PESQ and STOI for both noise types. However, for high SNR scenarios a phase estimation error is introduced at some parts of the signal. MAP and *TS soft mask* do not enhance the signal in terms of PESQ and STOI by that much, but they more constantly reduce the phase estimation error at high SNRs.

To conclude this Section, spectrograms of the most relevant phase enhancement methods are presented as a visualization of the enhancement process. Figure 6.14 shows the spectrogram of a female speaker of GRID corpus performing the sentence ”lay red with p 7 soon”. Figure 6.15 shows the spectrogram of a male speaker of GRID corpus performing the sentence ”lay green at l 3 again”. The spectrograms and corresponding PESQ and STOI values of the clean signal, noisy signal, *FS* $\bar{h} = 1$ enhanced signal, *TS soft mask* enhanced signal, STFTPI enhanced signal and MAP enhanced signal are presented for both speakers.

	ΔPESQ			ΔSTOI			ΔUnRMSE		
SNR level (dB)	0	5	10	0	5	10	0	5	10
FS $\bar{h} = 1$	0.156	0.198	0.181	0.024	0.024	0.012	**-0.242**	-0.225	0.082
oracle f_0, *FS* $\bar{h} = 1$	0.277	**0.272**	**0.217**	**0.038**	**0.032**	**0.015**	-0.36	-0.267	-0.081
TS $\bar{h} = 2$	0.172	0.203	0.184	0.017	0.017	0.005	-0.235	-0.193	-0.108
oracle f_0, *TS* $\bar{h} = 2$	0.269	0.261	0.205	0.03	0.024	0.009	-0.304	-0.26	-0.194
FS soft mask	0.142	0.176	0.171	0.018	0.02	0.01	-0.21	-0.211	-0.023
oracle f_0, *FS soft mask*	0.258	0.251	0.205	0.032	0.027	0.013	-0.302	-0.286	-0.089
TS soft mask	0.173	0.209	0.194	0.014	0.013	0.001	-0.223	-0.229	-0.086
oracle f_0, *TS soft mask*	**0.278**	0.269	**0.217**	0.027	0.02	0.004	-0.34	-0.299	-0.086
MAP	0.147	0.169	0.158	0.016	0.018	0.01	-0.227	-0.261	-0.135
oracle f_0, *MAP*	0.274	0.254	0.202	0.032	0.027	0.014	-0.312	**-0.303**	-0.187
STFTPI	0.131	0.165	0.127	-0.008	-0.004	-0.012	0.27	0.664	1.309
oracle f_0, *STFTPI*	0.267	0.239	0.154	0.017	0.005	-0.011	0.302	0.719	1.297
TSUP	0.123	0.14	0.132	0.014	0.016	0.009	-0.196	-0.21	-0.005
oracle f_0, *TSUP*	0.208	0.196	0.156	0.024	0.022	0.013	-0.297	-0.259	**-0.255**

Table 6.6: Delta (Left) ΔPESQ, (Middle) ΔSTOI, (Right) ΔUnRMSE for PQI based phase enhancement methods: FS $\bar{h} = 1$*,TS* $\bar{h} = 2$*,FS soft mask,TS soft mask and the benchmark methods: MAP, STFTPI, TSUP, with and without oracle* f_0*. The results were evaluated for 50 Speakers of GRID corpus with additive white babble for SNR =[0,5,10] and displayed in comparison to the noisy signal.*

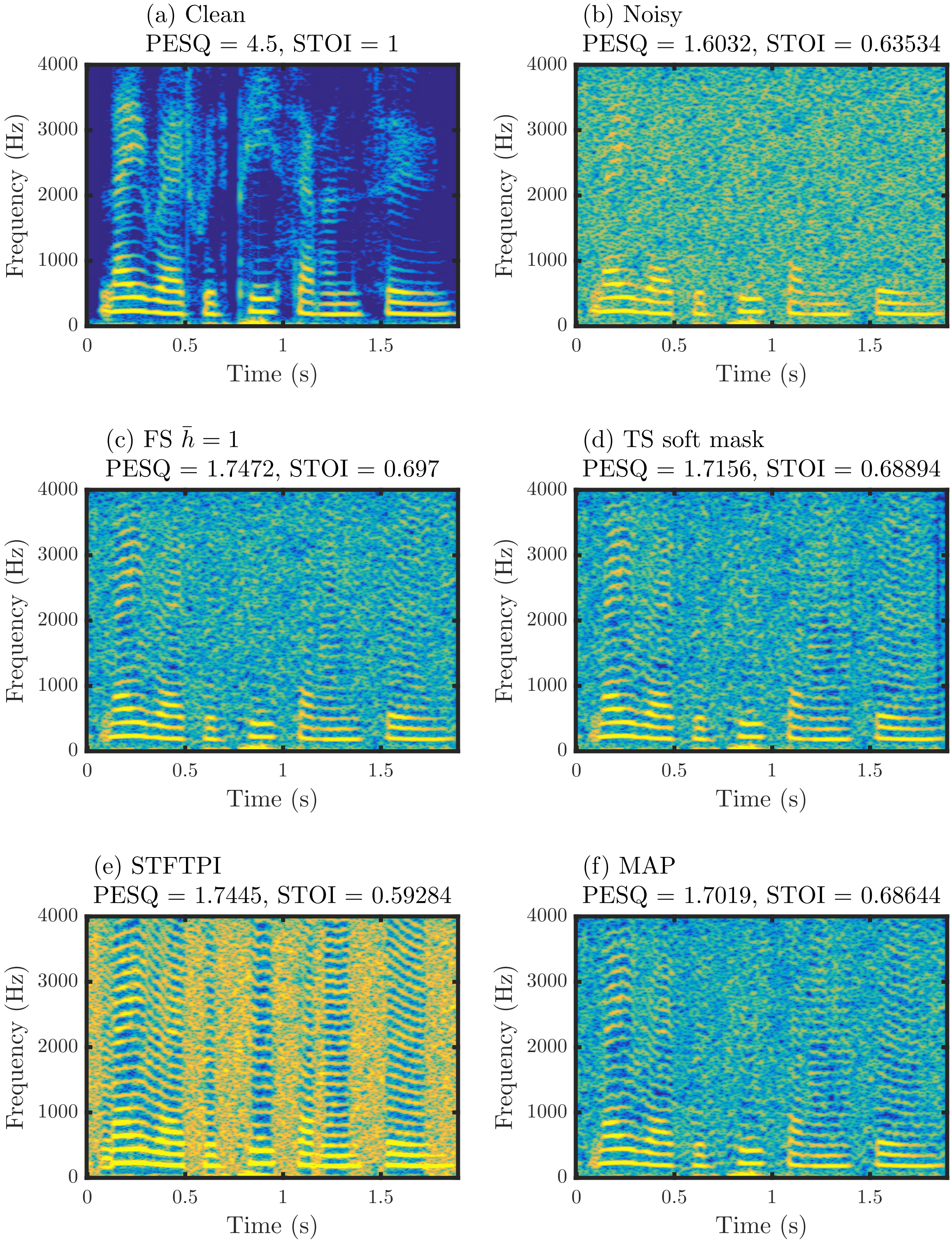

Figure 6.14: Spectrogram and corresponding PESQ and STOI values of: (a) clean signal, (b) noisy signal, (c) FS $\bar{h} = 1$ enhanced signal, (d) TS soft mask enhanced signal, (e) STFTPI enhanced signal and (f) MAP enhanced signal of a female speaker of GRID performing ”lay red with p 7 soon”.

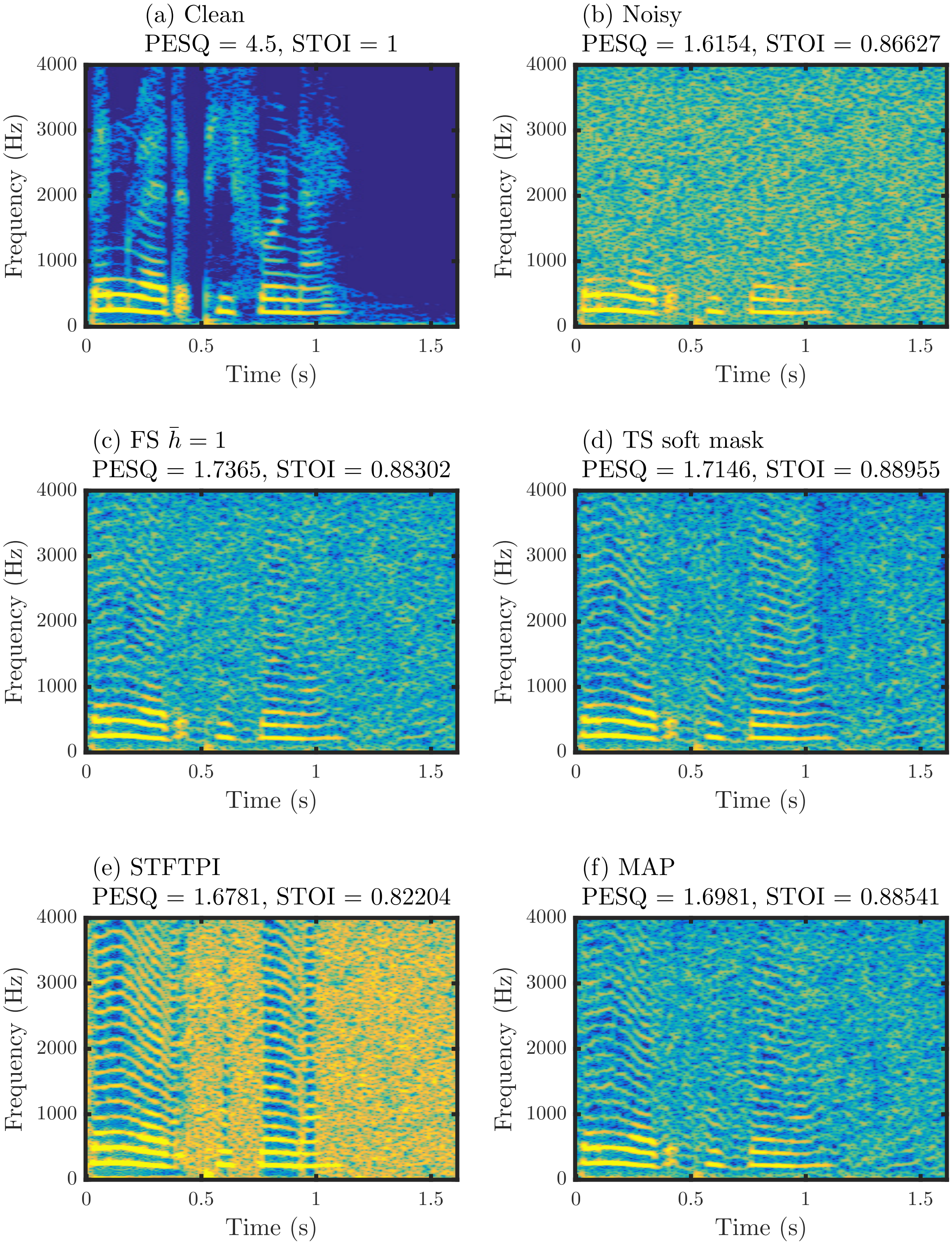

Figure 6.15: Spectrogram and corresponding PESQ and STOI values of: (a) clean signal, (b) noisy signal, (c) FS $\bar{h} = 1$ enhanced signal, (d) TS soft mask enhanced signal, (e) STFTPI enhanced signal and (f) MAP enhanced signal of a male speaker of GRID performing "lay green at l 3 again".

7 Conclusion and Future Outlook

In this work several phase enhancement methods were presented and compared in terms of their achievable perceptual quality. Different state-of-the-art phase estimators which were later documented as benchmark methods were explained in detail.

Due to its wrapping properties, the spectral phase was long believed to be structureless and insignificant for the use in speech enhancement systems. In the course of this work the Phase Quasi Invariant (PQI), a novel harmonic phase representation method for single channel speech enhancement, was introduced. This representation uncovers the hidden structure of the instantaneous phase, which makes spectral phase modification possible.

From this phase representation two harmonic phase estimation methods were derived, which aim to reduce the non-deterministic part of the instantaneous phase. The methods were implemented with different parameter setting and smoothing filter and were compared to three benchmark methods. The results obtained in the experiments demonstrated the potential of PQI based phase enhancement methods. FS $\bar{h} = 1$ performed best in reducing the noise and improving both perceived quality and speech intelligibility jointly. This behavior was observed for both noise types and for all SNR levels.

Conventional amplitude based enhancement methods, which still show a higher efficiency in terms of noise reduction than phase enhancement methods, have currently been used to their full potential, while the field of phase enhancement has yet to be explored more. Consequentially, future work may include the investigation of joint phase and amplitude enhancement methods. Also the new phase estimates could be efficiently used together with phase-aware amplitude estimators, to surpass the limits of the current single channel speech enhancement systems.

A

Appendix

In the following the paper ”Phase Estimation in Single-channel Speech Enhancement using Phase Invariance Constraints” which was written in the course of this work is presented. The paper got accepted for presentation in a poster session at ICASSP 2017.

PHASE ESTIMATION IN SINGLE-CHANNEL SPEECH ENHANCEMENT USING PHASE INVARIANCE CONSTRAINTS

Michael Pirolt[*], *Johannes Stahl*[*], *Pejman Mowlaee*[*],
Vasili I. Vorobiov[**], *Siarhei Y. Barysenka*[**], *Andrew G. Davydov*[**]

[*] Signal Processing and Speech Communication Lab, Graz University of Technology, Graz, Austria
[**] Belarusian State University of Informatics and Radioelectronics, Minsk, Belarus
michael.pirolt@student.tugraz.at, johannes.stahl@tugraz.at,
pejman.mowlaee@tugraz.at, viv314@gmail.com,
siarhei.barysenka@gmail.com, agdavydov81@gmail.com

ABSTRACT

Phase-aware signal processing has received increasing interest in many speech applications. The success of phase-aware processing is governed with having access to robust estimates of the clean spectral phase to be obtained from noisy observation. In this paper, we propose a novel harmonic phase estimator relying on the phase invariance property exploiting relations between harmonics using the phase structure. We present speech quality results achieved in speech enhancement to justify the effectiveness of the proposed phase estimator. compared to noisy phase and other phase estimation benchmarks.

***Index Terms*—** Phase estimation, phase invariance, speech enhancement, speech quality.

1. INTRODUCTION

Speech signal processing methods often ignore the processing of spectral phase information. Performance gain can be achieved when an enhanced spectral phase or some additional information about phase is incorporated. For general reviews on recent advances in phase-aware signal processing and its applications in speech communication we refer to [1–3].

In particular, in the field of noise reduction the importance of phase receives increasing attention by researchers. Some examples are, model-based short-time Fourier transform (STFT) phase improvement [4], maximum a posteriori harmonic (MAP) phase estimation [5], temporal smoothing of the unwrapped harmonic phase [6], and finally the reviews on phase estimation impact on enhancement have been investigated in [7]. Apart from improved signal reconstruction, spectral phase information can be also used to derive improved spectral amplitude estimators, see e.g. [8, 9].

The advances due to phase-aware processing are limited by the accuracy of the estimated phase. Therefore, a challenging research topic is to find novel approaches that help to achieve more robust and accurate estimators of the clean spectral phase from the noisy speech observation.

In this paper, we propose exploiting the relation between the phase of harmonics of a speech signal. The so-derived harmonic phase estimator results in improved perceived quality and speech intelligibility, and a low phase estimation error.

The rest of the paper is organized as follows. Section 2 presents the background on phase invariance and phase quasi-invariance properties. Section 3 presents the proposed phase enhancement scheme. Section 4 presents a proof-of-concept experiment and speech enhancement results and Section 5 concludes the work.

2. BACKGROUND ON PHASE INVARIANCE PROPERTY

2.1. Phase Invariant

Phase invariant constraint (PI) was first introduced by Zverev in ultrasonic dispersion measurements [10], where it was reported that the harmonic oscillation contains a phase structure which is invariant to the time reference. In harmonic signal, PI can be determined for any triplet of harmonic components if their frequencies satisfy the set of equations:

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

In these equations, F_0 denotes fundamental frequency. For the given polyharmonic signal $s(t)$ with time index t, that contains $h \in [1, H_t]$ harmonics with slowly varying amplitude $A(h, t)$ and phase $\Phi(h, t)$:

The work of Pejman Mowlaee, Michael Pirolt and Johannes Stahl was supported by the Austrian Science Fund (project number P28070-N33).

$$s(t) = \sum_{h=1}^{H_t} s(h,t) = \sum_{h=1}^{H_t} A(h,t)\cos\underbrace{(2\pi f_h(t)t + \Phi(h,t))}_{\Psi(h,t)} \tag{2}$$

the PI denoted by $\Delta\Psi(t)$, for $H_t = 3$ is given by:

$$\begin{aligned}\Delta\Psi(t) &= \frac{\Psi(1,t)+\Psi(3,t)}{2} - \Psi(2,t) \\ &= \frac{\Phi(1,t)+\Phi(3,t)}{2} - \Phi(2,t),\end{aligned} \tag{3}$$

where $\Psi(h,t)$ denotes instant phase. It is important to mention that cancellation of linear items $2\pi f_h(t)t$ can be achieved only when functions of instant phase $\Psi(h,t)$ are continuous and have no wraps. This can be ensured using the phase unwrapping procedure.

2.2. Phase Quasi-Invariant

Phase quasi-invariant constraint (PQI) was introduced by Vorobiov within the analysis of phase relations in speech [11]. The application of this constraint together with the PI was outlined for speech analysis [12, 13].

For the given polyharmonic signal $s(t)$ with fundamental frequency $F_0(t)$ that contains $h \in [1, H_t]$ harmonics with slowly varying amplitude $A(h,t)$ and phase $\Phi(h,t)$

$$s(t) = \sum_{h=1}^{H_t} s(h,t) = \sum_{h=1}^{H_t} A(h,t)\cos(2\pi h F_0(t)t + \Phi(h,t)), \tag{4}$$

the following relation $\Delta\Psi_{\bar h}(h,t)$ between components with $\bar h F_0(t)$ and $hF_0(t)$ frequencies, where $\bar h < h$, does not have linear components similar to PI in Eq. (3):

$$\begin{aligned}\Delta\Psi_{\bar h}(h,t) &= \Psi(\bar h,t) - \frac{\Psi(h,t)\cdot\bar h}{h} \\ &= \left(\Phi(\bar h,t) - \frac{\Phi(h,t)\cdot\bar h}{h}\right)\Bigg|_{\frac{2\pi\bar h}{h}}.\end{aligned} \tag{5}$$

The equation above is called *phase quasi-invariant* (PQI). It is also required to unwrap instant phase functions $\Psi(\bar h,t)$ and $\Psi(h,t)$ before calculating PQI. The unambiguous definition range of PQI is $[0, \frac{2\pi\bar h}{h})$ if value of harmonic phase $\Phi(\bar h,t)$ at instant $t = 0$ is in $[0, 2\pi)$, which can be easily achieved in speech processing.

Another signal representation based on the phase difference measure is the Relative Phase Shift (RPS) [14]. The relation between RPS and PQI can be depicted as:

$$RPS(h,t) = \Phi(h,t) - h\Phi(1,t) = -h\Delta\Psi_1(h,t). \tag{6}$$

Eq. (6) shows that the RPS can be represented by the negative PQI with $\bar h = 1$, multiplied with the harmonic index. While the RPS depicts the phase difference only between the fundamental frequency and its higher harmonics, the PQI is not limited to the fundamental frequency phase, as the reference harmonic $\bar h$ is free to choose.

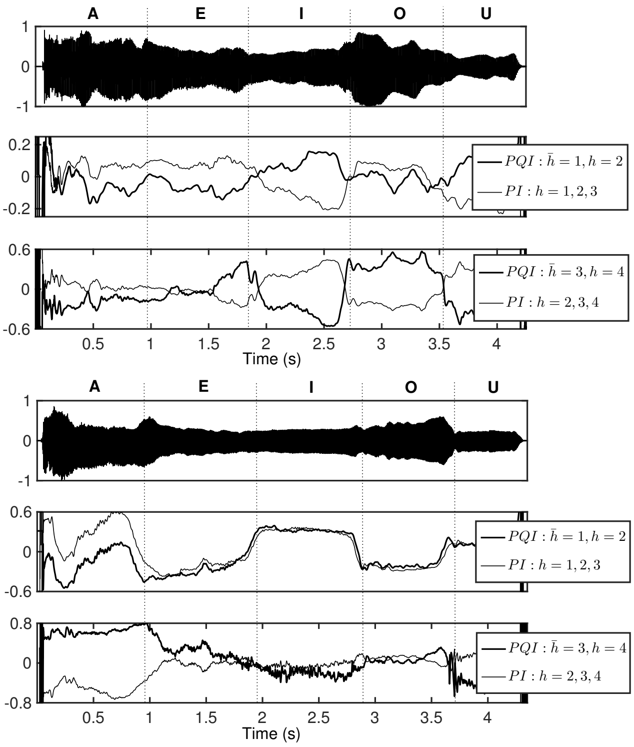

Fig. 1. The results of proof-of-concept experiments of PI and PQI for sustained A-E-I-O-U sequence. (Top) Male speaker. $F_0 = 118 \pm 2$ Hz along the whole record, (Bottom) Female speaker. $F_0 = 220 \pm 5$ Hz along the whole record.

2.3. Suitability for Phase-aware Speech Processing

To demonstrate the smoothness along time in voiced speech using PI and PQI constraints, records of sustained vowels A-E-I-O-U were analyzed in PI and PQI domain. The results of this analysis for male and female speakers are shown in Figure 1 up to an additive constant[1].

First, the instantaneous pitch estimation $F_0(t)$ was obtained. Next, the instant phase functions $\Psi(h,t)$ were calculated using Hilbert transform for filtered $hF_0(t)$ where $h \in [1,4]$. After that, the phase unwrapping procedure was applied for each $\Psi(h,t)$. Finally, the phase characteristics were calculated. In order to unify the scale of all these functions for representation purpose, each of PQI was normalized to the half of its unambiguous definition range, whereas PI was normalized to π.

The PI and PQI properties show similar trends, or they become similar after inversion of one in the pair. Some curves have phase jumps at instants where one vowel changes to an-

[1] The implementation can be found at [15].

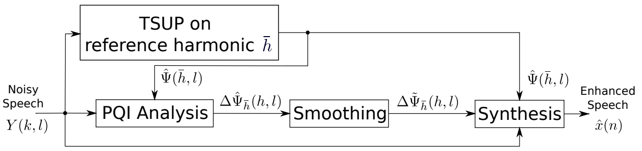

Fig. 2. Illustration of the proposed phase estimator.

other. These results support that PI and PQI constraints carry information about the structure of voiced speech, so they are favorable candidates for phase-aware speech processing.

3. PROPOSED PHASE ESTIMATOR

The idea is to apply temporal smoothing on the PQI extracted from the noisy speech signal in order to reduce its variance. This is motivated by the successful results reported in TSUP [6, 7] and will be justified within the proof-of-concept experiment presented in this Section. An overview of the proposed method is depicted in Figure 2.

3.1. PQI Framework

In this Section, the proposed phase enhancement framework is presented. As explained in Section 2, the PQI is based on the phase difference measures between two harmonics. Therefore we model the noisy signal as the sum of harmonics corresponding to the clean signal $x(n)$ with some noise added. The noisy signal is represented by an assembly of signal frames $y(n,l)$ where n denotes the discrete time index, N the frame length with discrete time index $n \in [0, N-1]$ and l denotes the frame index and we have:

$$y(n,l) = \underbrace{\sum_{h=1}^{H_l} A(h,l)\cos\left(h \cdot 2\pi \frac{F_0(l)}{f_s} n + \Phi(h,l)\right)}_{x(n)\ldots\text{clean signal}} + \nu(n,l), \tag{7}$$

where $\omega_0(l)$ denotes the normalized angular fundamental frequency at frame l, $\nu(n,l)$ denotes the noise and h denotes the harmonic index with $h \in [1, H_l]$ and H_l denotes the number of harmonics at frame l. The time instances at each frame t_l are calculated according to [16]:

$$t_l = t_{l-1} + \frac{1}{4 \cdot F_0(l-1)} \tag{8}$$

3.2. Calculation of PQI

The PQI values are calculated based on Eq. (5). Since the output of Eq. (5) gives us a cyclic random variable with the unambiguous definition range of $\left[\frac{-\pi\bar{h}}{h}, \frac{\pi\bar{h}}{h}\right)$, it is recommended to add a scaling factor after wrapping to ensure an unambiguous definition range of $[-\pi, \pi)$, please note that the PQI is independent of the fundamental frequency, therefore it yields:

$$\begin{aligned} \Delta\Psi_{\bar{h}}(h,l) &= \frac{h}{\bar{h}}\left(\Phi(\bar{h},l) - \frac{\Phi(h,l)\cdot\bar{h}}{h}\right)\Bigg|_{\frac{2\pi\cdot\bar{h}}{h}} \\ &= \frac{h}{\bar{h}}\left(\Psi(\bar{h},l) - \frac{\Psi(h,l)\cdot\bar{h}}{h}\right)\Bigg|_{\frac{2\pi\cdot\bar{h}}{h}}. \end{aligned} \tag{9}$$

The PQI can be evaluated for every arbitrary pair $\{h, \bar{h}\} \in [1, H_l]$. For all further observations, the harmonic index $\bar{h}$ is referred to as PQI reference harmonic, while h denotes the harmonic index. Furthermore, the reference harmonic $\bar{h}$ is chosen with 2 and therefore not changed during the process.

3.3. Temporal Smoothing of PQI

From Eq. (9), the harmonic phase of an arbitrary harmonic $h \in [1, H_l]$ can be reformulated using PQI and the corresponding reference harmonic phase $\bar{h}$:

$$\Psi(h,l) = \frac{h \cdot \Psi(\bar{h},l)}{\bar{h}} - \Delta\Psi_{\bar{h}}(h,l). \tag{10}$$

The PQI reference phase $\Psi(\bar{h},l)$ is of high importance, as corruption with noise leads to erroneous results for the corresponding harmonic phases throughout the harmonics. Therefore it is recommended to pre-enhance the reference phase. For the following observations, we used TSUP [6] solely on the reference phase.

The PQI values are then calculated based on the pre-enhanced reference phases $\hat{\Psi}(\bar{h},l)$:

$$\Delta\hat{\Psi}_{\bar{h}}(h,l) = \frac{h}{\bar{h}}\left(\hat{\Psi}(\bar{h},l) - \frac{\Psi(h,l)\cdot\bar{h}}{h}\right)\Bigg|_{\frac{2\pi\cdot\bar{h}}{h}}. \tag{11}$$

The differential phases obtained from Eq. (11) are then smoothed across time, by mean averaging:

$$\Delta\tilde{\Psi}_{\bar{h}}(h,l) = \angle\frac{1}{\mathcal{W}}\sum_{\tilde{l}\in\mathcal{W}} e^{j\Delta\hat{\Psi}_{\bar{h}}(h,\tilde{l})}, \tag{12}$$

where $\mathcal{W}$ denotes all frames that lie within a range of 100 milliseconds around frame l. This filter length was chosen empirically, after observing the PQI behavior over time.

3.4. Synthesizing Phase-Enhanced Speech

The signal synthesis is based on [5], as the enhanced harmonic phase is transformed to the STFT domain by modifying the frequency bins within the main-lobe width of the analysis window. We define $Y(k,l)$ as the DFT of the noisy signal with k as the corresponding frequency bin and K as the DFT length with $k \in [0, K-1]$. Further $|Y(k,l)|$ denotes the noisy spectral amplitude and the noisy STFT phase, i.e.,

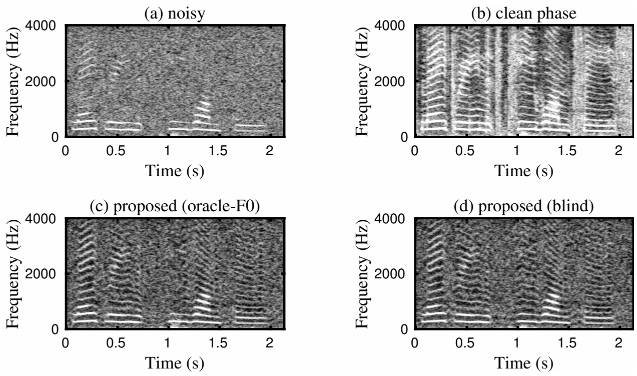

Fig. 3. Spectrogram of a female utterance with white noise at SNR $= 5$ dB. a: Noisy phase. b: Clean phase. c: Proposed method and clean F_0. d: Proposed method blind.

$\vartheta(k,l) = \angle Y(k,l)$. The enhanced STFT phase is then given by:

$$\hat{\vartheta}(\lfloor h\omega_0(l)K\rfloor + i, l) = \left(\frac{h \cdot \hat{\Psi}(\bar{h},l)}{\bar{h}} - \Delta\tilde{\Psi}_{\bar{h}}(h,l)\right), \quad (13)$$
$$\forall i \in [-N_p(l)/2, N_p(l)/2].$$

where $N_p(l)$ denotes the minimum value of either the main-lobe width of the analysis window N_w or the frequencies close to neighboring harmonic $N_p(l) = \min(N_w, \omega_0(l)K/(2\pi))$. Further we obtain the phase enhanced signal in STFT domain by:

$$\hat{X}(k,l) = |Y(k,l)|e^{j\hat{\vartheta}(k,l)}. \quad (14)$$

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

4. RESULTS

4.1. Experiment Setup

We randomly chose 50 utterances spoken by 20 speakers (10 female and 10 male) from GRID [17] and mixed them with white and babble noise from NOISEX-92 [18] at SNRs between 0 to 10 dB. As evaluation criteria we chose perceptual evaluation of speech quality (PESQ) [19], short-term objective intelligibility measure (STOI) [20] and unwrapped root mean square estimation error (UnRMSE) [21] in dB.

4.2. Speech Enhancement Results

Figure 3 shows the proof of concept experiment carried out on a female speech sample mixed with white noise at SNR $= 5$ dB. The phase-enhanced results using the proposed method shows an improved harmonic structure closer to that observed in the clean phase. This harmonic structure was lost in the noisy scenario where noisy spectral phase only available.

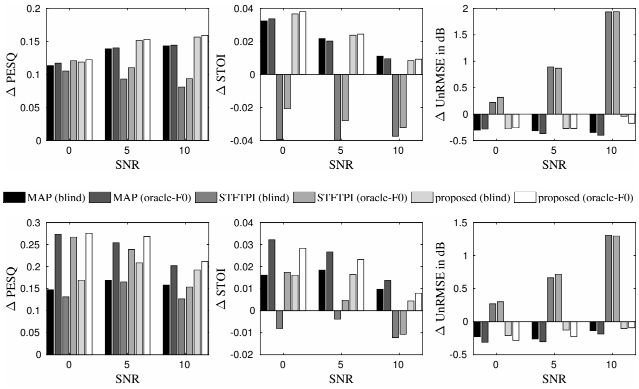

Fig. 4. PESQ improvement, STOI improvement and UnRMSE improvement in dB for (top) white and (bottom) babble noise.

The columns in Figure 4 show, respectively, the delta improvement compared to the noisy signal by means of (left) perceived quality, (middle) speech intelligibility, and (right) UnRMSE [21]. The results are averaged over utterances for white and babble noise. The reported results for noisy phase demonstrate the lower-bound. As benchmarks, we document the performance of STFTPI [4] and MAP [5], both in combination with PEFAC [22] as noise-robust F_0-estimator.

In white noise, the method is not that sensitive to F_0 estimation accuracy. However, in the babble noise scenario the achievable performance by the phase enhancement methods is dependent on F_0 estimation accuracy. Overall, the proposed method improves the perceived quality, speech intelligibility and phase estimation error for all SNRs and noise types. This is an important finding in that most speech enhancement methods are reported to degrade speech intelligibility or not capable for joint improvement of perceived quality and intelligibility. In terms of speech quality, the proposed method outperforms all benchmark methods at all SNRs. In terms of the speech intelligibility, the proposed method shows less impact at high SNRs. In terms of UnRMSE, the MAP estimate [5] is superior which is attributed to the fact it relies on prior information about SNR, which is not taken into account in the proposed estimator. For listening examples we refer to webpage [23].

5. CONCLUSION

The paper proposed a new harmonic phase estimator from noisy speech relying on relations between harmonics using the phase structure across harmonics. Temporal smoothing of the phase invariance representation allows for selective smoothing at harmonic level and contributes to improved speech quality when used at signal reconstruction. Here we concentrated on signal reconstruction, the enhanced spectral phase is also useful in speech recognition [24] and separation [25] tobe considered as future works.

6. REFERENCES

[1] P. Mowlaee, R. Saeidi, and Y. Stylianou, "Advances in phase-aware signal processing in speech communication," *Speech communication*, vol. 81, pp. 1–29, 2016.

[2] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase Processing for Single-Channel Speech Enhancement: History and recent advances," *IEEE Trans. Magn.*, vol. 32, no. 2, pp. 55–66, March 2015.

[3] P. Mowlaee, J. Kulmer, J. Stahl, and F. Mayer, "Phase-Aware Signal Processing in Speech Communication: History, Theory and Practice," *John Wiley & Sons*, 2016.

[4] M. Krawczyk and T. Gerkmann, "STFT Phase Reconstruction in Voiced Speech for an Improved Single-Channel Speech Enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 22, no. 12, pp. 1931–1940, Dec. 2014.

[5] J. Kulmer and P. Mowlaee, "Harmonic Phase Estimation in Single-Channel Speech Enhancement Using Von Mises Distribution and Prior SNR," in *Proc. ICASSP*, Apr. 2015, pp. 5063–5067.

[6] J. Kulmer and P. Mowlaee, "Phase Estimation in Single Channel Speech Enhancement Using Phase Decomposition," *IEEE Signal Process. Lett.*, vol. 22, no. 5, pp. 598–602, May. 2015.

[7] P. Mowlaee and J. Kulmer, "Phase Estimation in Single-Channel Speech Enhancement: Limits-Potential," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 23, no. 8, pp. 1283–1294, Aug. 2015.

[8] P. Mowlaee and R. Saeidi, "Iterative Closed-Loop Phase-Aware Single-Channel Speech Enhancement," *IEEE Signal Process. Lett.*, vol. 20, no. 12, pp. 1235–1239, Dec. 2013.

[9] T. Gerkmann, "Bayesian Estimation of Clean Speech Spectral Coefficients Given a Priori Knowledge of the Phase," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 62, no. 16, pp. 4199–4208, Aug. 2014.

[10] V. A. Zverev, "Modulation method of ultrasonic dispersion measurements (in Russian)," *The Papers of the USSR Academy of Sciences*, vol. 91/4, pp. 791–794, 1953.

[11] V. I. Vorobiov, "Inter-component phase processing of speech signals for their recognition and identification of announcers," in *Proceedings of the 18th session of the Russian Acoustical Society*. Russian Acoustical Society, 2006, vol. 3, pp. 48–51.

[12] V. I. Vorobiov, G. V. Davydov, and Y. V. Shamgin, "Phase relation between fundamental tones and vowel sounds obertones (in Russian)," in *The reports of BSUIR*. Belarusian State University of Informatics and Radioelectronics, 2006, vol. 2/14, pp. 64–68.

[13] V. I. Vorobiov, "Inter-component phase processing of speech signals in time and frequency domains," in *Proceedings of the 19th session of the Russian Acoustical Society*. Russian Acoustical Society, 2007, vol. 3, pp. 46–49.

[14] I Saratxaga, I Hernaez, D Erro, E Navas, and J Sanchez, "Simple representation of signal phase for harmonic speech models," *Electronics Letters*, vol. 45, no. 7, pp. 381–383, 2009.

[15] V. I. Vorobiov, A. G. Davydov, and S. Barysenka, "The interactive bispectrum calculation tool for MATLAB," *Software, available [Sep. 2016] from https://github.com/agdavydov81/bispectrum/*, 2016.

[16] G. Degottex and D. Erro, "A measure of phase randomness for the harmonic model in speech synthesis.," in *INTERSPEECH*, Haizhou Li, Helen M. Meng, Bin Ma, Engsiong Chng, and Lei Xie, Eds. 2014, pp. 1638–1642, ISCA.

[17] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, November 2006.

[18] A. Varga, H.J.M. Steeneken, M. Tomlinson, and D. Jones, "The NOISEX-92 Study on the Effect of Additive Noise on Automatic Speech Recognition," Tech. Rep., DRA Speech Research Unit, 1992.

[19] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs.," in *ICASSP*. 2001, pp. 749–752, IEEE.

[20] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech.," *IEEE Trans. Audio, Speech & Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[21] A. Gaich and P. Mowlaee, "On Speech Intelligibility Estimation of Phase-Aware Signal Processing for Automatic Speech Recognition," *INTERSPEECH*, vol. September, pp. 2553–2557, 2016.

[22] S. Gonzalez and M. Brookes, "PEFAC - A pitch estimation algorithm robust to high levels of noise," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 22, no. 2, pp. 518–530, 2014.

[23] M. Pirolt, J. Stahl, P. Mowlaee, V. I. Vorobiov, S. Y. Barysenka, and A. G. Davydov, "Phase Estimation in Single-channel Speech Enhancement Using Phase Invariance Constraints: supporting webpage with some audio examples. [Online]. Available: `http://www2.spsc.tugraz.at/people/pmowlaee/PQI.html`," Sept. 2016.

[24] J. Fahringer, T. Schrank, J. Stahl, P. Mowlaee, and F. Pernkopf, "Phase-Aware Signal Processing for Automatic Speech Recognition," in *INTERSPEECH*, 2016, pp. 3374–3378.

[25] F. Mayer and P. Mowlaee, "Improved phase reconstruction in single-channel speech separation," in *INTERSPEECH*, 2015, pp. 1795–1799.

Bibliography

[1] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*, John Wiley & Sons, 2006.

[2] A. V. Oppenheim and J. S. Lim, "The Importance of Phase in Signals," *Proceedings of the IEEE*, vol. 69, no. 5, pp. 529–541, May 1981.

[3] D. Wang and L. Lim, "The Unimportance of Phase in Speech Enhancement," *IEEE Trans. on Acoustic Speech and Signal Processing*, vol. 30, no. 4, pp. 679–981, 1982.

[4] H. Pobloth and W.B. Kleijn, "Squared Error as a Measure of Percieved Phase Distrotion," *J. Acoust. Soc. Am.*, vol. 114, no. 2, pp. 1081–1094, 2003.

[5] R. Schlüter and H. Ney, "Using Phase Spectrum Information for Improved Speech Recognition Performance," in *ICASSP*, 2001, pp. 133–136.

[6] T. Kleinschmidt, S. Sridharan, and M. Mason, "The Use of Phase in Complex Spectrum Subtraction for Robust Speech Recognition," *Computer Speech & Language*, vol. 25, no. 3, pp. 585–600, 2011.

[7] P. Mowlaee, R. Saeidi, and R. Martin, "Phase Estimation for Signal Reconstruction in Single-channel Source Separation," in *INTERSPEECH*. 2012, pp. 1548–1551, ISCA.

[8] M. Watanabe and P. Mowlaee, "Iterative Sinusoidal-based Partial Phase Reconstruction in Single-channel Source Separation," in *INTERSPEECH*, Frédéric Bimbot, Christophe Cerisara, Cécile Fougeron, Guillaume Gravier, Lori Lamel, François Pellegrino, and Pascal Perrier, Eds. 2013, pp. 832–836, ISCA.

[9] J. Roux and E. Vincent, "Consistent Wiener Filtering for Audio Source Separation," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 217–220, 2013.

[10] I. Hernáez, I. Saratxaga, J. Sánchez, E. Navas, and I. Luengo, "Use of the Harmonic Phase in Speaker Recognition," in *INTERSPEECH*, 2011, pp. 2757–2760.

[11] P. Rajan, T. Kinnunen, C. Hanilçi, J. Pohjalainen, and P. Alku, "Using Group Delay Functions from All-pole Models for Speaker Recognition," in *INTERSPEECH*, Frédéric Bimbot, Christophe Cerisara, Cécile Fougeron, Guillaume Gravier, Lori Lamel, François Pellegrino, and Pascal Perrier, Eds. 2013, pp. 2489–2493, ISCA.

[12] M. Krawczyk and T. Gerkmann, "STFT Phase Reconstruction in Voiced Speech for an Improved Single-Channel Speech Enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 22, no. 12, pp. 1931–1940, Dec. 2014.

[13] J. Kulmer and P. Mowlaee, "Harmonic Phase Estimation in Single-Channel Speech Enhancement Using Von Mises Distribution and Prior SNR," in *Proc. ICASSP*, Apr. 2015, pp. 5063–5067.

[14] J. Kulmer and P. Mowlaee, "Phase Estimation in Single Channel Speech Enhancement Using Phase Decomposition," *IEEE Signal Process. Lett.*, vol. 22, no. 5, pp. 598–602, May. 2015.

[15] P. Mowlaee and R. Saeidi, "Iterative Closed-Loop Phase-Aware Single-Channel Speech Enhancement," *IEEE Signal Process. Lett.*, vol. 20, no. 12, pp. 1235–1239, Dec. 2013.

[16] T. Gerkmann, "Bayesian Estimation of Clean Speech Spectral Coefficients Given a Priori Knowledge of the Phase," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 62, no. 16, pp. 4199–4208, Aug. 2014.

[17] P. C. Loizou, *Speech Enhancement: Theory and Practice*, CRC press, 2013.

[18] B. Schwerin, *Modulation Domain Based Processing for Speech Enhancement*, Ph.D. thesis, Griffith University, Brisbane, 2012.

[19] M. Nilsson, *Entropy and Speech*, Ph.D. thesis, Royal Institute of Technology (KTH), 2006.

[20] S. F. Boll, "A Spectral Subtraction Algorithm for Suppression of Acoustic Noise in Speech," in *ICASSP*, 1979, pp. 200–203.

[21] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-square Error Log-spectral Amplitude Estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[22] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*, vol. 2, MIT press Cambridge, MA, 1949.

[23] Y. Ephraim and H. L. Van Trees, “A Signal Subspace Approach for Speech Enhancement,” *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 251–266, 1995.

[24] Y. Zhang and Y. Zhao, “Spectral Subtraction on Real and Imaginary Modulation Spectra,” in *ICASSP*, 2011, pp. 4744–4747.

[25] M. Berouti, R Schwartz, and J. Makhoul, “Enhancement of Speech Corrupted by Acoustic Noise,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1979, vol. 4, pp. 208–211.

[26] P. C. Loizou, “Speech Enhancement Based on Perceptually Motivated Bayesian Estimators of the Magnitude Spectrum,” *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 5-2, pp. 857–869, 2005.

[27] I. Cohen, “Relaxed Statistical Model for Speech Enhancement and a Priori SNR Estimation,” *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 5-2, pp. 870–881, 2005.

[28] P. Vary, “Noise Suppression by Spectral Magnitude Estimation Mechanism and Theoretical Limits,” *Signal Processing*, vol. 8, no. 4, pp. 387–400, July 1985.

[29] K. K. Paliwal, K. K. Wójcicki, and B. J. Shannon, “The Importance of Phase in Speech Enhancement,” *Speech Communication*, vol. 53, no. 4, pp. 465–494, 2011.

[30] P. Mowlaee, J. Kulmer, J. Stahl, and F. Mayer, “Phase-Aware Signal Processing in Speech Communication: History, Theory and Practice,” *John Wiley & Sons*, 2016.

[31] M. Krawczyk and T. Gerkmann, “STFT Phase Improvement for Single Channel Speech Enhancement,” *Proceedings of International Workshop on Acoustic Signal Enhancement*, pp. 1–4, 2012.

[32] G. Degottex and D. Erro, “A Measure of Phase Randomness for the Harmonic Model in Speech Synthesis,” in *INTERSPEECH*, Haizhou Li, Helen M. Meng, Bin Ma, Engsiong Chng, and Lei Xie, Eds., 2014, pp. 1638–1642.

[33] P. Mowlaee and J. Kulmer, “Phase Estimation in Single-Channel Speech Enhancement: Limits-Potential,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 23, no. 8, pp. 1283–1294, Aug. 2015.

[34] A. V. Oppenheim and R. W. Schafer, *Discrete-time Signal Processing*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1989.

[35] Y. Agiomyrgiannakis and Y. Stylianou, “Wrapped Gaussian Mixture Models for Modeling and High-Rate Quantization of Phase Data of Speech.,” *IEEE Trans. Audio, Speech & Language Processing*, vol. 17, no. 4, pp. 775–786, 2009.

[36] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice Hall, 1997.

[37] N. I. Fisher, *Statistical Analysis of Circular Data*, Cambridge University Press, Cambridge, 10 1993.

[38] I. Saratxaga, I. Hernaez, D. Erro, E. Navas, and J. Sanchez, “Simple Representation of Signal Phase for Harmonic Speech Models,” *Electronics Letters*, vol. 45, no. 7, pp. 381–383, March 2009.

[39] I. Saratxaga, I. Hernáez, M. Pucher, E. Navas, and I. Sainz, “Perceptual Importance of the Phase Related Information in Speech,” in *INTERSPEECH*, 2012.

[40] G. Degottex and D. Erro, “A Uniform Phase Representation for the Harmonic Model in Speech Synthesis Applications,” *EURASIP J. Audio, Speech and Music Processing*, vol. 2014, pp. 38, 2014.

[41] G. Degottex, A. Roebel, and X. Rodet, “Function of Phase-distortion for Glottal Model Estimation,” *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, , no. June, pp. 4608–4611, 2011.

[42] V. I. Vorobiov, G. V. Davydov, and Y. V. Shamgin, “Phase Relation Between Fundamental Tones and Vowel Sounds Obertones (in Russian),” in *The reports of BSUIR*. Belarusian State University of Informatics and Radioelectronics, 2006, vol. 2/14, pp. 64–68.

[43] V. I. Varabyev and S. Y. Barysenka, “Application of Inter-Component Phase Processing Methods in Non-Stationary Vibration Analysis.,” *Technical Acoustics / Tekhnicheskaya Akustika*, , no. 5, pp. 1–6, 2014.

[44] P. L. De Leon, I. Hernaez, I. Saratxaga, M. Pucher, and J. Yamagishi, “Detection of Synthetic Speech for the Problem of Imposture,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 4844–4847.

[45] J. Sanchez, I. Saratxaga, I. Hernáez, E. Navas, D. Erro, and T. Raitio, "Toward a Universal Synthetic Speech Spoofing Detection Using Phase Information," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 810–820, April 2015.

[46] M. Pirolt, J. Stahl, P. Mowlaee, V. Vorobiov, S. Barysenka, and A. Davydov, " Phase Estimation in Single-Channel Speech Enhancement Using Phase Invariance Constraints," *in Proc. ICASSP*, 2016.

[47] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An Audio-visual Corpus for Speech Perception and Automatic Speech Recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, November 2006.

[48] A. Varga, H.J.M. Steeneken, M. Tomlinson, and D. Jones, "The NOISEX-92 Study on the Effect of Additive Noise on Automatic Speech Recognition," Tech. Rep., DRA Speech Research Unit, 1992.

[49] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual Evaluation of Speech Quality (PESQ)-A New method for Speech Quality Assessment of Telephone Networks and Codecs," in *ICASSP*. 2001, pp. 749–752, IEEE.

[50] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech," *IEEE Trans. Audio, Speech & Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[51] A. Gaich and P. Mowlaee, "On Speech Intelligibility Estimation of Phase-Aware Signal Processing for Automatic Speech Recognition," *INTERSPEECH*, vol. September, pp. 2553–2557, 2016.

[52] S. Gonzalez and M. Brookes, "PEFAC - A Pitch Estimation Algorithm Robust to High Levels of Noise," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 22, no. 2, pp. 518–530, 2014.