Graz University of Technology

Master Thesis

Advances in Phase Aware Speech Enhancement

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

by
Johannes Stahl, 0814103

Supervisors:
Mowlaee, Pejman, PhD

Assessors/Examiners:
Mowlaee, Pejman, PhD
Sontacchi, Alois, Dipl.-Ing. Dr.techn.
Pernkopf, Franz, Assoc.Prof. Dipl.-Ing. Dr.mont.

Graz, September 8, 2015

Abstract

State-of-the-art single-channel speech enhancement methods are typically focussed on compensating the influence of additive noise on magnitude only. Recent studies showed that modifying the noisy spectral phase can yield improvement in quality and intelligibility of the enhanced speech. This thesis presents a joint estimation framework of amplitude and phase based on the maximum a posteriori criterion (MAP). Previous studies on joint MAP estimation of amplitude and phase assumed a uniform distribution of the spectral phase. This resulted in phase-insensitive amplitude estimators together with the noisy phase as the MAP estimate of phase. Contrary, in this thesis a von Mises prior distribution of the spectral phase is considered. By maximising the joint likelihood-function of amplitude and phase, two interdependent estimators are achieved. Due to the non-linear linkage of the amplitude and the phase estimator, an iterative procedure is introduced, in order to achieve an improved complex spectrogram of speech. In a comparative study, the phase-insensitive and the phase-sensitive frameworks are evaluated, showing that an improved phase can contribute positively to speech quality and intelligibility.

Zusammenfassung

Moderne Einkanal-Sprachverbesserungsmethoden versuchen üblicherweise den Einfluss von additivem Rauschen auf die Magnitude von Sprache zu kompensieren. Jüngste Studien zeigten, dass durch Modifikation der verrauschten spektralen Phase die Sprachqualität und Verständlichkeit verbessert werden kann. In dieser Arbeit wird vorgeschlagen, die Amplitude und Phase gemeinsam nach der Maximum-a-posteriori-Methode (MAP) zu schätzen. Vorangegangene Studien zur MAP Schätzung von Amplitude und Phase nahmen eine Gleichverteilung der spektralen Phase an. Dies resultierte in einem von der Phase unabhängigen Amplitudenschätzer und der verrauschten Phase als MAP Schätzer der Phase. Im Gegensatz dazu wird in dieser Arbeit eine von Mises Verteilung der Phase angenommen. Das Ergebnis der Maximierung der gemeinsamen Likelihood-Funktion der Amplitude und der Phase sind zwei voneinander abhängige Schätzer. Aufgrund des nicht-linearen Zusammenhangs der beiden Schätzer wird eine iterative Methode vorgeschlagen um ein verbessertes komplexes Spektrogramm zu erhalten. Eine abschließende Vergleichsstudie zwischen der phaseninsensitiven und der phasensensitiven Methode zeigt, dass eine verbesserte Phase zu höherer Sprachqualität und Verständlichkeit beitragen kann.

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 used sources.

date	(signature)

Contents

1 Introduction

Single-channel speech enhancement addresses the need of compensating the impact of additive noise on the perceived quality and intelligibility of speech. The applications of a speech enhancement system are manifold, including hearing aids, mobile phones and automatic speech recognition. In the last decades, the research has been concentrating on enhancing the spectral amplitude of speech only, neglecting the influence of the spectral phase. This was justified by the findings in [1], where Wang and Lim reported that a more accurate phase does not help to increase the performance in terms of equivalent SNR, when reconstructed with an independently estimated amplitude. However, they admit that a more accurate phase estimate may be important for improving the amplitude estimate.

State-of-the-art amplitude-only enhancement methods are formulated in the STFT domain, motivated by its simplicity and computational efficiency. Using different distributions of the speech spectral magnitude (e.g. Rayleigh, Super-Gaussian), the spectral phase is generally assumed to be uniformly distributed. Thus MAP, ML and MMSE estimates of the phase are equal to the noisy phase [3], whilst the amplitude estimate depends on the assumed prior distribution. Different criteria, such as the MMSE [3] or the MAP [5] have been used to derive the corresponding amplitude estimators. Most of these estimators are a function of the prior and posterior SNRs, which implies the necessity of a noise PSD estimator. The prior SNR can be obtained by the decision directed approach presented in [3], as an alternative to the spectral subtraction method, which yields artefacts called musical noise.

Amplitude-only enhancement is mostly capable of enhancing the perceived quality on the expense of a degraded intelligibility, hence a trade-off between the two is needed. Recently, new studies on the importance of phase for speech quality and intelligibility have been conducted, highlighting the potential of incorporating phase estimation into speech enhancement systems [2, 23, 24].

The iterative method presented in [21] obtained a more accurate phase by synthesizing the signal within iterations, improving the consistency of the enhanced spectra. In [20], Mowalee and Saeidi used a phase-aware amplitude estimator together with a geometry based phase estimator, in order to iteratively enhance the speech.

In [40], the artefacts due to the noise phase components, are met by phase randomization. The randomization only takes place at low SNR regions, where mainly noise components are expected.

More recently, statistical phase estimators have been introduced. The phase can be assumed to consist of deterministic part (due to the fundamental frequency) and a stochastic part, modelled by a von Mises distribution in [9, 12]. In order to obtain a phase estimate, the stochastic

contribution to the phase is of interest which means that the deterministic linear phase has to be removed, yielding the unwrapped phase. Utilizing an unwrapped-phase-estimate for reconstruction and/or for amplitude estimation was reported to improve both, quality and intelligibility of the degraded speech [9, 12, 14, 15, 25].

This thesis examines the joint MAP estimators of amplitude and phase, assuming a von Mises prior distribution of the unwrapped phase. The two estimators are interdependent and therefore an iterative method is proposed to solve the non-linear equations.

The thesis is structured as follows; in chapter 2 previous methods and fundamentals of speech enhancement in general and phase-aware single channel speech enhancement are presented. Chapter 3 presents the derivation of the joint MAP estimators, as well as their implementation in an iterative framework. Two iterative methods are reported, yielding different results, which will be discussed in chapter 4. Chapter 5 concludes on the work and gives an outlook on possible future work.

2 Fundamentals

Typical single-channel speech-enhancement systems as shown in figure 2.1 consist out of three processing blocks referred to as Analysis, Modification and Synthesis (AMS).

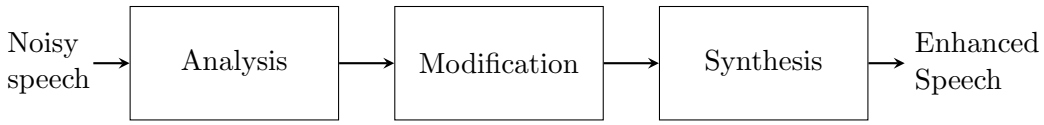

Figure 2.1: Block Diagram of the AMS-System

The focus of this thesis lies on the modification block of the AMS system, the block in which the actual speech-enhancement takes place.

2.1 Signal Model

To unify the notation for the rest of this thesis the following signal model based on the the assumption of additive noise is used:

$$y(n) = s(n) + d(n), \tag{2.1}$$

where $s(n)$ indicates the clean speech, $d(n)$ the additive noise and $y(n)$ the noisy speech, all sampled at time instance $n \cdot T$.
A noisy DFT-coefficient at frame number λ and frequency index k is obtained by segmenting the signal $y(n)$ into (overlapping) frames, multiplying each frame with the window function $w(n)$ and applying a DFT on each frame:

$$Y(\lambda, k) = \sum_{n=0}^{N-1} y(\lambda\Delta + n)w(n)\mathrm{e}^{\frac{-j2\pi nk}{N}}, \tag{2.2}$$

with N as the window length and Δ indicating the frameshift. $Y(\lambda, k)$ is the noisy DFT:

$$Y(\lambda, k) = S(\lambda, k) + N(\lambda, k) \tag{2.3}$$

As every DFT-coefficient consists out of an amplitude and a phase part, it may be rewritten as a complex exponential:

- $Y(\lambda,k) = R(\lambda,k)\mathrm{e}^{j\vartheta(\lambda,k)}$
- $S(\lambda,k) = A(\lambda,k)\mathrm{e}^{j\alpha(\lambda,k)}$
- $N(\lambda,k) = D(\lambda,k)\mathrm{e}^{j\phi(\lambda,k)}$

For the sake of simplicity in notation the frame index λ and the frequency index k will be dropped wherever possible in the further course of this thesis. The definitions of the prior SNR ξ and the posterior SNR ζ used in this thesis are:

$$\text{Prior SNR: } \xi \triangleq \frac{\sigma_s^2}{\sigma_d^2}, \tag{2.4}$$

$$\text{Posterior SNR: } \zeta \triangleq \frac{R^2}{\sigma_d^2}, \tag{2.5}$$

with the noise power spectral density (PSD) σ_d^2 and the speech PSD σ_s^2 defined as follows:

$$\text{Speech PSD: } \sigma_s^2 \triangleq \mathbb{E}\left(D^2\right), \tag{2.6}$$

$$\text{Nosie PSD: } \sigma_d^2 \triangleq \mathbb{E}\left(A^2\right), \tag{2.7}$$

where $\mathbb{E}$ indicates the expectation value operator.
An estimate is indicated by a hat (e.g. $\hat{\xi}$).

2.2 Statistical Models

Since different assumptions on the prior statistics of the spectral amplitude and phase yield different estimation rules, the distributions relevant for this thesis are described in the following.

2.2.1 The Rayleigh Amplitude Model

To analyse the statistical properties of the DFT coefficients of speech, the following two assumptions are commonly made [26]:

1. the window length N in eq. (2.2) approaches infinity: $N \to \infty$
2. N is much longer than the correlation-span of $s(n)$ (for clean speech: 10ms-30ms)

The *central limit theorem* states that the sum of a large number of independent random variables is asymptotically normal-distributed if the variance is finite and positive ([30], p 278). Motivated by this and considering that $s(n)$ is sufficiently random, one can model the real and imaginary parts of S as mutually independent, zero-mean Gaussian random variables for the case $k \notin \{0, \frac{N}{2}\}$. This model also implies that the variance splits equally onto the real and the imaginary part, so that $\mathrm{var}(\mathrm{Re}\{S\}) = \mathrm{var}(\mathrm{Im}\{S\}) = \frac{1}{2}\sigma_S^2$ [26]:

$$p_{\mathrm{Re}\{S\}}(u) = \frac{1}{\sqrt{\pi\sigma_S^2}}\mathrm{e}^{-\frac{u^2}{\sigma_S^2}}, \tag{2.8}$$

$$p_{\mathrm{Im}\{S\}}(v) = \frac{1}{\sqrt{\pi\sigma_S^2}}\mathrm{e}^{-\frac{v^2}{\sigma_S^2}}. \tag{2.9}$$

Since the real and imaginary part are independent from each other, the joint probability density can be achieved by simply multiplying the two distributions:

$$p_{\mathrm{Re}\{S\},\mathrm{Im}\{S\}}(u,v) = p_{\mathrm{Re}\{S\}}(u)p_{\mathrm{Im}\{S\}}(v) = \frac{1}{\pi\sigma_S^2}\mathrm{e}^{-\frac{u^2+v^2}{\sigma_S^2}}. \tag{2.10}$$

The density functions of the amplitude $A = |S|$ and the phase $\alpha = \angle S$ of $S = u + jv$ can be derived by converting eq. (2.10) to polar coordinates ([30], p 203):

$$p(A) = \begin{cases} \frac{2A}{\sigma_S^2}\mathrm{e}^{-\frac{A^2}{\sigma_S^2}} & \text{if } A \in [0,\infty) \\ 0 & \text{otherwise} \end{cases} \tag{2.11}$$

$$p(\alpha) = \begin{cases} \frac{1}{2\pi} & \text{if } \alpha \in [-\pi,\pi) \\ 0 & \text{otherwise} \end{cases} \tag{2.12}$$

It is important to note that eq. (2.11) represents a *Rayleigh* distribution and eq. (2.12) a *Uniform* distribution and that they are jointly independent from each other. Thus, their joint distribution is equal to the product of the two densities:

$$p(A,\alpha) = p(A)\,p(\alpha) = \frac{A}{\pi\sigma_S^2}\mathrm{e}^{-\frac{A^2}{\sigma_S^2}} \tag{2.13}$$

On the supposition of additive Gaussian noise, the PDF of Y, conditioned on A and α can be evaluated ([5], [26]):

$$p(Y|A,\alpha) = \frac{1}{\pi\sigma_d^2}\mathrm{e}^{-\frac{|Y-A\mathrm{e}^{j\alpha}|}{\sigma_d^2}} \tag{2.14}$$

By integrating over α, the conditional PDF of the noisy speech amplitude R is obtained [31]:

$$p(R|A) = \frac{2R}{\sigma_d^2}\mathrm{e}^{-\frac{R^2+A^2}{\sigma_d^2}} I_0\left(\frac{2AR}{\sigma_d^2}\right), \tag{2.15}$$

which corresponds to a *Rice* PDF with $I_0(.)$ being the modified Bessel function of the first kind and the order zero.

2.2.2 The Super-Gaussian Amplitude Model

The prerequisite that the frame length N is much longer than the span of correlation of speech is not met by common frame lengths ranging from 10 ms to 30 ms ([5]). This leads to less Gaussian distributions of real and imaginary parts of the speech DFT coefficients. Therefore several alternatives to the Gaussian speech PDF have been introduced, including Gamma distributed [32] and Laplacian distributed [33] speech priors. In 2005, Lotter and Vary presented a parametric Super-Gaussian density function, which provides the ability to cope with different frame lengths resulting in different distributions of the speech coefficients [5]:

$$p(A) = \frac{\mu^{\nu+1}}{\Gamma(\nu+1)}\frac{A^\nu}{\sigma_S^{\nu+1}}\mathrm{e}^{-\mu\frac{A}{\sigma_S}}, \tag{2.16}$$

where μ and ν are the shape parameters of the distribution and $\Gamma(.)$ denotes the Gamma function. In [5], the authors obtained the set (μ, ν) via fitting the analytic parametric distribution of eq. (2.16) to an empirical distribution of the speech coefficients. The parameters were chosen by

means of the Kullback-Leibler divergence between the two distributions resulting in $\nu = 0.126$ and $\mu = 1.74$, considering a frame length of 32 ms. Especially, the high density for low spectral magnitude regions are well modelled by this distribution (see figure 2.2).

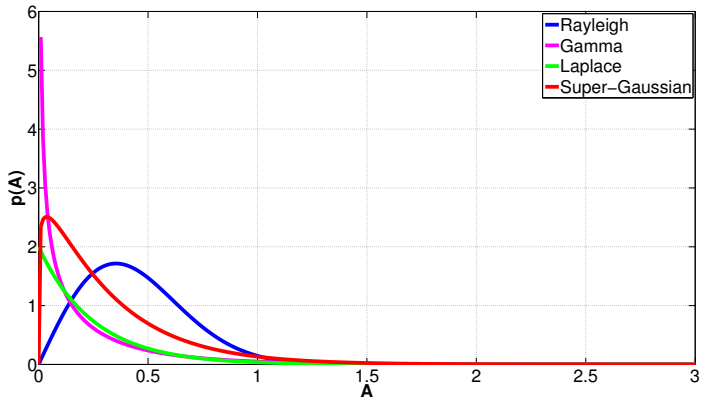

Figure 2.2: Different PDFs of the amplitude of the speech DFT coefficients

2.2.3 Phase Model

The above discussed statistical models assume a uniform prior distribution of the phase. In the last years, this hypothesis has been strongly doubted, e.g. [9, 11, 25], and a more general model has been used to describe the statistical properties of the unwrapped phase α. The *von Mises* distribution is a circular distribution and is characterized by the following equation:

$$p(\alpha) = \frac{\mathrm{e}^{\kappa \cos(\alpha - \alpha_\mu)}}{2\pi I_0(\kappa)} \tag{2.17}$$

Here, α_μ denotes the mean and κ the concentration parameter. The concentration parameter illustrates, how strong the distribution is concentrated around its mean; the higher κ is, the narrower the distribution $p(\alpha)$ gets. It is important to note that the *Uniform* distribution is a special case of the *von Mises* distribution, as for $\kappa \to 0$, $\mathrm{e}^{\kappa \cos(\alpha - \alpha_\mu)} \to 1$ and $I_0(\kappa) \to 1$, resulting in $p(\alpha) = \frac{1}{2\pi}$. The concentration parameter is assumed to be large (expressing high certainty) for highly voiced speech and rather small for unvoiced speech or noise dominated regions. How to obtain the parameters of the κ and α_μ is explained in section 2.3. Figure 2.3 clarifies the influence of κ onto the *von Mises* distribution.
The *von Mises* distribution is a special case of the more general *von Mises-Fisher* distribution [28], modelling the distribution of a p-dimensional vector $\boldsymbol{x}$ on a $p-1$-dimensional sphere:

$$p(\boldsymbol{x}) = \frac{\kappa^{\frac{p}{2}-1} \mathrm{e}^{\kappa \boldsymbol{\mu}^T \boldsymbol{x}}}{(2\pi)^{\frac{p}{2}} I_{\frac{p}{2}-1}(\kappa)}. \tag{2.18}$$

To achieve the *von Mises* distribution in eq. (2.17) p is set to $p = 2$ and $\boldsymbol{x} = \begin{pmatrix} \cos\alpha \\ \sin\alpha \end{pmatrix}$ and $\boldsymbol{\mu} = \begin{pmatrix} \cos\alpha_\mu \\ \sin\alpha_\mu \end{pmatrix}$. The special case $p = 3$ is called *Fisher* distribution and is commonly used in order to analyse spherical data sets.

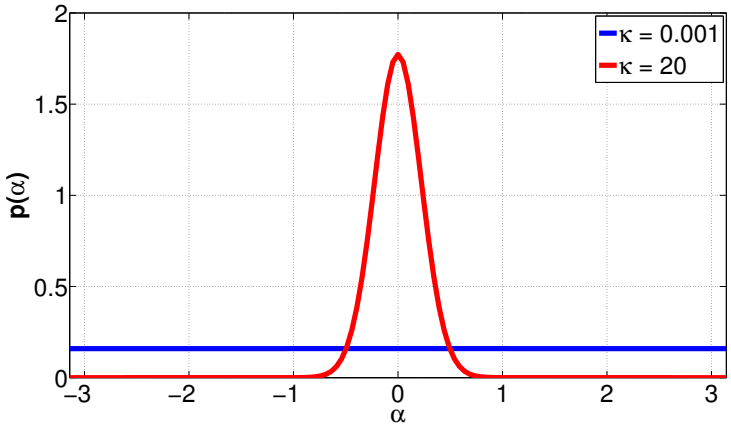

Figure 2.3: Two zero-mean von Mises probability-density functions with concentration parameters $\kappa = 0.001$ and $\kappa = 20$

2.3 Circular Statistics

In order to obtain the *von Mises* parameters κ and α_μ, the DFT coefficients can be analysed with respect to their circular behaviour. A complex exponential with an unit amplitude $z(n) = \mathrm{e}^{\jmath\theta(n)}$ can be analysed according to the sample mean vector [29]:

$$\bar{\rho} = \frac{1}{N}\sum_{n=1}^{N} z(n) \tag{2.19}$$

The absolute value of the sample mean vector is given as [29]:

$$\bar{Z} = |\bar{\rho}| \tag{2.20}$$

Similar, the sample mean for the angle is defined [29]:

$$\bar{\mu} = \angle\bar{\rho} \tag{2.21}$$

The circular variance σ^2_{circ} serves as a measure of directional spread, defined as follows [29]:

$$\sigma^2_{circ} = 1 - \bar{Z} = 1 - \frac{I_1(\kappa)}{I_0(\kappa)} \tag{2.22}$$

The concentration parameter κ gives information about how concentrated the complex exponential is around the mean angle. The circular variance is bounded to the interval [0 1] and as it can be directly estimated from the observed data, κ can in turn be estimated by approximating the inverse function of eq. (2.22).

2.4 Amplitude only Single-Channel Speech Enhancement

Conventional amplitude only speech enhancement methods assume that the spectral phase follows a *Uniform* distribution, hence the MMSE [3], ML and MAP [5, 8] estimates of the phase are equal to the noisy phase. Although this thesis relies on the assumption that the unwrapped spectral phase follows a *von Mises* distribution rather than a *Uniform* distribution, certain concepts of amplitude only estimation are used for the implementation of the proposed estimator in Chapter 3.

2.4.1 Short Time Spectral Amplitude Estimator

Ephraim and Malah presented a Minimum Mean Square Error (MMSE) Short-Time Spectral Amplitude (STSA) Estimator in [3]. Using the assumptions made in section 2.2.1, their model yields the MMSE STSA estimator of form:

$$\hat{A} = \Gamma(1.5)\frac{\sqrt{v}}{\zeta}\mathrm{e}^{-\frac{v}{2}}\left((1+v)I_0\left(\frac{v}{2}\right) + vI_1\left(\frac{v}{2}\right)\right)R, \tag{2.23}$$

with

$$v = \frac{\xi}{1+\xi}\zeta. \tag{2.24}$$

2.4.2 Prior SNR Estimation

Eq.(2.23) is clearly a function of the prior SNR ξ and posterior SNR ζ, hence, they need to be estimated. Given a noise PSD estimate, the posterior SNR is obtained as follows:

$$\hat{\zeta}(\lambda) = \frac{R^2(\lambda)}{\hat{\sigma}_d^2(\lambda)}. \tag{2.25}$$

The prior SNR, as defined in eq. (2.4), depends on the speech PSD, which is not given. A simple way to estimate the prior SNR is the method of power spectral subtraction:

$$\hat{\xi}(\lambda) = \max\left[\hat{\zeta}(\lambda) - 1, 0\right]. \tag{2.26}$$

There have been proposed several variants of the spectral subtraction method in the literature, all suffering form artefacts called musical noise. This is why Ephraim and Malah proposed to estimate the a priori SNR with a *decision-directed* approach, deduced from weighing the power spectral subtraction (the ML estimate of the prior SNR) against an estimate of the speech PSD obtained by incorporating the previous estimated speech coefficient:

$$\hat{\xi}(\lambda) = \alpha\frac{\hat{A}^2(\lambda-1)}{\sigma_d^2(\lambda-1)} + (1-\alpha)\max\left[\hat{\zeta}(\lambda) - 1, 0\right], \tag{2.27}$$

where α is the smoothing parameter in the interval $[0, 1]$, chosen heuristically with $\alpha = 0.97$. The decision-directed approach was reported to help suppressing the musical noise efficiently [3].
Since the decision-directed approach is sensitive to amplitude onsets it is prone to track noise bursts that are not identified to be noise by the noise PSD estimator. To cope with this problem Breithaupt et al. proposed an approach based on cepstro-temporal smoothing in [16]. The prior SNR is smoothed recursively in the cepstral domain; higher cepstral coefficients (except those, that represent the fundamental frequency f_0) are smoothed stronger than lower coefficients, representing the speech envelope. To this end a fundamental frequency estimation has to be carried out, as the cepstral bins containing f_0 information should stay untouched. The selective amount of smoothing is set by a smoothing parameter similar to α in eq. (2.27). Estimating the prior SNR in the cepstral domain has the advantage that it takes into account prior knowledge about the speech production. From a fundamental frequency estimate the smoothing can be focussed onto regions where speech is not likely to be present. The authors of [16] reported a clearer speech signal and higher robustness to non-stationary noise types compared to the recursive smoothing in frequency domain.

2.5 Phase Aware Speech Enhancement

Conventional speech enhancement methods only provide an estimate of the spectral amplitude. Thus, the enhanced amplitude is used together with the noisy phase for reconstruction (figure 2.4). In order to obtain a better estimate of the complex speech coefficients, the information carried by the phase has been incorporated in recent publications. A distinction between two ways of doing this can be made:

1. Estimating the phase and applying it at the synthesis stage, replacing the noisy phase (see section 2.5.1), scheme in figure 2.5
2. Using a phase-estimate as additional information for the amplitude estimation or joint estimation of amplitude and phase (see section 2.5.2)

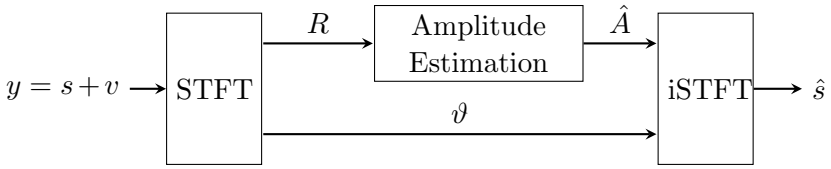

Figure 2.4: Conventional speech enhancement, illustrated by a block diagram

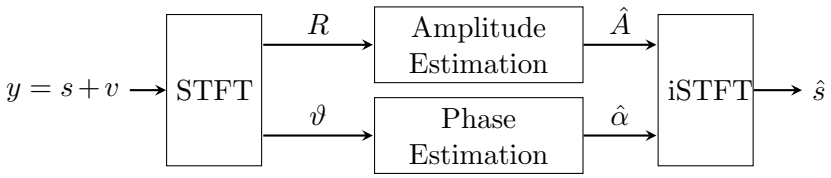

Figure 2.5: Amplitude enhancement together with phase enhancement. The independent estimates are used together for synthesis. The phase estimator can incorporate knowledge about the amplitude too (e.g. section 2.5.1).

2.5.1 Phase Estimation in Single Channel Speech Enhancement

In 1984, Griffin and Lim showed that the modified STFT of a signal does not fit to a time domain signal, if the phase is not taken into account. Thus, they proposed an iterative procedure to estimate the phase given the speech spectral amplitude, based on the inconsistency criterion eq. (2.63). While the phase is updated within iterations, the amplitude is constrained to stay constant over iterations [4]. If the clean speech amplitude spectrogram is perfectly known, this approach yields perceptually good results. In real world scenarios, the amplitude is estimated, hence the performance is limited [14]. Nevertheless, the method presented in [4] emphasizes that the limits of amplitude-only enhancement can be pushed by incorporating knowledge about the spectral phase.

This section provides an overview of phase estimation methods considered to be important for the proposed method in Chapter 3.

STFT Phase reconstruction in Voiced Speech for an Improved Single-Channel Speech Enhancement (STFTPI)

The STFTPI method [14, 15] is used as prior information on the phase for the phase-aware amplitude estimators described in section 2.5.2 and 2.5.2. The basic principle of the STFTPI

method relies on the assumption that the speech follows a harmonic structure in voiced frames . The authors proposed a harmonic model for the speech signal in these voiced frames:

$$s(n) = \sum_{h=0}^{H} 2A_h \cos\left(n\omega_h + \phi_h\right), \tag{2.28}$$

with H denoting the number of harmonics (i.e. the order of the model), ϕ_h the phase offset and ω_h the normalized angular frequency at harmonic index h:

$$\omega_h = 2\pi\frac{f_h}{f_s} = 2\pi\frac{(h+1)f_0}{f_s}. \tag{2.29}$$

The harmonic model is combined with the common STFT-framework by assigning $\omega(k)$ (the angular frequency at STFT bin k) to ω_h of the nearest harmonic h. This is justified by the assumption that in voiced speech, the complex exponential at h dominates the adjacent STFT bins k. The described assignment is expressed by:

$$\omega_h(k) = \arg\min_{\omega_h} \left|\omega_h - \omega(k)\right|. \tag{2.30}$$

For reconstruction purposes the authors presented two approaches:

1. phase reconstruction along time

2. phase reconstruction along frequency

The temporal derivative of the phase yields the instantaneous frequency IF. An approximation of the IF is obtained by the phase difference of two consecutive frames. The reconstruction along time makes use of the assumption that the fundamental frequency changes slowly over time and does not change within a frame's length; the recursive reconstruction of the phase along time is given by:

$$\phi(\lambda, k) = princ\{\phi(\lambda - 1, k) + \omega_h(\lambda, k)L\}, \tag{2.31}$$

with L and $princ$ indicating the frame shift and the principal argument of the phase, respectively.
Alternatively, the reconstruction along frequency is achieved by using the prerequisite of the dominance of the harmonic h in the adjacent frequency bands. This along with the assumption of a linear phase window is used to reconstruct the phase along frequency.

Since the method proposed in [14] does not take into account regions lying between the harmonics, important information is neglected. As the inter-harmonic frequency bands do provide important components for sounds unlike vowels such as fricatives and plosives, the harmonic model and the dominance-assumption do not fit well to real speech either. As a consequence of the over-harmonisation produced by these shortcomings, the resulting enhanced speech was reported to show buzzyness [11, 12]. An additional source of unnaturalness is the model-order H, not explicitly estimated but limited to fit the frequency range up to 4 (kHz):

$$H f_0 \leq 4(\text{kHz}) \tag{2.32}$$

This of course does not reflect the actual order of the harmonic model and is only a coarse estimate, relying on the fact that higher order harmonics appear being less prominent.

Phase Estimation in Single Channel Speech Enhancement Using Phase Decomposition

In 2015, Kulmer and Mowlaee presented a phase estimation method relying on the decomposition of the harmonic phase at time instant $t(l)$, based on the source-filter concept of speech [25]:

$$\psi(h,l) = \underbrace{h\sum_{l'=0}^{l}\omega_0(l')\left(t(l')-t(l'-1)\right)}_{\text{linear phase}} + \underbrace{\underbrace{\angle V(h,l)}_{\text{minimum phase}} + \underbrace{\psi_d(h,l)}_{\text{dispersion phase}}}_{\Psi(h,l)}, \tag{2.33}$$

where $\Psi(h,l)$ denotes the unwrapped phase, consisting of the minimum phase $\angle V(h,l)$ and the dispersion phase $\psi_d(h,l)$. The fundamental frequency of the excitation is modelled by the linear phase, whereas the minimum phase captures the phase of the vocal tract filter. The dispersion phase component relates to the stochastic part of the excitation signal. The phase wrapping, which is a major problem in phase estimation because due to the linear phase part, the phase is wrapped along time. The goal of the phase decomposition method is to decouple the instantaneous phase into linear and unwrapped parts. The unwrapped phase is then smoothed in order to reduce the impact of noise, which distorts the unwrapped phase. This requires an estimate of the linear phase, which is achieved by the underlying fundamental frequency denoted by ω_0 in eq. (2.33). When subtracting the linear phase estimate, the unwrapped phase $\Psi(h,l)$ is obtained. The temporal smoothing is implemented by the circular mean on the frames lying within a 20 ms time-span (see eq. (2.21)). For reconstruction, the temporally smoothed phase $\hat{\Psi}(h,l)$ is used:

$$\boxed{\hat{\psi}(h,l) = h\sum_{l'=0}^{l}\hat{\omega}_0(l')\left(t(l')-t(l'-1)\right) + \hat{\Psi}(h,l).} \tag{2.34}$$

In order to transform the harmonic phase into a STFT phase, the phase of the DFT bins lying within the main lobe width of the analysis window is set to $\hat{\psi}(h,l)$, since they are statistically dependent. In contrast to the STFTPI method, where the phase of every DFT bin is assigned to the nearest harmonic phase, the phase of bins lying outside the main lobe width stay untouched. The time domain signal is then reconstructed by applying the inverse DFT and the overlap-and-add procedure. The authors of [25] chose Blackman as the analysis window. It shows a high side-lobe rejection, minimizing the influence of noise components lying near the harmonics. The performance analysis showed a consistent quality improvement (predicted by PESQ) and even improved intelligibility (predicted by STOI) for the f_0 known scenario, where the f_0 was obtained from the clean reference speech signal.

MAP Phase Estimator

Based on the assumption of a *von Mises* distribution of the unwrapped phase component $\Psi(h,l)$ the authors of [9] proposed a MAP estimator of the harmonic phase. The signal model is given as follows:

$$y(n) = \sum_{h=1}^{H_l} A(h,n)\cos\left(h\omega_0(n)n + \Psi(h,n)\right) + d(n), \tag{2.35}$$

The MAP estimate of $\Psi(h,l)$ at frame index l is given by:

$$\hat{\Psi}_{MAP}(h,l) = \underset{\Psi(h,l)}{\arg\max}\frac{p\left(\mathbf{y}(l)|\Psi\right)p(\Psi)}{p\left(\mathbf{y}(l)\right)}, \tag{2.36}$$

with $p(\Psi)$ being the distribution of the unwrapped phase as defined in eq. (2.17), $\mathbf{y}(l)$ being the noisy observation vector at frame index l and $p\left(\mathbf{y}(l)|\Psi\right)$ is:

$$p\left(\mathbf{y}(l)|\Psi)\right) = \frac{1}{\sqrt{2\pi\sigma_d^2}^N} \mathrm{e}^{-\frac{1}{\sigma_d^2}\sum_{n=0}^{N-1}\left(y(n)-A\cos(h\omega_0 n+\Psi)\right)^2}. \tag{2.37}$$

Solving for the given distributions yields the MAP phase estimator at harmonics [9]:

$$\boxed{\hat{\Psi}_{MAP} = \tan^{-1}\left(\frac{-\frac{2A}{\sigma_d^2}\sum_{n=0}^{N-1} y(n)\sin\left(h\omega_0 n\right) + \kappa\sin\left(\alpha_\mu\right)}{\frac{2A}{\sigma_d^2}\sum_{n=0}^{N-1} y(n)\cos\left(h\omega_0 n\right) + \kappa\cos\left(\alpha_\mu\right)}\right),} \tag{2.38}$$

where N indicates the framelength and $y(n)$ is the nth entry of $\mathbf{y}(l)$. The parameters of the *von Mises* are estimated by applying the decomposition principle [25] onto the noisy observation in order to obtain an unwrapped phase estimate. The true parameters κ and α_μ of the distribution are approximated from the noisy observation according to eq. (2.22) and eq. (2.21). The mean value α_μ is then obtained by adding the linear phase. The STFT representation needed for synthesis is again obtained by the principle used for the phase decomposition method. Whereas in terms of predicted quality in stationary noise the MAP phase estimator is only slightly better performing than the STFTPI method, non-stationary noise types and evaluation in terms of intelligibility show a big improvement [9]. Especially the lower sensitivity to the f_0 estimation errors in the signal plus noise scenario is important to note, as both, the phase decomposition principle, as well as, the STFTPI method are very sensitive to such errors.

The MAP phase estimator already takes into account amplitude information and therefore is different to the methods that fit to the scheme in figure 2.5.

2.5.2 Phase-Aware Amplitude Estimation

The aforementioned methods use the obtained phase estimate at the reconstruction stage only. In the following, approaches to incorporate a phase estimate for amplitude estimation are discussed.

MMSE-optimal spectral amplitude estimation given the STFT-phase

In 2013, Gerkmann and Krawczyk proposed a phase-aware amplitude estimator optimal in the MMSE sense, where they assumed that phase is given [13]. This is contrary to the uniform assumption on phase, made in previous publications. The problem of estimating the speech coefficient's amplitude given the phase is formulated as follows:

$$\hat{A^\beta} = \mathbb{E}\left(A^\beta|R,\vartheta,\alpha\right) = \int_{-\infty}^{\infty} A^\beta p\left(A|R,\vartheta,\alpha\right)\mathrm{d}A, \tag{2.39}$$

where β is a compression parameter that represents a generalization of the logarithmic amplitude compression [17]. The estimation of (logarithmically) compressed spectral amplitudes has been reported to be perceptually beneficial in [41]. To obtain the posterior $p\left(A|R,\vartheta,\alpha\right)$, Bayes' Theorem is used, defined as:

$$p(B|C) = \frac{p(C|B)p(B)}{p(C)}. \tag{2.40}$$

Further, the assumption that the clean speech amplitude is independent of the clean speech phase (which means $p(\alpha)$ can be factorized out) gives:

$$\hat{A^{\beta}} = \frac{\int_{-\infty}^{\infty} A^{\beta} p\left(R, \vartheta | A, \alpha\right) p\left(A\right) \mathrm{d}A}{\int_{-\infty}^{\infty} p\left(R, \vartheta | A, \alpha\right) p\left(A\right) \mathrm{d}A} \tag{2.41}$$

The assumption that the real and imaginary parts of the complex noise coefficients are independent leads to the same conditional PDF of Y as in section 2.2.1. The speech-coefficient's amplitude distribution is assumed to follow a χ-distribution with shape parameter μ:

$$p(A) = \frac{2}{\Gamma(\mu)} \left(\frac{\mu}{\sigma_s^2}\right)^{\mu} A^{2\mu-1} \mathrm{e}^{-\frac{\mu}{\sigma_s^2} A^2}. \tag{2.42}$$

For the case that $\mu = 1$, eq. (2.42) equals the Rayleigh distribution in eq. (2.11). By inserting eq. (2.42) and eq. (2.14) into eq. (2.39) the solution is obtained:

$$\boxed{\hat{A} = \left(\mathbb{E}\left(A^{\beta} | R, \vartheta, \alpha\right)\right) = \sqrt{\frac{1}{2}\frac{\xi}{\mu+\xi}\sigma_d^2} \left(\frac{\Gamma(2\mu)+\beta}{\Gamma(2\mu)} \frac{D_{-(2\mu+\beta)}(\nu)}{D_{-(2\mu)}(\nu)}\right)^{\frac{1}{\beta}},} \tag{2.43}$$

where $D_{\cdot}(\nu)$ denotes the parabolic cylinder function and ν is given as follows:

$$\nu = -\sqrt{2\frac{\xi}{\mu+\xi}\zeta} \cos\left(\vartheta - \alpha\right). \tag{2.44}$$

As can be seen, the parameter ν contains the phase deviation $\Delta\phi = \vartheta - \alpha$ which is incorporating the phase information into the amplitude estimation. The lower the SNR, the larger the phase deviation gets, thus more attenuation is applied. To deal with the fact that the clean phase is unknown in practical scenarios, the authors proposed to use the phase estimation presented in [14] in order to evaluate ν. In [10] the same authors use the voicing probability P_{H_v} achieved by the f_0-estimator PEFAC [19] to weight the phase-aware [13] and phase unaware [53] gainfunctions (denoted by the subscripts):

$$\hat{A}_{[10]} = P_{H_v} \hat{A}_{[13]} + (1 - P_{H_v}) \hat{A}_{[53]} \tag{2.45}$$

This weighing helps to avoid employing an unreliable phase estimate at unvoiced frames while profiting from the additional phase information at voiced frames. For the described estimation scheme, the parameters of the speech amplitude's distribution were chosen with $\beta = \mu = 0.5$. An improvement in PESQ has been reported, especially for the case of voiced speech.

Bayesian Estimation of Clean Speech Spectral Coefficients Given a Priori Knowledge of the Phase

Unlike the method described above in [12], Gerkmann assumed the phase information obtained from [15] to be uncertain. Hence he proposed a joint amplitude and phase estimator yielding the CUP estimator (Complex spectral speech coefficients given Uncertain Phase information). The CUP estimator is derived by solving

$$\hat{S}^{(\beta)} = \mathbb{E}\left(A^{\beta} \mathrm{e}^{j\alpha} | Y, \tilde{\alpha}\right) = \int_0^{\infty} \int_0^{2\pi} A^{\beta} \mathrm{e}^{j\alpha} p\left(A, \alpha | Y, \tilde{\alpha}\right) \mathrm{d}\alpha \mathrm{d}A, \tag{2.46}$$

where $\tilde{\alpha}$ denotes the uncertain prior phase information, which is obtained by the STFTPI algorithm proposed in [14, 15]. The joint posterior $p\left(A, \alpha | Y, \tilde{\alpha}\right)$ is again obtained by using

Bayes' theorem:

$$p\left(A,\alpha|Y,\tilde{\alpha}\right)=\frac{p\left(Y|\tilde{\alpha},A,\alpha\right)p\left(\tilde{\alpha},A,\alpha\right)}{\int_0^{2\pi}\int_0^{\infty}p\left(Y|\tilde{\alpha},A,\alpha\right)p\left(\tilde{\alpha},A,\alpha\right)\mathrm{d}A\mathrm{d}\alpha} \tag{2.47}$$

To solve eq. (2.47) two assumptions are made:
If the clean phase α is given, $\tilde{\alpha}$ does not contain further information, so that:

$$p\left(Y|A,\alpha,\tilde{\alpha}\right)=p\left(Y|A,\alpha\right), \tag{2.48}$$

which is expressed by eq. (2.14) since the noise coefficients are assumed to be complex gaussian.
The second assumption is that speech amplitude and phase are independent of each other, which helps to further simplify eq. (2.47):

$$p\left(\tilde{\alpha},A,\alpha\right)=p(A)p(\alpha,\tilde{\alpha})=p(A)p(\tilde{\alpha})p(\alpha|\tilde{\alpha}). \tag{2.49}$$

The distribution of α around the prior information $\tilde{\alpha}$ $p(\alpha|\tilde{\alpha})$ can be modelled by a *von Mises* distribution with κ as a measure for the certainty of the prior information $\tilde{\alpha}$. The same χ-distribution as in [13] for the speech PDF is assumed. Given these models, eq. (2.47) can be solved yielding the desired estimator of complex spectral speech coefficients given uncertain phase information:

$$\boxed{\hat{S}^{(\beta)}=\left(\sqrt{\frac{1}{2}\frac{\xi}{\mu+\xi}\sigma_d^2}\right)^{\beta}\frac{\Gamma\left(2\mu+\beta\right)}{\Gamma\left(2\mu\right)}\frac{\int_0^{2\pi}\mathrm{e}^{j\alpha}\mathrm{e}^{\frac{\nu^2}{4}}D_{(-2\mu-\beta)}(\nu)p(\alpha|\tilde{\alpha})\mathrm{d}\alpha}{\int_0^{2\pi}\mathrm{e}^{\frac{\nu^2}{4}}D_{(-2\mu)}(\nu)p(\alpha|\tilde{\alpha})\mathrm{d}\alpha},} \tag{2.50}$$

where ν is given as in eq. (2.44).

In order to compensate the amplitude compression β the final speech coefficients are obtained by:

$$\hat{S}=\frac{\left|\hat{S}^{(\beta)}\right|^{\frac{1}{\beta}}}{\left|\hat{S}^{(\beta)}\right|}\hat{S}^{(\beta)}. \tag{2.51}$$

Since solving the integrals in eq. (2.50) is computationally very expensive, the author of [12] implemented the CUP estimator by incorporating a look up table with the four dimensions prior SNR ξ, posterior SNR ζ, concentration parameter κ and phase difference $\Delta\phi=\alpha-\vartheta$. The entries of the look-up table are obtained by numerical integration, which provides satisfactory precision since the integration intervals are bounded to $[0,2\pi]$. The chosen resolution for the look up table has not been reported.

The experiment setup includes the phase reconstruction along frequency as described in [14] for obtaining the prior phase information $\tilde{\alpha}$, relying on the fundamental frequency estimate of the PEFAC algorithm [19]. The degree of certainty of the prior information is given by κ, which controls the influence of $\tilde{\alpha}$. Thus, κ needs to be estimated. Gerkmann suggested to use the voicing probability $P_{H_v}(\lambda)$, given by the PEFAC algorithm:

$$\kappa(\lambda,k)=\begin{cases}4P_{H_v}(\lambda) & \text{if } \frac{kf_s}{N}<4000Hz\\ 2P_{H_v}(\lambda) & \text{if } \frac{kf_s}{N}\geq 4000Hz\end{cases} \tag{2.52}$$

The noise PSD σ_d^2 is estimated according to [42]. The obtained estimate is then plugged into

$\zeta = \frac{R^2}{\sigma_d^2}$, denoting the posterior SNR. The prior SNR estimation is implemented in a decision directed way [3]. In contrast to the authors in [3], the smoothing factor α is chosen with 0.96 instead of 0.97, in order to have less speech distortions. While this setup leads to PESQ improvement only, replacing the decision directed approach by the cepstro-temporal smoothing based prior SNR estimation approach presented in [16] leads to a joint improvement in terms of quality and intelligibility, predicted by STOI.

2.6 Iterative Speech Enhancement

Since in Chapter 3 the derived estimation rules for amplitude and phase will be combined in an iterative procedure, the following section presents previous iterative methods, considered to be important for the proposed iterative algorithm.

2.6.1 Iterative Wiener Filtering

In [34], Oppenheim and Lim proposed an Iterative Wiener filter (IWF) solution to jointly optimize the Linear predictive coefficients (LPC) parameters of a noise-degraded speech signal, the gain g and noise-free speech estimate s_0 in a maximum a posteriori (MAP) sense. The authors use the LP coefficients to obtain the speech PSD needed for the Wiener Filter:

$$H(\omega) = \frac{\sigma_S^2(\omega)}{\sigma_S^2(\omega) + \sigma_d^2(\omega)}, \tag{2.53}$$

with the speech PSD $\sigma_S^2(\omega)$, given by:

$$\boxed{\sigma_S^2(\omega) = \frac{g^2}{|1 - \sum_{k=1}^{K} a_k \mathrm{e}^{-jk\omega}|^2},} \tag{2.54}$$

where g is the excitation-gain and a_k are the LP coefficients from the all-pole-model of speech of order K:

$$s(n) = \sum_{k=1}^{K} a_k s(n-k) + g w(n). \tag{2.55}$$

Since the performance of the LPC procedures decreases in the presence of noise [35], the obtained coefficients a_k are inaccurate if obtained from the noisy observation. In order to refine the coefficient estimates, Oppenheim and Lim proposed an iterative procedure, where a_k is estimated from the filtered speech yielding a new speech PSD estimate. Thus, the filter in eq. (2.53) is updated accordingly for the next iteration. If the conditional probability density function of a_k is uni-modal, the method always converges to a global maximum [34]. Otherwise, the convergence strongly depends on the initial values of a_k. Besides this, the method has two major drawbacks [36]:

1. the computational load is increased by iterations

2. a convergence criterion was not found, so that a heuristic maximum number of iterations had to be employed

The Block diagram in figure 2.6 illustrates the basic principle of the Iterative Wiener filter, with i being the iteration index.

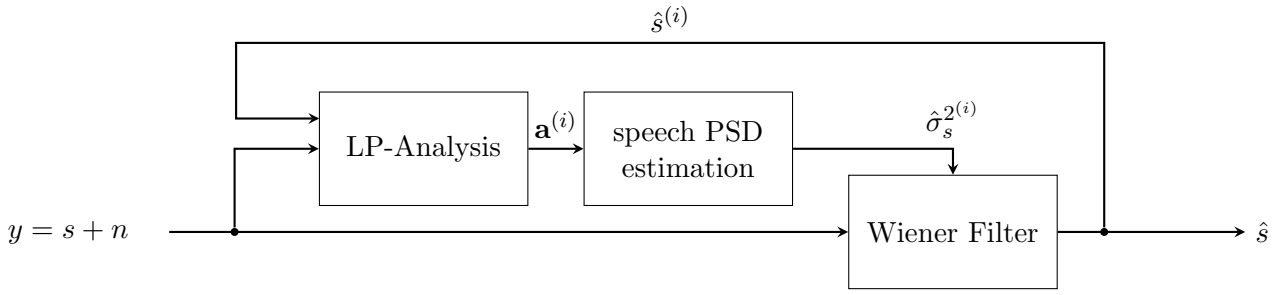

Figure 2.6: Block Diagram of the Iterative Wiener Filter, $\boldsymbol{a}^{(i)} = [a_1^{(i)} a_2^{(i)} \dots a_K^{(i)}]^T$

2.6.2 Constrained Iterative Wiener Filtering

The Lim and Oppenheim IWF solution suffers from formant drifts and pole-jitter over iterations. To circumvent this undesired behaviour, Hansen and Clements, in [36], extended the IWF idea by spectral constraints in order to obtain a more speech-like output. They reported consistent improvement over the IWF for different speech classes in terms of both, perceived speech quality and speech recognition accuracy for additive white and cockpit coloured noise at different signal-to-noise ratios. The spectral constraints Hansen and Clements introduced can be divided into two groups:

1. constraints across time (inter-frame)
2. constraints across iterations (intra-frame)

For the sake of computational efficiency, the constraints are applied in the LSP (Line Spectral Pairs) domain. The LSP representation originates in rearranging the LPC polynomial [26]:

$$A(z) = 1 - \sum_{k=1}^{K} a_k z^{-k} = 0.5\left(P(z) + Q(z)\right), \tag{2.56}$$

where $z = \mathrm{e}^{j\omega}$ and

$$P(z) = A(z) + z^{-(K+1)})A\left(z^{-1}\right), \tag{2.57}$$

$$Q(z) = A(z) - z^{-(K+1)}A\left(z^{-1}\right), \tag{2.58}$$

with $P(z)$ and $Q(z)$ characterizing the mirror and the anti-mirror polynomials, respectively. As the roots of the polynomials $P(z)$ and $Q(z)$ occur in complex conjugated pairs, half as many roots have to be computed as directly calculating the roots of $A(z)$. If the roots of $P(z)$ and $Q(z)$ interleave, $A(z)$ is considered to be stable, which is very easy to check. The bandwidth of a resonance is reflected in the distance of a pole of $P(z)$ to the closest root of $Q(z)$. Two examples of constraints based on the LSP representation are described in the following:
(i) An inter-frame-constraint based on median-smoothing over 5 frames and flooring the aforementioned separation of neighbouring roots by a minimum distance, keeps the algorithm from resulting in unreasonable bandwidths for speech LPC poles. (ii) To ensure smooth transitions of the formants across iterations, the same distance-parameter is used as an intra-frame constraint, indicating if a root corresponds to a formant; in this case the tracks of the formant across iterations would be smoothed accordingly. Hansen and Clements were not able to find a blind termination criterion for the iterative procedure, as the optimal number of iterations (with respect to predicted speech quality) varied for different noise and SNR scenarios. Hence,

they chose the iteration number empirically.
The scheme in figure 2.7 illustrates how the constraints are used to adapt the shadow filter $H(\omega)$.

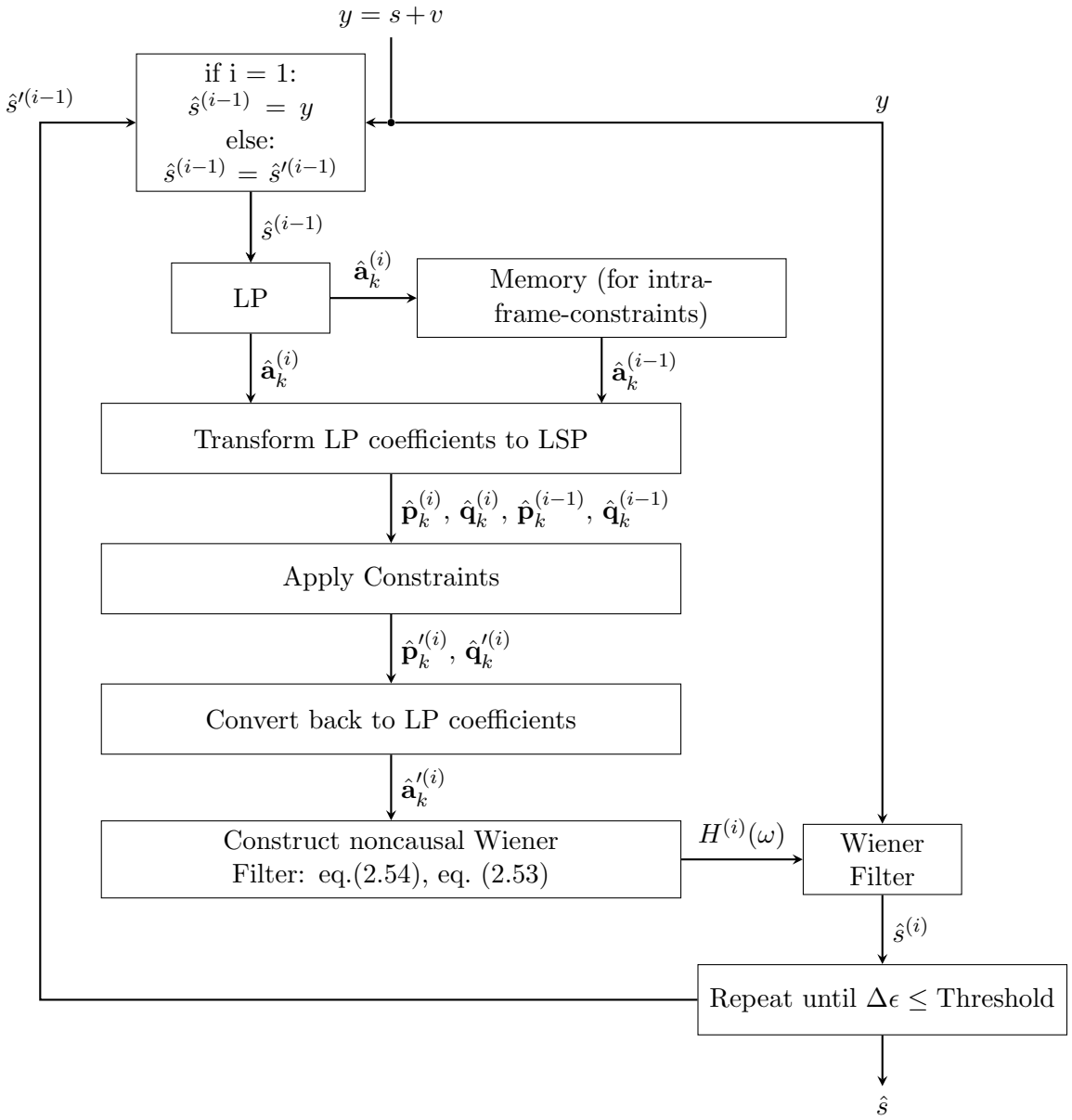

Figure 2.7: Block Diagram of the Constrained Iterative Wiener Filter. With LP coefficient vector $\boldsymbol{a}^{(i)} = [a_1^{(i)} a_2^{(i)} \dots a_K^{(i)}]^T$ and the coefficient vectors of the Line Spectrum Pairs $P(z)$ and $Q(z)$: $\boldsymbol{p}^{(i)} = [p_1^{(i)} p_2^{(i)} \dots p_K^{(i)}]^T$,$\boldsymbol{q}^{(i)} = [q_1^{(i)} q_2^{(i)} \dots q_K^{(i)}]^T$ at iteration index i. $\hat{\boldsymbol{p}}_k'^{(i)}$, $\hat{\boldsymbol{p}}_k'^{(i)}$, $\hat{\boldsymbol{q}}_k'^{(i)}$ and $\hat{s}'^{(i-1)}$ are the estimates obtained by applying the inter- and intra-frame constraints.

2.6.3 Iterative Wiener Filtering using Complex LPC Speech Analysis

The conventional LPC model relies on the assumption that speech is stationary within the analysis frame, hence it cannot model variations of the speech spectrum. To capture the time-varying nature of speech, Funaki et al. proposed a time-varying complex auto regressive (TV-

CAR) model in [37]. To this end, an analytic speech signal is assumed:

$$s^c(n) = \frac{s(n) + j s_H(n)}{\sqrt{2}}, \tag{2.59}$$

where c indicates the complex nature of the target signal and the subscript H denotes the Hilbert transform of the observed signal $y(n)$. The LP coefficients in eq. (2.55) are adapted according to the new signal model:

$$s^c(n) = \sum_{k=1}^{p} a_k^c s^c(n-k) + g w^c(n), \tag{2.60}$$

where a_k^c are now the complex-valued AR coefficients, modelled by a complex basis expansion. The authors argued that the speech PSD estimate is refined by the assumption of an analytic target signal by replacing a_k in eq.(2.54) with a_k^c. In [38], Funaki proposed to replace the LP in the iterative Wiener filter by the TV-CAR method, resulting in the scheme shown in figure 2.8

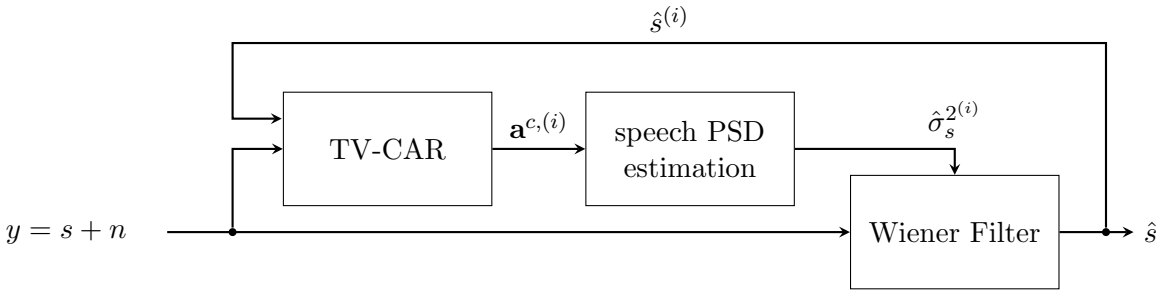

Figure 2.8: Block Diagram of the iterative Wiener Filter using the TV-CAR method instead of the conventional LP, $\mathbf{a}^{c,(i)} = [a_1^{c,(i)} a_2^{c,(i)} \dots a_K^{c,(i)}]^T$

2.6.4 Iterative Closed-Loop Phase-Aware Single-Channel Speech Enhancement

Whereas the two aforementioned iterative methods are working in the linear prediction domain, in 2013 Mowlaee and Saeidi [20] presented an iterative approach that combines a phase aware amplitude estimator together with a phase estimator. The two estimators are used to estimate the amplitude and phase spectra iteratively [20]:

To provide an initial amplitude estimate, the conventional Wiener filter is used. This estimate is exploited to estimate a phase spectrum for speech components with a local SNR lower than 6 (dB). This is justified by the finding that for spectral components with SNRs larger than 6 (dB), the noisy phase is a decent estimate of the clean phase from a perceptual point of view [22]. To refine the spectral amplitude estimate, the enhanced phase spectrum is fed back into an amplitude estimator, now phase aware as defined in eq. (2.43). The parameters μ and β in eq. (2.43) are set to $\mu = 0.5$, modelling a more heavy-tailed prior than the χ-distribution [13] and $\beta = 1$, respectively. The phase-aware amplitude estimator is of the form:

$$\boxed{\hat{A} = \sqrt{\frac{1}{2}\frac{\xi}{\mu+\xi}\sigma_d^2}\left(\frac{\Gamma(2)}{\Gamma(1)}\frac{D_{-2}(\nu)}{D_{-1}(\nu)}\right),} \tag{2.61}$$

where ν takes into account the phase deviation $\Delta\phi = \vartheta - \alpha$:

$$\nu = -\sqrt{2\frac{\xi}{\mu+\xi}\zeta}\cos(\vartheta - \alpha). \tag{2.62}$$

In this case α is assumed to be known by its estimate $\hat{\alpha}$, which was achieved by the geometry method [39]. The so-obtained amplitude and phase estimates together build a complex speech spectrogram estimate $\hat{\mathbf{S}}^{(i)}$ consisting of the entries $\hat{S}^{(i)}(\lambda,k) = \hat{A}^{(i)}(\lambda,k)\mathrm{e}^{\hat{\alpha}^{(i)}(\lambda,k)}$, where i denotes the iteration index. As a stopping criterion, the authors used the inconsistency constraint derived in [21]:

$$F(\hat{\mathbf{S}}^{(i)}) = \mathrm{STFT}\left(\mathrm{iSTFT}\left(\hat{\mathbf{S}}^{(i)}\right)\right) - \hat{\mathbf{S}}^{(i)}, \qquad i : \text{iteration index.} \tag{2.63}$$

The consistency criterion takes care of the fact that a modified STFT-spectrogram does not in general correspond to an existing time domain signal. The iterative application of the phase and the amplitude estimator improves the consistency of the modified spectrogram and was shown to saturate across iterations, so that its use as a stopping criterion was well justified. The results reported in terms of PESQ and segmental SNR showed superior performance compared to [3]. Figure 2.9 pictures the processing steps by a block diagram.

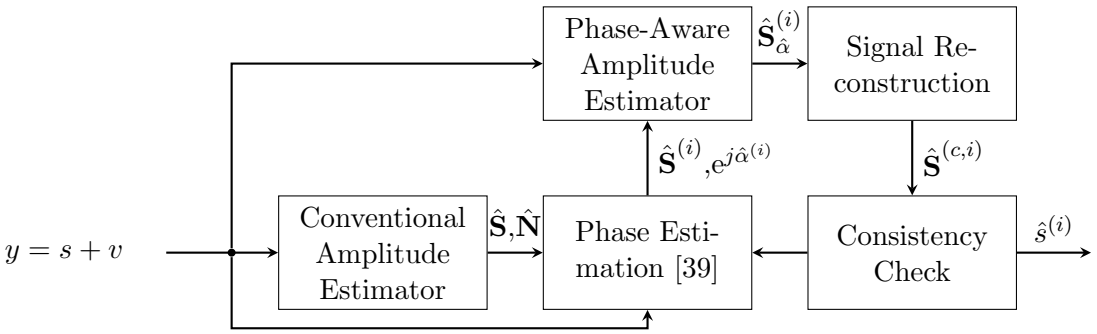

Figure 2.9: Block Diagram of the iterative closed loop method proposed in [20]. $\hat{\boldsymbol{S}}^{(c,i)}$ denotes the STFT of the reconstructed signal, $STFT\left(iSTFT\left(\hat{\boldsymbol{S}}^{(i)}\right)\right)$, used for the consistency criterion. The spectrogram estimate $\hat{\boldsymbol{S}}^{(i)}$ consists of the amplitude estimates from eq.(2.61) and the phase estimate achieved by the geometry method presented in [39].

2.7 Joint MAP Estimation of Speech Amplitude and Phase Assuming Uniform Prior Distribution of the Phase

One novelty of this thesis lies in incorporating the assumption of a non-uniform prior distribution of the speech phase component into a joint Maximum a Posteriori (JMAP) estimation rule. There have been JMAP estimators in the past, assuming uniform-phase prior; two of these noise suppression rules will be presented in the following. They both assume the same prior distribution on the spectral phase but they make different assumptions about the spectral amplitude distributions namely:

1. Rayleigh Distribution (by Godsill and Wolfe in [8])
2. Super-Gaussian Distribution (by Lotter Vary in [5])

2.7.1 Godsill-Wolfe

In [8], Wolfe and Godsill presented three (computationally) efficient alternatives to the Ephraim-Malah estimation rule in [3]. Amongst two other methods they introduced a *Joint maximum a*

Posteriori Spectral Amplitude and Phase Estimator, where they assume independent, zero mean, complex Gaussian Distributions of the speech and noise DFT-coefficients with variances σ_s^2 and σ_d^2:

$$S \sim \mathcal{N}_2(0, \sigma_s^2), \tag{2.64}$$
$$N \sim \mathcal{N}_2(0, \sigma_d^2). \tag{2.65}$$

The complex noisy speech DFT coefficients $Y_k = R_k \mathrm{e}^{j\vartheta_k}$ can be expressed by the sum of the clean speech and the noise components:

$$Y_k = S_k + N_k \tag{2.66}$$

The described model yields the same marginal and joint distributions already mentioned in section 2.2, equations eq. (2.11)-eq. (2.13). Godsill and Wolfe obtain the joint MAP estimator by maximising the posterior distribution $p(A_k, \alpha_k|Y_k)$ (for the sake of notational convenience the frequency index k will be neglected in the following):

$$p(A, \alpha|Y) = p(Y|A, \alpha)p(A, \alpha) = \frac{A}{\pi^2 \sigma_s^2 \sigma_d^2} \mathrm{e}^{-\frac{|R\mathrm{e}^{j\vartheta} - A\mathrm{e}^{j\alpha}|}{\sigma_d^2} - \frac{A^2}{\sigma_s^2}} \tag{2.67}$$

Due to the monotonically increasing behaviour of the ln(.), it can be applied to eq. (2.67) in order to simplify the differentiating needed to find its maximum.

$$J = \ln(p(A, \alpha|Y)) = -\frac{|R\mathrm{e}^{j\vartheta} - A\mathrm{e}^{j\alpha}|}{\sigma_d^2} - \frac{A^2}{\sigma_s^2} + \ln(A) - \ln\left(\pi^2 \sigma_s^2 \sigma_d^2\right) \tag{2.68}$$

The joint MAP estimates of the phase and the amplitude are now obtained by differentiating eq. (2.68) with respect to both variables and setting the terms to zero:

$$\boxed{\hat{\alpha}_{Godsill/Wolfe}^{JMAP} = \vartheta} \tag{2.69}$$

$$\boxed{\hat{A}_{Godsill/Wolfe}^{JMAP} = \frac{\xi + \sqrt{\xi^2 + 2(1+\xi)\frac{\xi}{\zeta}}}{2(1+\xi)} R} \tag{2.70}$$

Due to the assumption of a uniform phase prior probability density function, the joint MAP phase estimate is equal to the observed noisy phase ϑ. The amplitude estimate is a function of prior and posterior SNR. For analysis purposes the corresponding gain function that fulfils $\hat{A} = G \cdot R$ is often used:

$$\boxed{G_{Godsill/Wolfe}^{JMAP} = \frac{\xi + \sqrt{\xi^2 + 2(1+\xi)\frac{\xi}{\zeta}}}{2(1+\xi)}} \tag{2.71}$$

2.7.2 Lotter-Vary

In 2005, Lotter and Vary proposed a JMAP estimation method in [5] where they jointly maximize the posterior probability of amplitude and phase assuming uniform prior distribution of the

phase:

$$\hat{A} = \underset{A}{\operatorname{argmax}} \left(p(A,\alpha|Y)\right) = \underset{A}{\operatorname{argmax}} \left(\frac{p(Y|A,\alpha)p(A,\alpha)}{p(Y)}\right) \tag{2.72}$$

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmax}} \left(p(A,\alpha|Y)\right) = \underset{A}{\operatorname{argmax}} \left(\frac{p(Y|A,\alpha)p(A,\alpha)}{p(Y)}\right) \tag{2.73}$$

Using the Super-Gaussian speech model in section 2.2.2 for $P(A)$ and the independence assumption of phase and amplitude $p(A,\alpha) = p(A)p(\alpha)$ the log-likelihood function is obtained:

$$\log(p(Y|A,\alpha)p(A,\alpha)) = \log\left(\frac{\mu^{\nu+1}}{2\pi^2\sigma_d^2\sigma_s^{\nu+1}\Gamma(\nu+1)}\right) - \frac{\left|Y - A\mathrm{e}^{j\alpha}\right|}{\sigma_d^2} + \nu\log(A) - \mu\frac{A}{\sigma_s} \tag{2.74}$$

Differentiating and setting to zero yields the desired estimates. The joint MAP estimate of the phase is the noisy phase:

$$\hat{\alpha}^{JMAP}_{Lotter/Vary} = \vartheta \tag{2.75}$$

The JMAP gain function for amplitude enhancement is given as follows:

$$\boxed{\hat{G}^{JMAP}_{Lotter/Vary} = u + \sqrt{u^2 + \frac{\nu}{2\zeta}},} \tag{2.76}$$

with:

$$\boxed{u_{Lotter/Vary} = \frac{1}{2} - \frac{\mu}{4\sqrt{\zeta\xi}}} \tag{2.77}$$

There are no special functions, such as the Bessel-functions or the parabolic cylinder function involved in the estimation rule, thus the computational complexity of this joint MAP method is very low compared to MMSE estimators (e.g. [3, 12]).

3 Proposed Contributions

Based on the methods presented in Chapter 2, the following Chapter introduces a novel approach to estimate phase and amplitude in a joint way. The prior knowledge of the distribution of the unwrapped phase is taken into account by incorporating it into the joint maximum a posteriori framework.

3.1 Proposed Method Assuming Non-Uniform Prior Distribution of Phase

The derivation of the proposed Joint MAP estimator follows the same principle as illustrated in section 2.7.2 and 2.7.1. It maximizes the maximum a posteriori probability of $\hat{A}$ and $\hat{\alpha}$ given the observation $Y = R\mathrm{e}^{j\vartheta}$. The assumed amplitude-distribution is similar to the distribution derived by Lotter and Vary:

$$p(A) = \frac{\mu^{\nu+1}}{\Gamma(\nu+1)} \frac{A^{\nu}}{\sigma_s^{\nu+1}} \mathrm{e}^{-\frac{A\mu}{\sigma_s}}, \tag{3.1}$$

whereas the distribution of the unwrapped phase is considered to be a Von Mises distribution with concentration parameter κ and circular mean α_μ instead of a uniform distribution:

$$p(\alpha) = \frac{\mathrm{e}^{\kappa \cos(\alpha - \alpha_\mu)}}{2\pi I_0(\kappa)}. \tag{3.2}$$

The assumption of independence of amplitude and phase yields:

$$p(A,\alpha) = p(A)p(\alpha) = \frac{A^{\nu}\mu^{\nu+1}}{\sigma_s^{\nu+1}\Gamma(\nu+1)2\pi I_0(\kappa)} \mathrm{e}^{\kappa\cos(\alpha-\alpha_\mu)-\frac{A\mu}{\sigma_s}}. \tag{3.3}$$

Taking into account the statistical model in section 2.2.1 gives us:

$$p(R,\vartheta|A,\alpha) = \frac{1}{\pi\sigma_d^2} \mathrm{e}^{\frac{|R\mathrm{e}^{j\vartheta} - A\mathrm{e}^{j\alpha}|}{\sigma_d^2}}. \tag{3.4}$$

Similar to [5], the values for $\hat{A}$ and $\hat{\alpha}$, that maximise $p(A,\alpha|R,\vartheta)$, are searched. From Bayes' Theorem we know that $p(A,\alpha|R,\vartheta) = \frac{p(R,\vartheta|A,\alpha)p(A,\alpha)}{p(R,\vartheta)}$. The maximisation is achieved by partial

differentiating with respect to the variables A and α and setting the obtained derivatives to zero. Since $p(R,\vartheta)$ is independent of A and α, it can be dropped in the further course of the derivation. This yields the following likelihood function to maximise:

$$L(A,\alpha) = p(R,\vartheta|A,\alpha)p(A,\alpha) = \frac{A^{\nu}\mu^{\nu+1}}{2\pi^2\sigma_s^{\nu+1}\Gamma(\nu+1)I_0(\kappa)\sigma_d^2}\mathrm{e}^{\kappa\cos(\alpha-\alpha_\mu)-\frac{A\mu}{\sigma_s}+\frac{|R\mathrm{e}^{j\vartheta}-A\mathrm{e}^{j\alpha}|^2}{\sigma_d^2}}. \quad (3.5)$$

To simplify the differentiation of eq. (3.5) the log-likelihood function $\ln(L(A,\alpha))$ is evaluated instead of $L(A,\alpha)$. This is possible, because the natural logarithm function is monotonically increasing and therefore the solutions will not change due to the transformation:

$$\{\hat{A}^{MAP},\hat{\alpha}^{MAP}\} = \arg\max_{A,\alpha} L(A,\alpha) = \arg\max_{A,\alpha} \ln(L(A,\alpha)). \quad (3.6)$$

Additionally, all terms that are independent of A and α can be discarded, so that the equation that remains to be solved is of a much more simple form than eq. (3.5):

$$\ln(L(A,\alpha)) = \nu\ln(A) - \frac{\mu}{\sigma_s}A + \frac{|R\mathrm{e}^{\vartheta}-A\mathrm{e}^{j\alpha}|^2}{\sigma_d^2} + \kappa\cos(\alpha-\alpha_\mu), \quad (3.7)$$

with $|R\mathrm{e}^{j\vartheta}-A\mathrm{e}^{j\alpha}|^2 = R^2+A^2-2AR\cos(\vartheta-\alpha)$ and further neglecting R^2:

$$\boxed{\ln(L(A,\alpha)) = \nu\ln(A) - \frac{\mu}{\sigma_s}A + \frac{A^2}{\sigma_d^2} - \frac{2AR}{\sigma_d^2}\cos(\vartheta-\alpha) + \kappa\cos(\alpha-\alpha_\mu)}. \quad (3.8)$$

As an approximation the amplitude and the phase are assumed to be independent, thus the partial derivatives of eq. (3.8) can be taken in order to obtain the MAP estimates of A and α, respectively.

3.1.1 The MAP Spectral Phase Estimator

In order to obtain the MAP-Phase estimate we set the derivative of $\ln(L(A,\alpha))$ with respect to α to zero:

$$\frac{\partial\ln(L(A,\alpha))}{\partial\alpha} = \frac{2AR}{\sigma_d^2}\sin(\vartheta-\alpha) - \kappa\sin(\alpha-\alpha_\mu) \stackrel{!}{=} 0. \quad (3.9)$$

Inserting $\sin(a-b) = \sin(a)\cos(b) - \cos(a)\sin(b)$ gives us:

$$\frac{2AR}{\sigma_d^2}\left(\sin(\vartheta)\cos(\alpha) - \cos(\vartheta)\sin(\alpha)\right) = \kappa\left(\sin(\alpha)\cos(\alpha_\mu) - \cos(\alpha)\sin(\alpha_\mu)\right), \quad (3.10)$$

$$\frac{\sin(\alpha)}{\cos(\alpha)} = \frac{\frac{2AR}{\sigma_d^2}\sin(\vartheta)+\kappa\sin(\alpha_\mu)}{\frac{2AR}{\sigma_d^2}\cos(\vartheta)+\kappa\cos(\alpha_\mu)}. \quad (3.11)$$

$$(3.12)$$

Resulting in:

$$\boxed{\hat{\alpha}^{MAP} = \arctan\left(\frac{\beta\sin(\vartheta)+\kappa\sin(\alpha_\mu)}{\beta\cos(\vartheta)+\kappa\cos(\alpha_\mu)}\right)}, \quad (3.13)$$

where $\beta = \frac{2AR}{\sigma_d^2}$. As the clean spectral amplitude A is not known the MAP estimate is incorporated instead, so that $\beta = \frac{2\hat{A}^{MAP}R}{\sigma_d^2}$.

Eq. (3.13) shows that the observation ϑ and the circular mean α_μ are traded off against each other by the weighting factors β and κ. At high SNRs, β becomes large and the the noisy observation ϑ is considered to be more reliable. For highly voiced regions, where the concentration parameter κ is very large, the weight lies on the circular mean α_μ.

Relation to previous MAP Phase Estimators

It can be shown that the STFT MAP Phase estimator in eq. (3.13) evaluated at harmonics is equivalent to the harmonic MAP Phase estimator in [9], once the harmonics (ω) are resolved by the DFT (i.e. $\omega = \frac{2\pi k}{N_{FFT}}$). To illustrate this, the following identities are used (exceptionally, to show the relation between the two estimators, the frequency is indexed again):

$$\mathrm{Im}\{Y(k)\} = R(k)\sin(\vartheta(k)) \tag{3.14}$$
$$\mathrm{Re}\{Y(k)\} = R(k)\cos(\vartheta(k)) \tag{3.15}$$

Eq. (3.14) and eq. (3.15) as well as the DFT definition eq. (2.2) of $Y(k)$ (neglecting the window and the frame shift) can be plugged into eq. (3.13):

$$\hat{\alpha}^{MAP}(k) = \tan^{-1}\left(\frac{\frac{2A(k)}{\sigma_d^2(k)}\mathrm{Im}\left\{\sum_{n=0}^{N-1} y(n)\mathrm{e}^{\frac{-j2\pi nk}{N}}\right\} + \kappa(k)\sin\left(\alpha_\mu(k)\right)}{\frac{2A(k)}{\sigma_d^2(k)}\mathrm{Re}\left\{\sum_{n=0}^{N-1} y(n)\mathrm{e}^{\frac{-j2\pi nk}{N}}\right\} + \kappa(k)\cos\left(\alpha_\mu(k)\right)}\right), \tag{3.16}$$

$$\hat{\alpha}^{MAP}(k) = \tan^{-1}\left(\frac{-\frac{2A(k)}{\sigma_d^2(k)}\sum_{n=0}^{N-1} y(n)\sin\left(\frac{2\pi nk}{N}\right) + \kappa(k)\sin\left(\alpha_\mu(k)\right)}{\frac{2A(k)}{\sigma_d^2(k)}\sum_{n=0}^{N-1} y(n)\cos\left(\frac{2\pi nk}{N}\right) + \kappa(k)\cos\left(\alpha_\mu(k)\right)}\right). \tag{3.17}$$

In order to obtain samples of the noisy observation at harmonics, its DFT representation is sampled at frequency ω by convolution with the δ function:

$$Y(\omega) = \sum_{n=0}^{N-1} y(n)e^{\frac{-j2\pi nk}{N}} * \delta(\frac{2\pi nk}{N} - \omega n) = \sum_{n=0}^{N-1} y(n)e^{-j\omega n}. \tag{3.18}$$

Using the harmonic $Y(\omega)$ in eq. (3.18) results in the MAP phase estimate at harmonics:

$$\hat{\alpha}_h^{MAP} = \tan^{-1}\left(\frac{-\frac{2A_h}{\sigma_{d,h}^2}\sum_{n=0}^{N-1} y(n)\sin\left(\omega n\right) + \kappa_h\sin\left(\alpha_{\mu,h}\right)}{\frac{2A_h}{\sigma_{d,h}^2}\sum_{n=0}^{N-1} y(n)\cos\left(\omega n\right) + \kappa_h\cos\left(\alpha_{\mu,h}\right)}\right), \tag{3.19}$$

which is the same estimator as derived in [9]. The amplitude A, the noise PSD σ_d^2 as well as κ and α_μ need to be evaluated at harmonics only too, which is denoted by the subscript h.

3.1.2 The MAP Spectral Amplitude Estimator assuming a Super-Gaussian Distribution

The phase estimate is clearly a function of the spectral amplitude, in this case of the MAP-estimate of the amplitude $\hat{A}^{MAP}$, which will be derived in the following. To this end, the derivative of $\ln(L(A,\alpha))$ with respect to A is set to zero:

$$\frac{\partial \ln(L(A,\alpha))}{\partial A} = \frac{\nu}{A} - \frac{\mu}{\sigma_s} + 2\frac{A}{\sigma_d^2} - 2\frac{R}{\sigma_d^2}\cos(\vartheta - \alpha) \stackrel{!}{=} 0, \tag{3.20}$$

which leads to the quadratic equation:

$$A^2 - A\left(R\cos(\vartheta-\alpha) - \frac{\sigma_d^2\mu}{2\sigma_s}\right) + \nu\frac{\sigma_d^2}{2} = 0, \tag{3.21}$$

$$A^2 - AR\left(\cos(\vartheta-\alpha) - \frac{\sigma_d^2\mu}{2R\sigma_s}\right) + R^2\nu\frac{\sigma_d^2}{2R^2} = 0, \tag{3.22}$$

$$A^2 - AR\left(\cos(\vartheta-\alpha) - \frac{\mu}{2\sqrt{\xi\zeta}}\right) + R^2\frac{\nu}{2\zeta} = 0. \tag{3.23}$$

By defining $2u = \cos(\vartheta-\alpha) - \frac{\mu}{2\sqrt{\xi\zeta}}$ the solution of eq. (3.23) is:

$$\hat{A}^{MAP} = uR + \sqrt{R^2u^2 - R^2\frac{\nu}{2\zeta}}, \tag{3.24}$$

$$\hat{A}^{MAP} = G^{MAP}R. \tag{3.25}$$

$$\tag{3.26}$$

The corresponding estimation rule:

$$\boxed{G^{MAP} = u + \sqrt{u^2 - \frac{\nu}{2\zeta}}}. \tag{3.27}$$

Furthermore, since there is no access to the clean phase α:

$$\boxed{u = \frac{\cos(\vartheta - \hat{\alpha}^{MAP})}{2} - \frac{\mu}{4\sqrt{\xi\zeta}}}. \tag{3.28}$$

The MAP amplitude estimate is a function of the MAP phase estimate and vice versa. This interdependency is used to implement a joint estimation rule, based on iterations in section 3.2 and 3.3.

3.1.3 The MAP Spectral Amplitude Estimator assuming a Rayleigh distribution

Using the *Rayleigh* amplitude model as presented in section 2.7.1 together with a *von Mises* distribution of the phase yields the following posterior:

$$p(A,\alpha|Y) = p(Y|A,\alpha)p(A,\alpha) = \frac{A}{\pi^2\sigma_s^2\sigma_d^2 I_0(\kappa)}\mathrm{e}^{-\frac{|R\mathrm{e}^{j\vartheta} - A\mathrm{e}^{j\alpha}|}{\sigma_d^2} - \frac{A^2}{\sigma_s^2} + \kappa\cos(\alpha-\alpha_\mu)} \tag{3.29}$$

Taking the ln(.) and neglecting all terms independent of A and α leads the function to maximize:

$$L(A,\alpha) = \ln(A) - \frac{A^2}{\sigma_d^2} + \frac{2AR}{\sigma_d^2}\cos(\vartheta-\alpha) + \kappa\cos(\alpha-\alpha_\mu) - \frac{A^2}{\sigma_s^2} \tag{3.30}$$

The MAP phase estimator is independent of the assumed prior distribution of the amplitude, therefore only the phase-aware amplitude estimator changes (obtained by differentiating with respect to A and setting to zero):

$$\boxed{G_2^{MAP} = \frac{\xi\cos(\vartheta-\alpha) + \sqrt{\xi^2\cos^2(\vartheta-\alpha) + 2(1+\xi)\frac{\xi}{\zeta}}}{2(1+\xi)}}. \tag{3.31}$$

If the phase deviation $\vartheta - \alpha$ gets zero, the obtained estimator yields the phase-insensitive estimator of Godsill and Wolfe in [8], now being a special case of the phase-aware estimation rule.

It is interesting to note that if the complex coefficients Y and S are orthogonal to each other (which means that $\cos(\vartheta - \alpha) = 0$), the amplitude estimate gets independent of the observation R as it does not contain any information about the underlying amplitude A:

$$\hat{A}^{MAP} = G_2^{MAP} R = R\frac{\sqrt{2(1+\xi)\frac{\xi}{\zeta}}}{2(1+\xi)} = R\sqrt{\frac{\xi\sigma_d^2}{R^2 2(1+\xi)}} = \sqrt{\frac{\xi\sigma_d^2}{2(1+\xi)}} \tag{3.32}$$

3.1.4 Relation to previous MAP Amplitude Estimators

In this section the relation of the phase-sensitive MAP amplitude estimator to its corresponding phase-insensitive MAP estimator will be examined (corresponding means that the same prior distribution of the STFT amplitudes is assumed). To this end it will be assumed that the noise is known. The JMAP amplitude estimator of Lotter and Vary [5] eq. (2.76) can be considered as a special case of the derived estimation rule in eq. (3.27).
The gainfunction in both cases yields:

$$G = u + \sqrt{u^2 + \frac{\nu}{2\cdot\zeta}}, \tag{3.33}$$

where for [5]:

$$u_{Lotter/Vary} = \frac{1}{2} - \frac{\mu}{4\cdot\sqrt{\zeta\cdot\xi}}, \tag{3.34}$$

and for the proposed method:

$$u_{proposed} = \frac{\cos\Delta\phi}{2} - \frac{\mu}{4\cdot\sqrt{\zeta\cdot\xi}}, \tag{3.35}$$

with the phase deviation $\Delta\phi = \vartheta - \alpha$. The statistical values of the prior SNR ξ and the posterior SNR ζ will be replaced by their instantaneous values in this section:

$$SNR_{prior} = \frac{A^2}{D^2}, \tag{3.36}$$

$$SNR_{post} = \frac{R^2}{D^2}. \tag{3.37}$$

If $\cos(\vartheta - \alpha) = 1$ (which corresponds to the case $\vartheta - \alpha = 0 \rightarrow \vartheta = \alpha$) the two estimators are the same. To illustrate, how different the two amplitude-estimators behave (especially in low SNR-regions), we plot the weights of the gainfunctions for different prior SNRs SNR_{prior} across the posterior SNR SNR_{post} (figure 3.2 and figure 3.3). As both SNRs are now given and therefore the triangle of the speech-amplitude, the noise-amplitude and the amplitude of the observation is well defined (figure 3.1), geometry implicitly provides the corresponding phase deviation, which can be used for evaluating the corresponding weights.

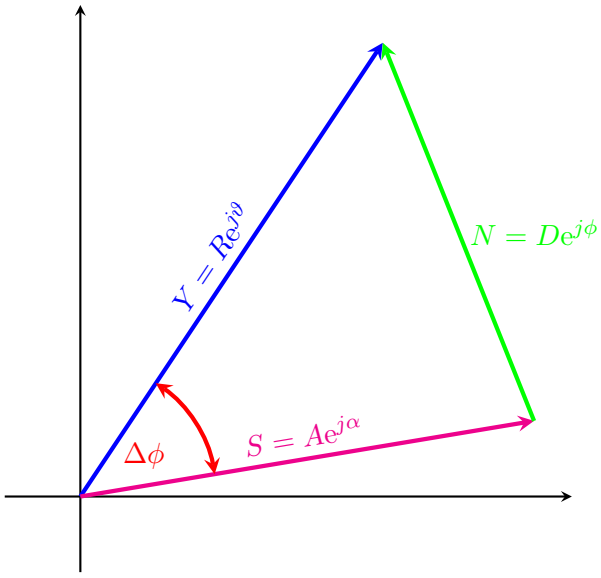

Figure 3.1: The triangle of observation, clean speech and noise

Applying the law of cosines yields the following equation:

$$D^2 = R^2 + A^2 - 2AR\cos\Delta\phi \tag{3.38}$$

Which can be rewritten (dividing by D^2) as:

$$1 = SNR_{post} + SNR_{prior} - 2\sqrt{SNR_{post}SNR_{prior}}\cos\Delta\phi \tag{3.39}$$

and solved for $\cos\Delta\phi$:

$$\boxed{\cos\Delta\phi = \frac{SNR_{post} + SNR_{prior} - 1}{2\sqrt{SNR_{post}SNR_{prior}}}}. \tag{3.40}$$

The geometric relations also show that only certain combinations of prior and posterior SNRs are valid, as the absolute value of the cosine cannot be larger than 1:

$$\left|\frac{SNR_{post} + SNR_{prior} - 1}{2\sqrt{SNR_{post}SNR_{prior}}}\right| \leq 1 \tag{3.41}$$

Assuming the prior SNR is given, from eq. (3.41) we obtain a valid region for the posterior SNR:

$$\boxed{SNR_{prior} - 2\cdot\sqrt{SNR_{prior}} + 1 \leq SNR_{post,valid} \leq SNR_{prior} + 2\cdot\sqrt{SNR_{prior}} + 1}. \tag{3.42}$$

Lotter-Vary

Figure 3.2 shows that for high SNRs the gain functions suppress very similar, whilst for lower SNR regions the phase aware amplitude estimator shows more noise suppression. The still very often used STSA estimator of Ephraim and Malah [3] is also included in the analysis to show how the different criteria (MMSE vs. MAP) and the prior distribution choice can affect the estimation rule.

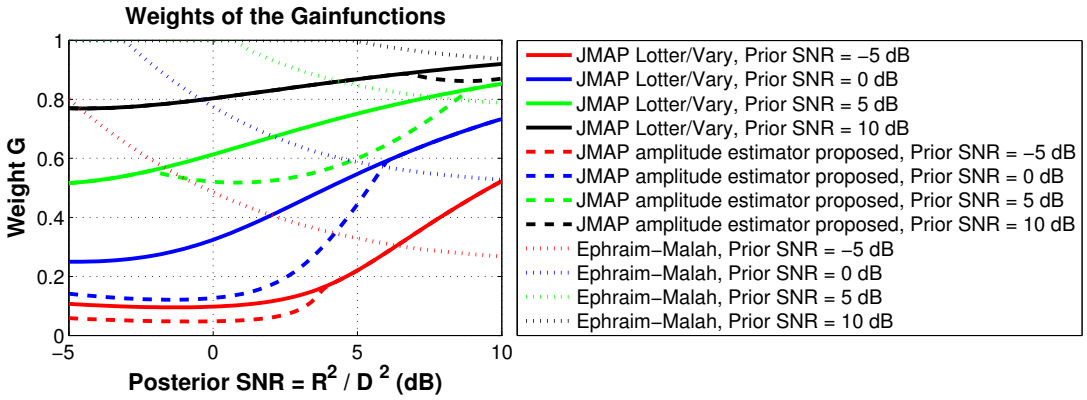

Figure 3.2: Weights of the gainfunctions, JMAP: Super-Gaussian amplitude prior

Godsill-Wolfe

A similar picture is given in figure 3.3, where the phase-insensitive and the phase-sensitive JMAP amplitude estimators derived from a Rayeigh amplitude model are compared with each other. Again lower SNR regions tend to be more suppressed by the phase-aware than by the phase-unaware method.

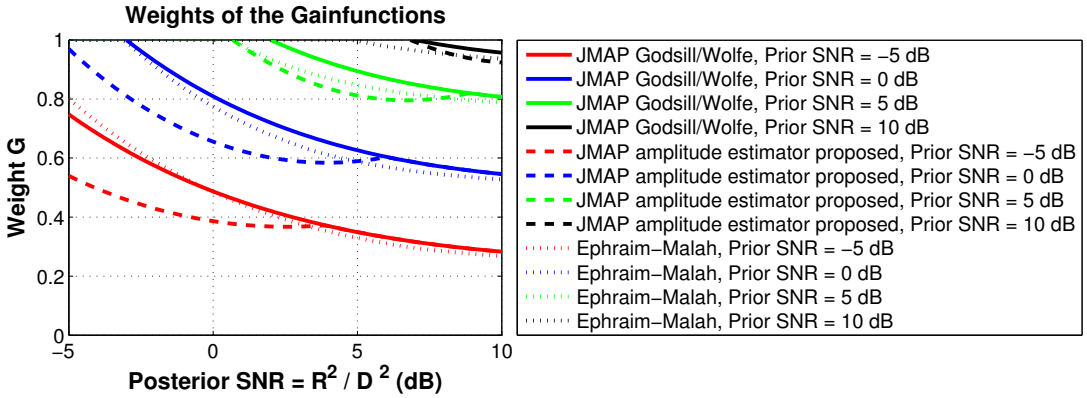

Figure 3.3: Weights of the gainfunctions, JMAP: Rayleigh amplitude prior

3.2 Joint Estimation in STFT Domain

Since the amplitude estimate depends on the phase and vice versa the following section describes how to connect the two estimators. The goal is to have a joint estimation framework that can improve both, phase and amplitude of noise corrupted speech. The estimation rules in eq. (3.13) and eq. (3.24) are non-linear, therefore it is a challenging task to jointly solve them analytically. Hence, the estimators are connected by applying them iteratively. Interconnecting the two estimators is not straightforward either, as there arise several questions. The most obvious question is, whether to start with an amplitude or a phase estimate and which initial values to use within the first iteration. The JMAP amplitude estimator of Lotter and Vary [5] is well examined and known to perform superior to MMSE estimators in terms of noise reduction in speech present segments [49] of a noisy signal. Thus, the choice to initialize α with the noisy phase ϑ in order to obtain a first (phase-unaware) amplitude estimate seems to be a reasonable choice. However, an initial phase estimate obtained by smoothing the unwrapped phase as in [25] could be an alternative.

Beside the question, which estimator to start with, it is also important to clarify, if the iterative procedure should be applied framewise or for the whole signal. This problem arises, when estimating the prior SNR decision directed [3] from frame to frame. The two approaches will be referred to as:

- the *inner loop* (framewise estimation of amplitude and phase)
- the *outer loop* (iterating over the whole signal)

3.2.1 Parameter Estimation

In the following, the assignment of the parameters, needed for the JMAP estimator, will be discussed.

The *Super-Gaussian* amplitude distribution is modelled by the parameters μ and ν, which have been chosen with $\mu = 1.74$ and $\nu = 0.126$ in [5] as they minimise the Kullback-Leibler divergence between the empirical distribution of speech coefficients and the parametric *Super-Gaussian* distribution. The Kullback-Leibler divergence is an information theoretical measure, which describes the discriminative power between two random variables emitted by different probability density functions. The parameters μ and ν can also be chosen to fit other distributions such as the Gamma distribution, however, this thesis will stick to the parameters found to give an optimal fit in an information theoretical way by Lotter and Vary.

The parameters of the *von Mises* distribution are directly estimated from the noisy observation. The sample mean angle α_μ is calculated as described in eq. (2.21), while the concentration parameter κ is estimated using the following approximation of the inverse of eq. (2.22) [43]:

$$\kappa = \begin{cases} 2\bar{Z} + \bar{Z}^3 + 5\frac{\bar{Z}^5}{6} & \text{if } \bar{Z} < 0.53 \\ -0.4 + 1.39\bar{Z} + \frac{0.43}{1-\bar{Z}} & \text{if } 0.53 \leq \bar{Z} < 0.85 \\ \frac{1}{\bar{Z}^3 - 4\bar{Z}^2 + 3\bar{Z}} & \text{if } 0.85 \leq \bar{Z} \end{cases} \tag{3.43}$$

κ- and α_μ-Estimation for the STFT-Domain

At harmonics, the unwrapped phase is obtained by removing the linear phase part due to the fundamental frequency, hence the circular statistics of one harmonic can easily be estimated. In the STFT domain, the unwrapped phase is approximated by a de-trended phase. In order to obtain a trend phase, the STFT phase is smoothed with a 20 ms moving average filter. A frameshift of one sample ensures that no wrapping-jumps of the phase are missed (up to $\frac{f_s}{2}$). Further the small frameshift provides a large number of data points, needed to obtain reliable statistics. The trend phase is subtracted from the instantaneous phase yielding an unwrapped phase. After compensating the phase shift due to the window, the circular statistics methods described above are applied. The resulting κ estimate along time and frequency is pictured in figure 3.4, left panel.

An alternative way of obtaining the κ estimate is to estimate it first at harmonics and assigning the values at harmonic h to the STFT-bins lying within its main lobewidth. The outcome of this method is illustrated in figure 3.4, right panel. As a a consequence of the underlying harmonic model, κ shows a corresponding structure, which is more likely to fit speech if the model order is estimated correctly. The mean value α_μ can be assigned similar from harmonics to DFT-bins.

One way to deal with the model order problem was presented in [12]; Gerkmann proposed to assign κ proportional to the voicing probability (see eq. (2.52)), using different factors for low and high frequency regions.

Since κ can be assumed to be SNR dependent (the higher the concentration parameter, the higher the SNR) another empirical estimation was examined; Assigning κ to a multiple of the posterior SNR:

$$\hat{\kappa} = P\hat{\zeta}, \tag{3.44}$$

where the factor P relating κ to the posterior SNR needs to be empirically chosen. The mean value α_μ can in turn be estimated by the de-trending method or by assigning it from harmonics. In section 3.2.5 the approaches are compared with respect to their effects on noise suppression.

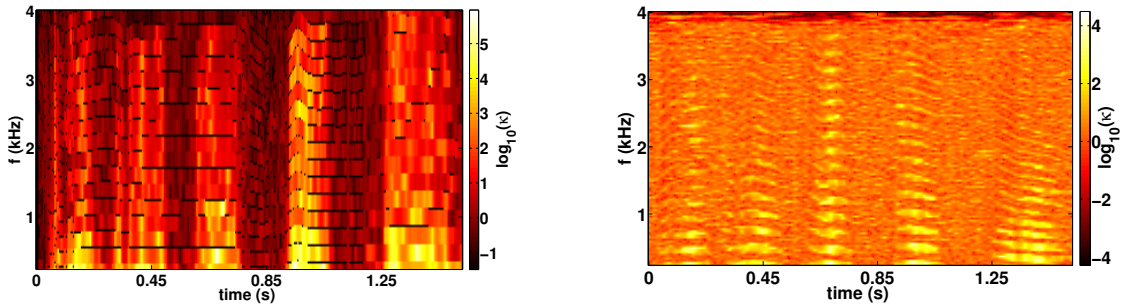

Figure 3.4: Time-frequency representations of κ, obtained by (left) unwrapping the phase at harmonics, (right) unwrapping the phase by detrending; both for cean speech

3.2.2 Analysis/Synthesis Setup

The STFT setup for the iterative JMAP follows closely the setup chosen by Lotter and Vary, since the parameters of the amplitude distribution depend strongly on that setup. Only the window was adapted, because the Blackman window was found to be suitable for phase enhancement since it has a high side-lobe rejection. However, this is on the cost of a wider main lobe [50]. The main lobe width of a Blackman window with length N is given as follows [50]:

$$\Delta\omega_{mlw} = \frac{12\pi}{N} \tag{3.45}$$

The ability to seperate harmonics from each other depends on the fundamental frequency and on the main lobe width:

$$\frac{\Delta\omega_{mlw}}{2} f_s \leq 2\pi f_0 \tag{3.46}$$

For a chosen window length of 32 ms at a sampling frequency of $f_s = 8$ (kHz), the smallest frequency value that can be resolved by the window is:

$$f_{0,min} = \frac{12\pi}{4 \cdot 256 \cdot \pi \,(\text{ samples})} 8000 \frac{(\text{ samples})}{(\text{s})} = 46.88\,(\text{ Hz}) \tag{3.47}$$

These equations assume that the window is centred around the harmonic, which is usually not the case in the STFT domain. However, as an approximation the chosen window along with the window length is appropriate for the estimator. This is the case for speech, where fundamental frequencies $f_0 > 80$ (Hz) can be expected. While the parameter estimation is carried out at a frame shift of 1 sample, the estimation itself uses a frame shift of $\frac{1}{8}$ of the frame length.

Figure 3.5 illustrates the impact of the window onto the separation of the harmonics.

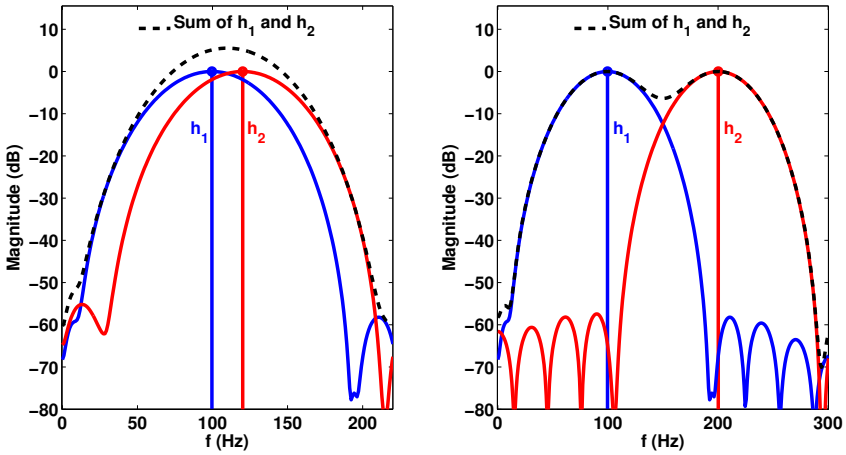

Figure 3.5: Two harmonics windowed by a Blackman window with length 32 (ms). The left panel shows two harmonics that cannot be separated due to the window ($f_{h_1} = 99.61$ (Hz), $f_{h_2} = 120.12$ (Hz)); two separable harmonics are shown on the right panel ($f_{h_1} = 99.61$ (Hz), $f_{h_2} = 200.20$ (Hz)). (blue, solid) harmonic 1, (red, solid) harmonic 2, (black, dashed) sum of the two harmonics.

3.2.3 The Stopping Criterion

To find the optimal number of iterations, a stopping criterion is needed. One requirement such a stopping criterion has to fulfil is that it should reflect the main purpose of speech-enhancement; improving the quality and/or the intelligibility of the noise-corrupted speech. In order to asses the quality and/or intelligibility improvement, the reference signal is needed. Since the clean speech is not accessible in real world applications, directly optimizing on the quality or intelligibility is not possible. In [34] this problem stayed unsolved and a heuristic stopping criterion had to be employed, whereas in [20] the consistency criterion showed sufficient convergence.

Since the estimates of phase and amplitude should both maximise the same log-likelihood function given in eq. (3.8), an intuitive way of assessing the convergence of the iterative method is the tracking of the log-likelihood function across iterations. Figure 3.6 shows how the function gets saturated after one iteration.

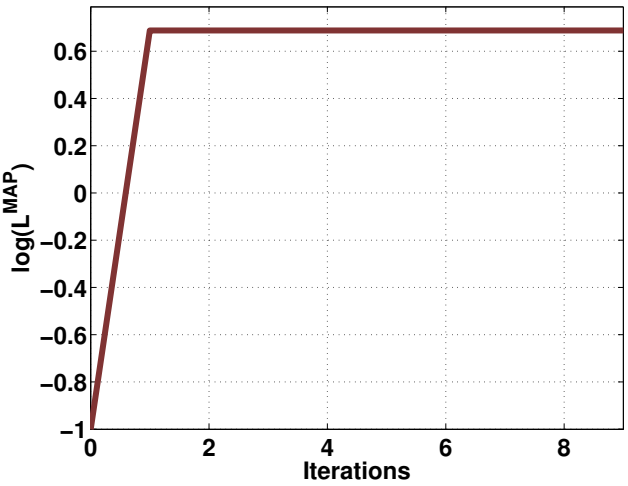

Figure 3.6: Saturation of the log-likelihood function across iterations, utterance: GRID-corpus [51], "bin blue at l 4 soon", speaker A50

To get a better insight in how the iteration number influences the estimates directly, the change of the MAP-phase and amplitude (i.e. gainfunction) across iterations was analysed:

$$\Delta\hat{\alpha}_i^{MAP} = \frac{2}{N\lambda_{max}} \sum_{\lambda=1}^{\lambda_{max}} \sum_{k=1}^{\frac{N}{2}} \left(\hat{\alpha}_i^{MAP}(\lambda,k) - \hat{\alpha}_{i-1}^{MAP}(\lambda,k)\right)^2 \quad , \; i = 1,2,..., \tag{3.48}$$

$$\Delta G_i^{MAP} = \frac{2}{N\lambda_{max}} \sum_{\lambda=1}^{\lambda_{max}} \sum_{k=1}^{\frac{N}{2}} \left(G_i^{MAP}(\lambda,k) - G_{i-1}^{MAP}(\lambda,k)\right)^2 \quad , i = 1,2,..., \tag{3.49}$$

with λ, λ_{max}, i, k and N being the frame index, the frame number, the iteration index, the frequency index and the framelength. The start value $\hat{\alpha}_0^{MAP}(\lambda,k)$ is initialised with the noisy phase ϑ. Figure 3.7 shows how the estimates converge. However, iterating by means of maximising the joint posterior probability does not inherently provide the best result in terms of quality or intelligibility. To this end the maximum number of iterations has been chosen heuristically, motivated by the performance in quality and intelligibility, which turned out to be the best after two iterations. This choice again is well justified by the convergence behaviour shown in figure 3.7, which illustrates that after two iterations there is hardly any change in the estimates.

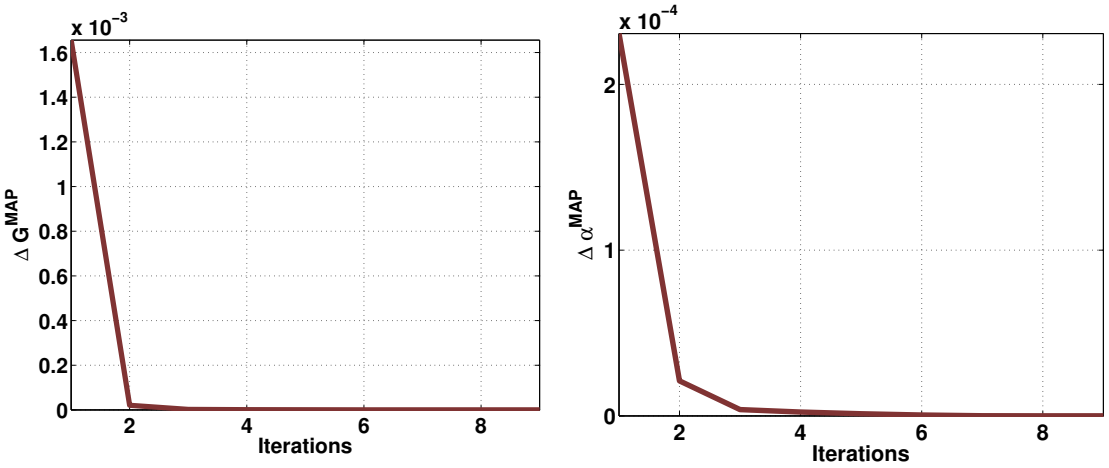

Figure 3.7: Convergence across iterations of (left) gain function (right) phase estimate

3.2.4 Noise PSD Estimation

The noise PSD estimate is obtained using the *Unbiased MMSE-Based Noise Power Estimator* proposed in [42]. It takes into account the Speech Presence Probability into the MMSE estimation of noise. The method has been reported to outperform state-of-the art techniques such as the minimum-statistics approach [18].

3.2.5 Iterative Estimation

In the STFT-domain, the iterative procedure performed best, when using the *inner loop* implementation i.e. the decision directed prior SNR at frame λ is estimated with respect to the amplitude estimate $\hat{A}(\lambda-1)$, obtained by the iterative procedure. Starting with an initial phase-unaware amplitude estimate, two iterations are carried out. The corresponding block diagram is given in figure 3.8. Bold letters denote a full STFT spectrogram, whilst λ indicates the frame number for the decision directed prior SNR estimation. The maximum value of the iteration index i is the empirically found maximum number of iterations, denoted by a capital

I. The operator DD(.) is used to clarify that the prior SNR ξ is estimated using the decision directed approach described in eq. (2.27) [3]. The κ used for the phase estimation is obtained by the de-trending method described in section 3.2.1. Estimating it at harmonics yields higher noise suppression but more musical noise (see spectral peaks in figure 3.9). If κ is assigned proportional to the posterior SNR and α_μ is estimated at harmonics the musical noise gets de-emphasized on the cost of more speech suppression. The resulting performance depends highly on the assignment of the factor P connecting κ and the posterior SNR so that no consistent assignment could be found. As a trade-off between the two methods, the κ and α_μ assignment by the de-trending method yielded the best results.

Figure 3.10 illustrates how the estimator works in magnitude domain as well as in phase domain. While the circular variance gives information on how dense the circular data is around its mean, the group delay is defined as the negative derivative of the phase with respect to the frequency and has been reported to be correlated with the perceived quality of speech [46] ; its structure is shown in the lower panel of figure 3.10.

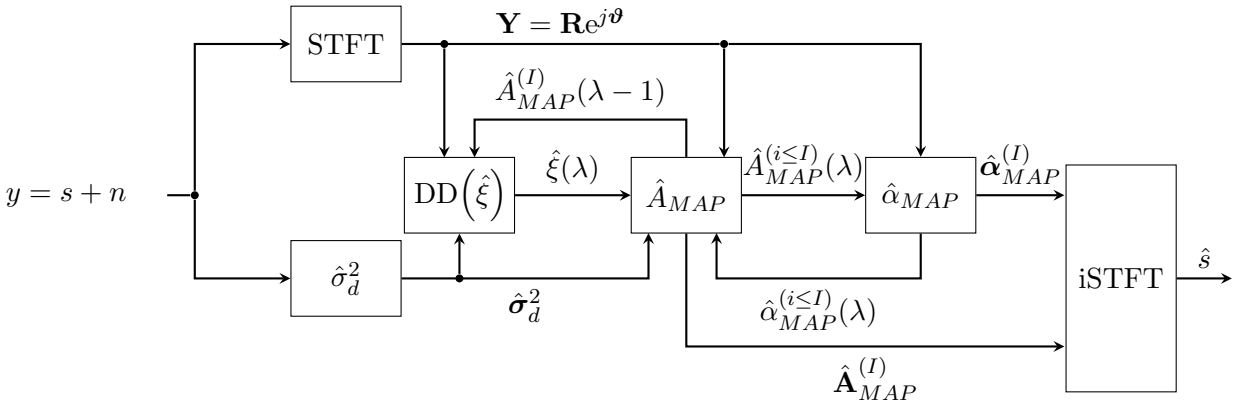

Figure 3.8: Block Diagram of the inner loop

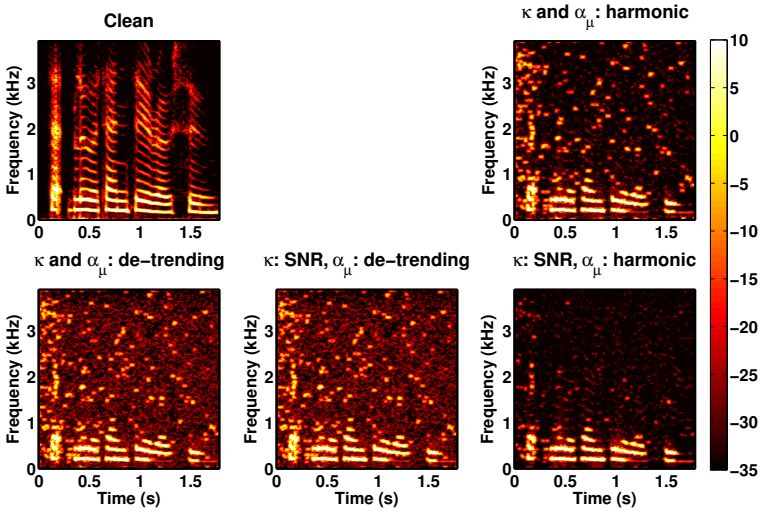

Figure 3.9: Different methods of assigning κ and α_μ and their impact on the spectrogram. The musical noise is de-emphasized by assigning κ to a multiple of the posterior SNR ($P = 40$) and α_μ to the harmonic mean phase. This is on the cost of more speech damping at higher frequencies.

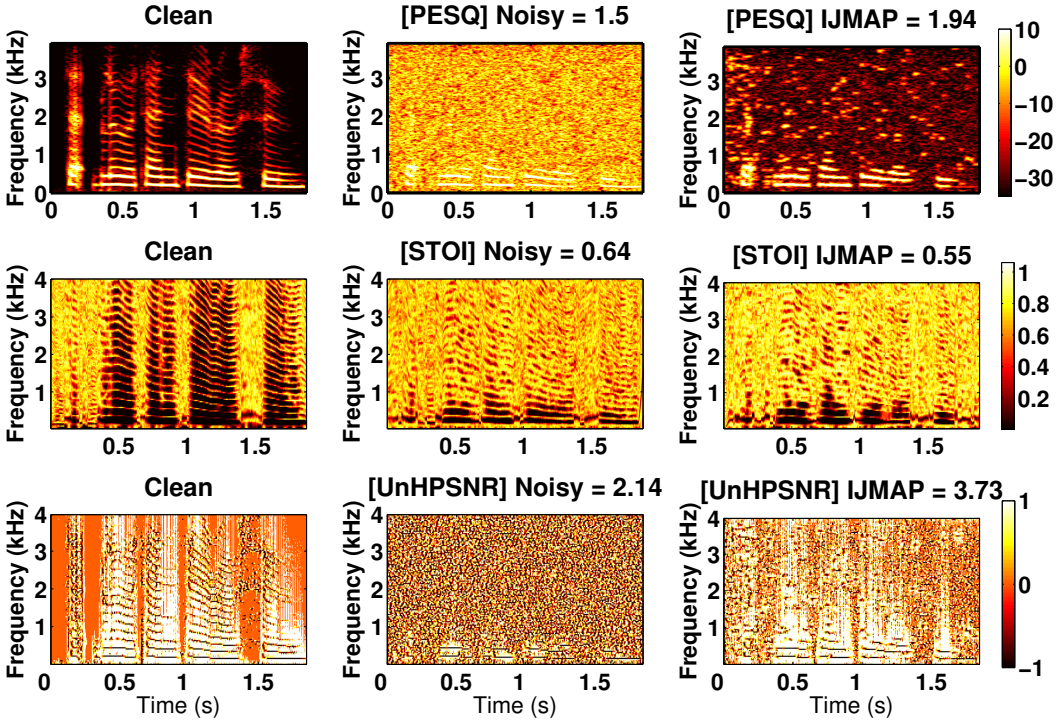

Figure 3.10: Spectrogram, circular variance and group delay, 2 Iterations. The scores predicting quality (PESQ), intelligibility (STOI) and phase recovery (HPSNR) are shown on the top of each panel.

From figure 3.10 it can be seen, that especially the group delay structure is well restored by the iterative method.

3.2.6 Evaluation

Since listening tests are very time-consuming, instrumental metrics, predicting the performance of enhanced speech, are commonly used. In order to choose an optimal number of iterations with respect to predicted intelligibility and quality, the performance across iterations has been monitored. For evaluation purposes the GRID corpus [51] together with different noise-types taken from the NOISEX-92 database [52] has been used. After two iterations, degradation in predicted quality and intelligibility has been observed, so that the maximum iteration number was chosen accordingly with two. The metrics used for this evaluation are described in the following. A detailed summary of the obtained results, using the settings in Table 3.1 (settings and parameters of the iterative joint MAP estimator), is given in Chapter 4.

Parameter	Settings
sampling frequency f_s	8 kHz
framelength	32 ms
frameshift, JMAP estimator	4 ms
frameshift, parameter estimation	0.125 ms
window type	Blackman
FFT-length	256
ν (shape parameter of the *Super-Gaussian* distribution)	0.126
μ (shape parameter of the *Super-Gaussian* distribution)	1.74
Number of Iterations	2
Floor of the gain-function	-25 dB

Table 3.1: Simulation system settings

Evaluation of Speech Quality

The PESQ (*Perceptual Evaluation of Speech Quality*) is recommended by ITU-T [44] for the evalution of perceptual quality. It uses a cognitive model of quality to compare the outputs of a perceptual model, fed with the original and the degraded speech signal. Its range is bounded to the interval $[-0.5, 4.5]$, where 4.5 denotes the maximum performance with respect to quality. Although, originally introduced to evaluate speech processed through networks, it showed the highest correlation with overall speech quality in [45], where the PESQ-score was investigated in terms of speech enhancement. Nevertheless, the PESQ has to be applied with caution, as it is sensitive to over-harmonization of the signal. Thus, speech enhancement algorithms that force a strictly harmonic structure in enhanced signals (e.g. STFTPI [14, 15]) perform outrageous in PESQ, while resulting in a buzzy speech quality [12, 46].

Evaluation of Speech Intelligibility

The *Short Time Objective Intelligibility* (STOI) measure proposed in [48] was developed for methods, where noisy speech is processed by a time-frequency weighting. The correlation coefficient of $\rho = 0.95$ with intelligibility demonstrates it to be highly appropriate, in order to instrumentally predict the intelligibility. Most amplitude-only enhancement methods only improve quality but degrade intelligibility. At the same time a better phase estimate was reported to yield a better intelligibility performance in [24, 50]. The STOI is obtained by segmenting the STFT of the input signals (the noisy signal and its underlying clean speech signal) in 15 one-third-bands, with the lowest center frequency being 150 (Hz) (justified by the fact that this is approximately fitting the cochlea characteristics). Silent regions are removed and the remaining bands are normalized and smoothed along time (around 300 ms [18]). Both input signals are processed the same way and at this stage, a short-time sample correlation coefficient between the two is calculated. Averaging the resulting correlation coefficients along time and sub-bands yields the STOI-measure, bounded to the interval $[-1, 1]$. The STOI was successfully used in [11, 50] to evaluate and predict the speech intelligibility obtained by phase-aware speech enhancement methods.

Evaluation of Phase Recovery

Besides the metrics for quality prediction presented in [46], Gaich and Mowlaee proposed two new metrics in [47] (the unwrapped HPSNR and the unwrapped RMSE) in order to capture

the phase impact on speech intelligibility. The unwrapped harmonic phase SNR (UnHPSNR) is defined as follows:

$$\text{UnHPSNR} = \frac{\sum_{h,\lambda} A^2(h,\lambda)}{\sum_{h,\lambda} A^2(h,\lambda)\left(1-\cos\left(\Psi(h,\lambda)-\hat{\Psi}(h,\lambda)\right)\right)}, \tag{3.50}$$

with $\Psi(h,\lambda)$ and $\hat{\Psi}(h,\lambda)$ denoting the clean unwrapped phase and its estimate at harmonic h and frame index λ, respectively. The unwrapped phase values are achieved by applying a pitch synchronous framing and removing the linear phase part introduced by the fundamental frequency as proposed in [6]. The unwrapped root mean square error (UnRMSE) is given by the following equation:

$$\text{UnRMSE} = \sqrt{\frac{\sum_{h,\lambda} A^2(h,\lambda)\left(\Psi(h,\lambda)-\hat{\Psi}(h,\lambda)\right)^2}{\sum_{h,\lambda} A^2(h,\lambda)}} \tag{3.51}$$

Both metrics use the clean amplitude as a weighting which means that the perceptually important harmonic regions are emphasized. Only the phase recovery is taken into account, therefore clean phase yields an infinite UnHPSNR and an UnRMSE of 0, if reconstructed with the clean amplitude. These upper bounds cannot be reached with noisy amplitude (see [21]). In [47], the UnRMSE was reported to have the highest correlation with subjective intelligibility compared to state of the art methods for speech corrupted with white noise .

3.3 Phase Estimation at Harmonics

The STFT Domain is not necessarily the best choice for speech enhancement as it is a redundant representation in speech inactive regions. The authors of [9, 14] argue that the phase of regions near a harmonic is dominated by the phase of the harmonic itself. Therefore, combining a harmonic phase estimator together with the STFT amplitude estimator is the next option to be examined. As a consequence of the different frame setups, found to be optimal for the phase estimator proposed in [9] and the amplitude estimator, the frame-setup recommended in [9] was chosen. After assigning the phase values within the main lobe width of the analysis window, the STFT representation of the phase is obtained.

The harmonic representation used for the phase estimator is obtained by the HMPD matlab library [6] which uses a pitch synchronous framing, hence the output DFT lengths of the phase estimator depend on the underlying fundamental frequency f_0. After applying the phase-aware MAP amplitude estimator on these frames, an overlap-and-add synthesis procedure is carried out. This helps to have a consistent spectrogram for the next iteration. The estimation procedure is initialised with a phase estimation step. The framework used to implement the joint estimator is the *outer loop* (schematic in figure 3.11).

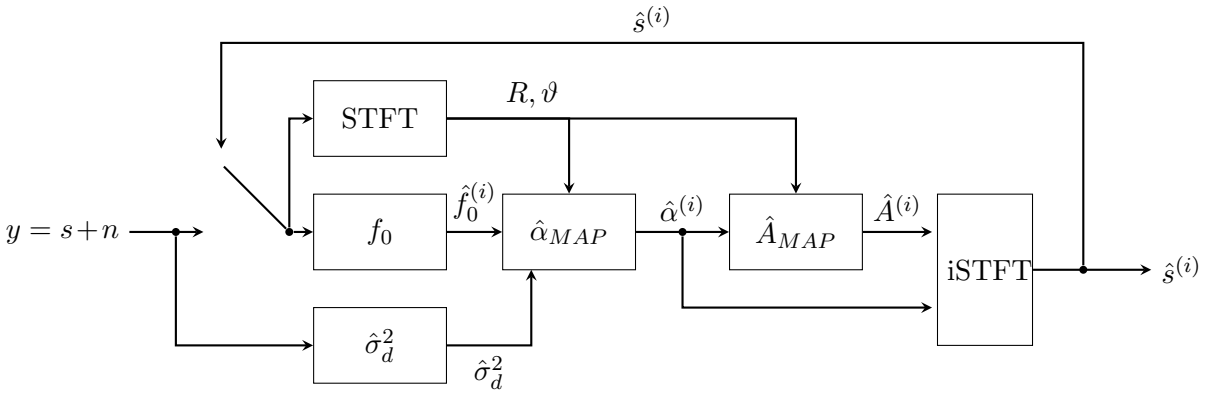

Figure 3.11: The block diagram of the outer loop.

Since the iterative procedure starts with the phase estimation step, the amplitude A_h in eq. (3.19), needs to be initialized. This is achieved by using the noisy DFT-amplitude as a first amplitude estimate. (R) is interpolated to the harmonics in time and frequency (R_h) so that:

$$\hat{A_h}^{(0)} = R_h \tag{3.52}$$

In the second iteration, the phase-aware amplitude estimate can be used. With iteration index (i) it follows

$$\hat{\alpha}_h^{(i)} = \tan^{-1}\left(\frac{-\frac{2\hat{A}_h^{(i-1)}}{\sigma_{d,h}^2}\sum_{n=0}^{N-1} y(n)\sin\left(\omega^{(i)}n\right) + \kappa^{(i)}\sin\left(\alpha_\mu^{(i)}\right)}{\frac{2\hat{A}_h^{(i-1)}}{\sigma_{d,h}^2}\sum_{n=0}^{N-1} y(n)\cos\left(\omega^{(i)}n\right) + \kappa^{(i)}\cos\left(\alpha_\mu^{(i)}\right)}\right) \tag{3.53}$$

Since the enhanced time-domain signal is fed back for iterations, the estimates of the *von Mises* parameters κ and α_μ are also updated within iterations. The synthesis is carried out by a iSTFT with pitch-synchronous frame-lengths. To incorporate the estimated harmonic phase, it is assumed that the phase of the DFT-bins is dominated by the harmonic phase $\hat{\alpha}_h^{(i)}$ if they jointly lye within the main lobe width of the analysis window [9]. Therefore these phase-values are set to the harmonic phase. This procedure is illustrated for one frame in figure 3.12.

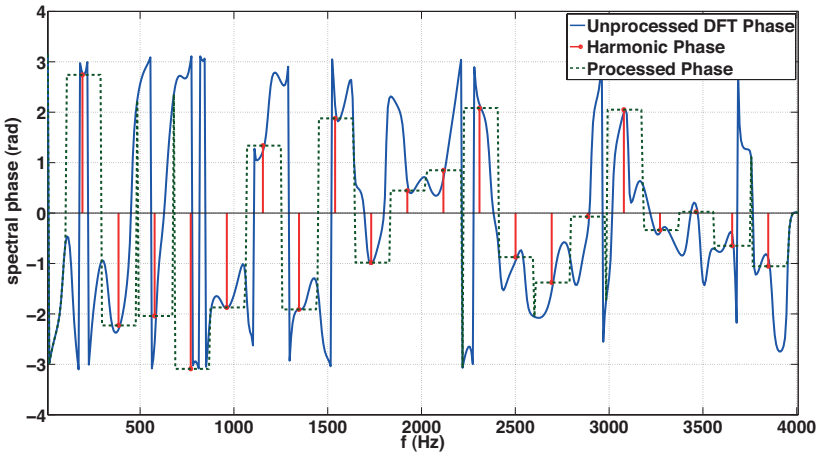

Figure 3.12: Assignment of harmonic phase to DFT phase. (blue) original DFT-phase, (red) harmonic values, (green, dashed) new DFT phase with harmonic phase assigned to the DFT-bins lying within the main lobe width of the analysis window

After the phase assignment the joint MAP amplitude estimation is carried out followed by the iSTFT. The estimator's performance tends to drop after more than two iterations (figure 3.13), which again led to an empirical choice of maximum iteration number of $I = 2$.

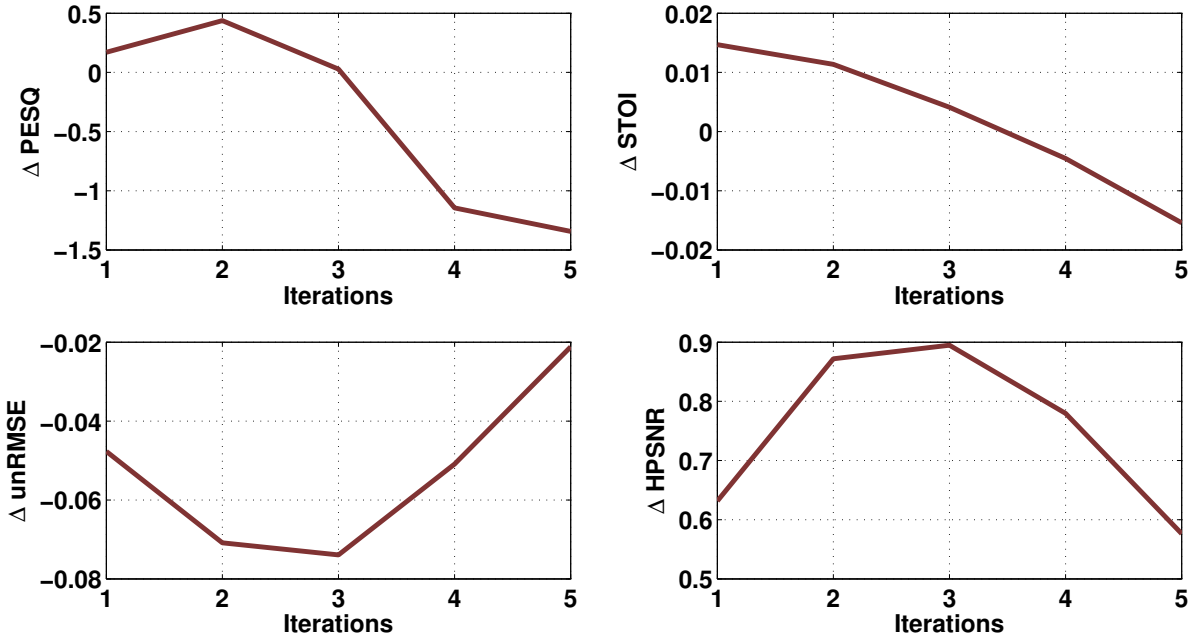

Figure 3.13: Performance in terms of PESQ, STOI, unwrapped RMSE and HPSNR drops after more than two iterations (Δ denotes the difference between the enhanced and noisy utterance). Averaged over: 50 utterances (GRID [51]), 4 noise types (white, pink, pink modulated, babble) [52], SNR: −5, 0, 5 and 10 (dB)

The f_0 estimator used to obtain the fundamental frequencies of the speech signal is the PEFAC estimator, known to be robust against non stationary noise types such as wind-noise [19]. To update the fundamental frequency within iterations, a smoothing dependent on the voicing probability P_v is applied. This helps to stabilize the fundamental frequency-estimate across iterations, in order to avoid a fundamental frequency drift across iterations:

$$\hat{f}_0^{(i)} = P_v \hat{f}_0^{(i)} + (1 - P_v)\hat{f}_0^{(i-1)} \tag{3.54}$$

Figure 3.14 and figure 3.15 illustrate how the harmonic structure of speech is restored across two iterations. The strong over-harmonisation of the amplitude after the second iteration yields a high buzzyness in the signal. The circular variance and group-delay structures fit well to the clean-speech, the obtained phase recovery is well reflected in the UnRMSE results in Chapter 4.

Parameter	Settings
sampling frequency f_s	8 kHz
frame length	pitch synchronous
frame shift	pitch synchronous
window type	Blackman
FFT-length	pitch synchronous
ν (shape parameter of the *Super-Gaussian* distribution)	0.126
μ (shape parameter of the *Super-Gaussian* distribution)	1.74
Number of Iterations	2
Floor of the gain-function	-25 dB

Table 3.2: Outer Loop Settings

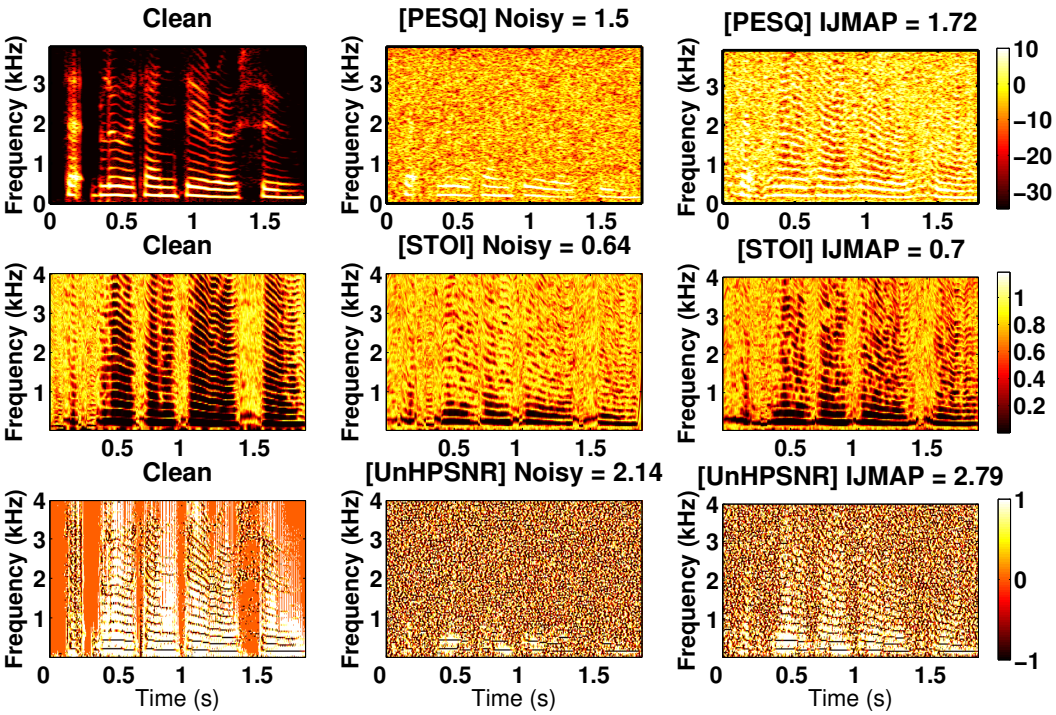

Figure 3.14: Spectrogram, circular variance and group delay 1 Iteration. The scores predicting quality (PESQ), intelligibility (STOI) and phase recovery (HPSNR) are shown on the top of each panel.

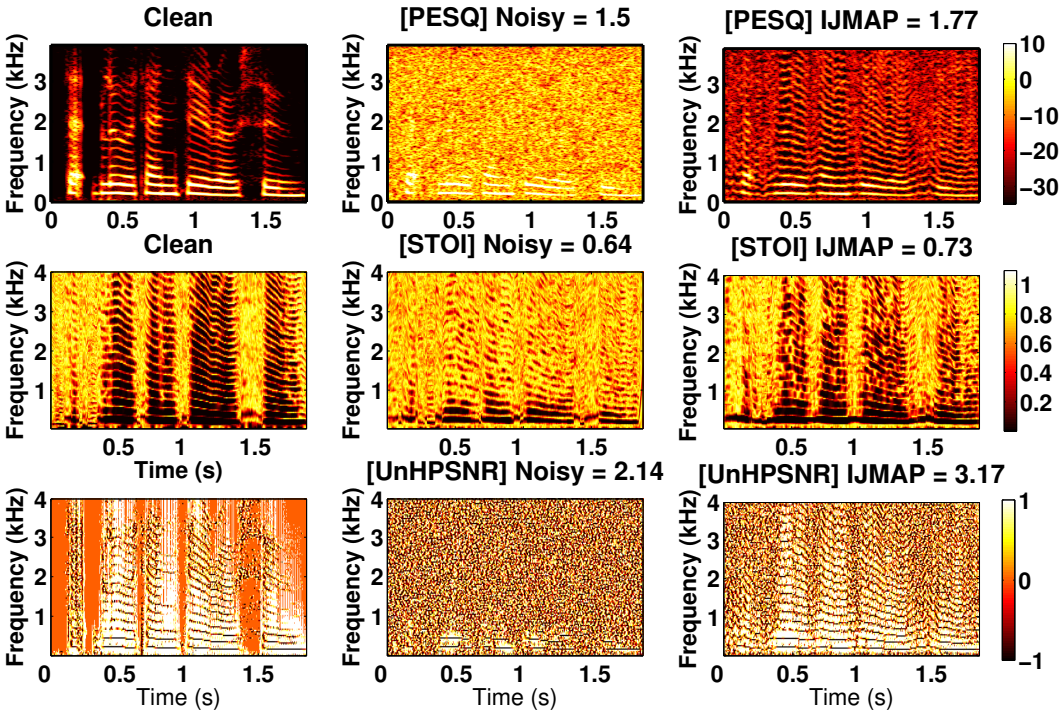

Figure 3.15: Spectrogram, circular variance and group delay 2 Iterations. The scores predicting quality (PESQ), intelligibility (STOI) and phase recovery (HPSNR) are shown on the top of each panel.

4 Results

The following Chapter presents the obtained performance of the proposed estimators, predicted by the instrumental metrics PESQ, STOI, UnHPSNR and UnRMSE (see section 3.2.6). The proposed methods are evaluated with different settings to illustrate upper and lower bounds and the importance of a good phase estimate for amplitude estimation as well as for reconstruction. The scenarios examined are:

- noisy phase for amplitude estimation and noisy phase for reconstruction (which is equivalent to the phase insensitive estimator of Lotter-Vary)
- noisy phase for amplitude estimation and MAP phase for reconstruction (*inner loop*)
- MAP phase for amplitude estimation and noisy phase for reconstruction (*inner loop*)
- MAP phase for amplitude estimation and MAP phase for reconstruction (*inner loop*, proposed)
- clean phase for amplitude estimation and noisy phase for reconstruction (*inner loop*)
- clean phase for amplitude estimation and MAP phase for reconstruction (*inner loop*)
- MAP phase for amplitude estimation and MAP phase for reconstruction (*outer loop*, proposed)
- MAP phase for amplitude estimation and MAP phase for reconstruction, f_0 from clean speech (*outer loop*)
- clean phase for amplitude estimation and clean phase for reconstruction, f_0 from clean speech (*inner loop*)

Since the frame-setup used for the *outer loop* is pitch-synchronous, the different scenarios used for the *inner loop* cannot be combined without any interpolation (which results in degraded performance). Thus, the only additional scenario for the *outer loop* is the oracle-f_0 scenario, which shows how sensitive the method is to f_0-estimation errors.
The results obtained by the *inner loop* implementation show consistent but rather small improvement. One reason is surely the sub-optimal assignment of the prior distribution parameters κ and α_μ. Nevertheless, the promising upper bounds of the performance motivate for further investigation of this problem. In contrast, the *outer loop* implementation shows very large improvement in objective scores. In white noise it outperforms even the clean phase scenario in

terms of PESQ (figure 4.2), which indicates the sensitivity of the PESQ to over-harmonisation. Intelligibility, predicted by UnRMSE and STOI is also improved by the *outer loop*.
In the following, the results are presented in the form of bar-plots, one for each noise-type and metric. Figure 4.1 shows the corresponding legend for the bar-plots and Table 4.1 summarises the evaluation-setup. The results are reported in the form of Δ-scores, where Δ denotes the difference between the scores obtained by the processed speech and the scores of the noisy speech.

Evaluation-Setup	
global SNR	-5, 0, 5, 10 (dB)
noise types	white, pink, pink modulated, babble from [52]
number of utterances	50
speech database	GRID [51]

Table 4.1: Simulation system settings

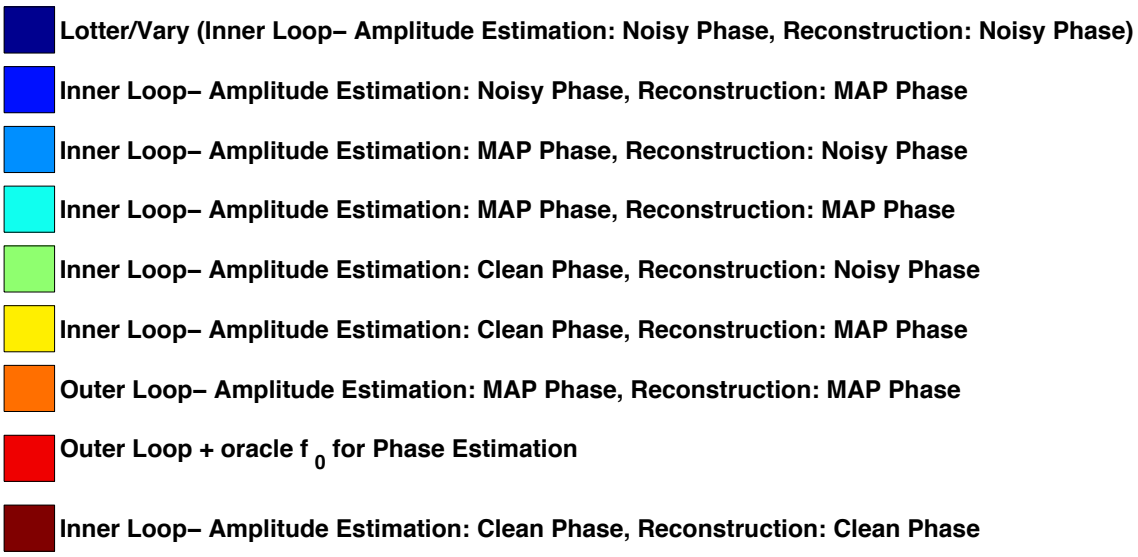

Figure 4.1: Legend

4.1 White Noise

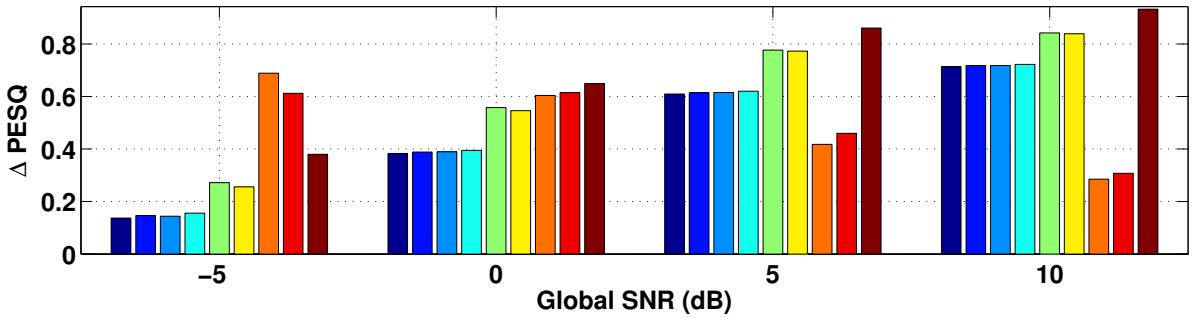

Figure 4.2: PESQ-scores for white noise. The outer loop outperforms the clean phase, which emphasizes the sensitivity of PESQ to over-harmonisation.

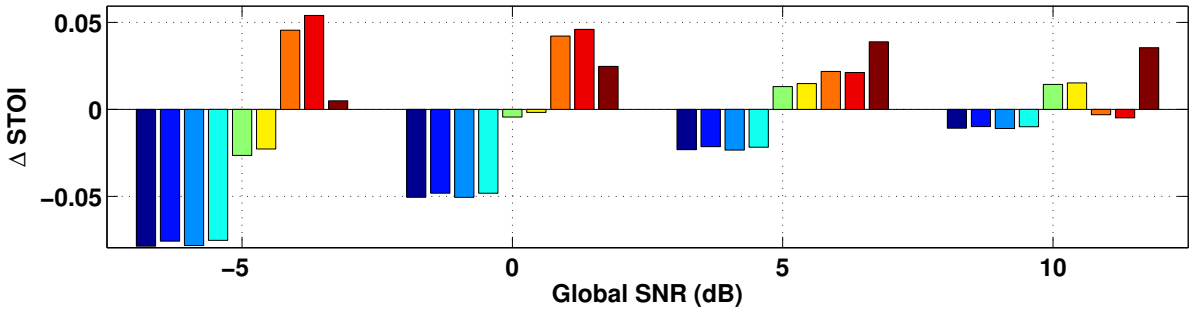

Figure 4.3: STOI-scores for white noise.

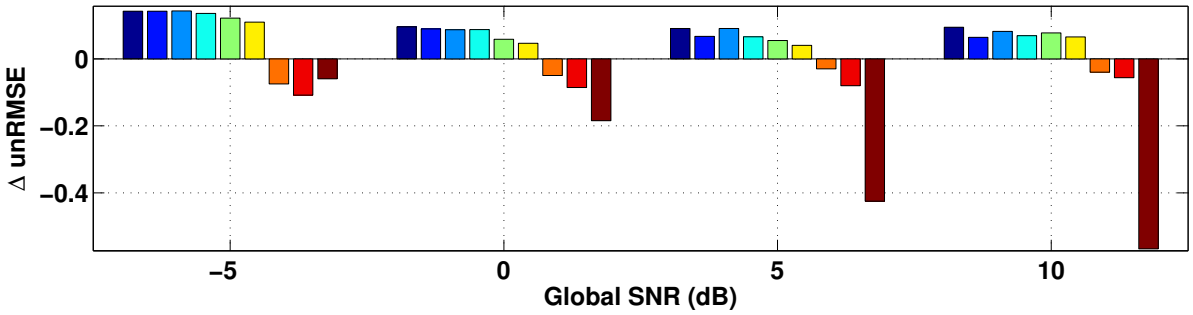

Figure 4.4: UnRMSE-scores for white noise.

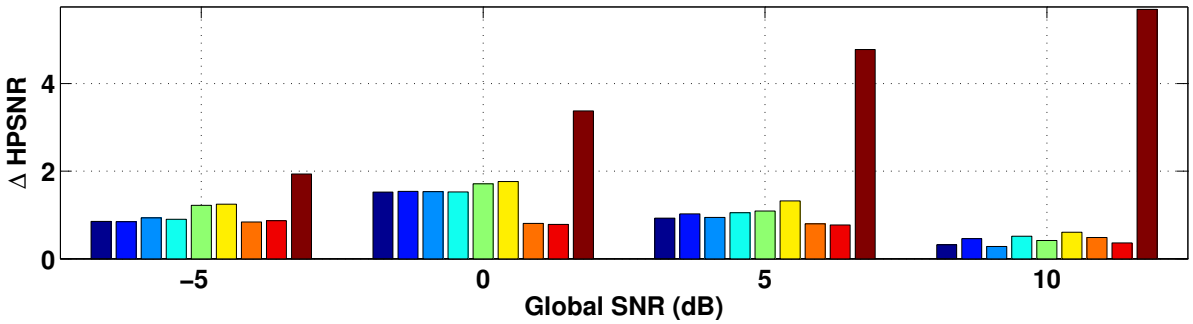

Figure 4.5: UnHPSNR-scores for white noise.

4.2 Pink Noise

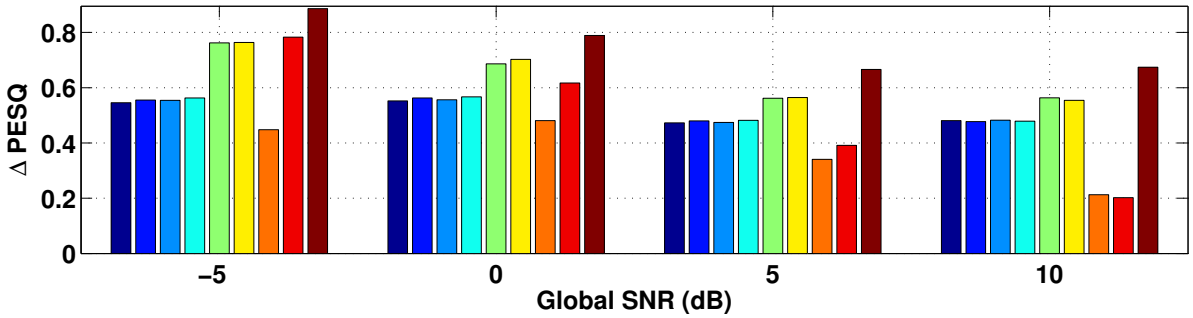

Figure 4.6: PESQ-scores for pink noise.

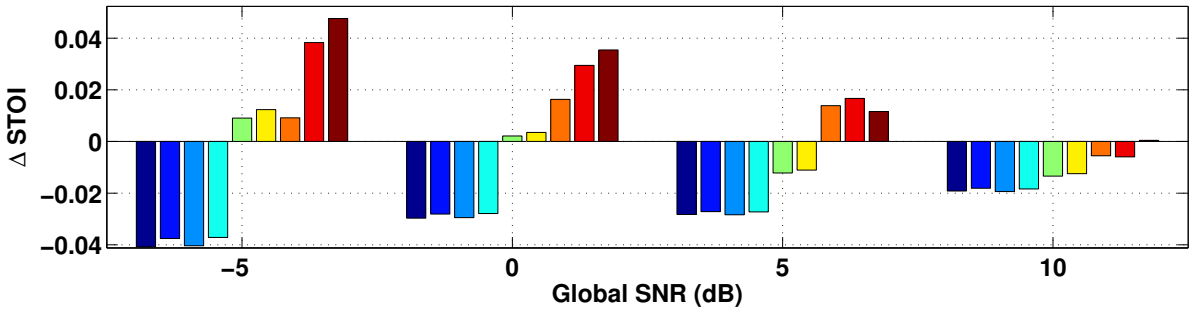

Figure 4.7: STOI-scores for pink noise.

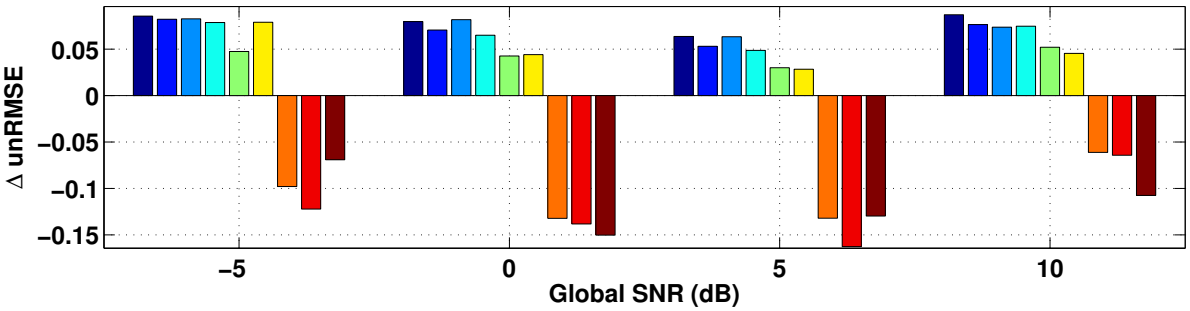

Figure 4.8: UnRMSE-scores for pink noise.

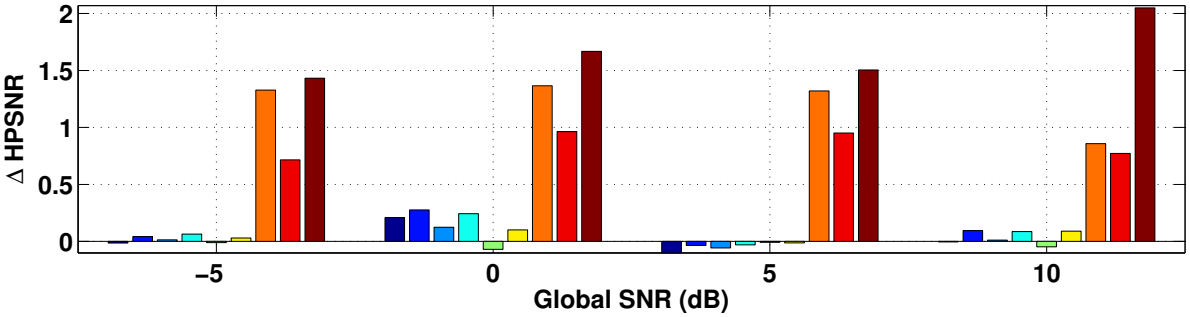

Figure 4.9: UnHPSNR-scores for pink noise.

4.3 Pink Modulated Noise

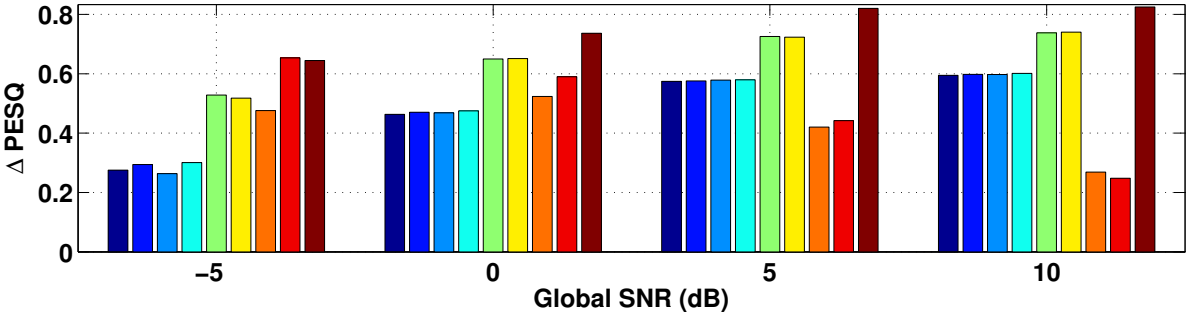

Figure 4.10: PESQ-scores for modulated pink noise

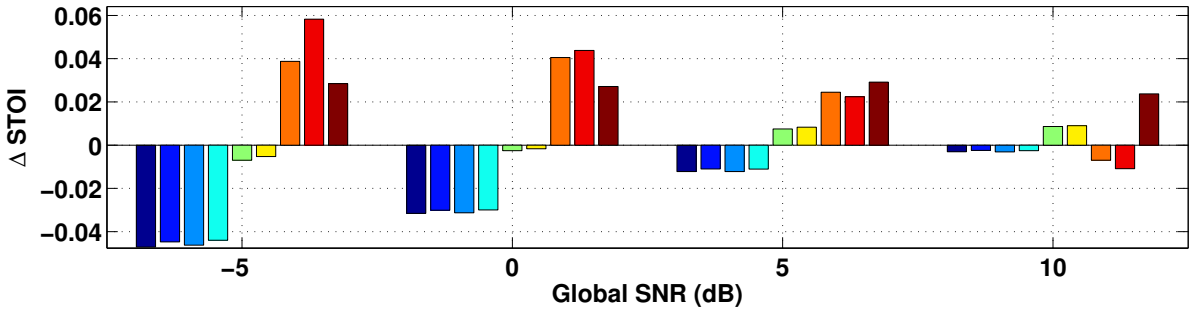

Figure 4.11: STOI-scores for modulated pink noise

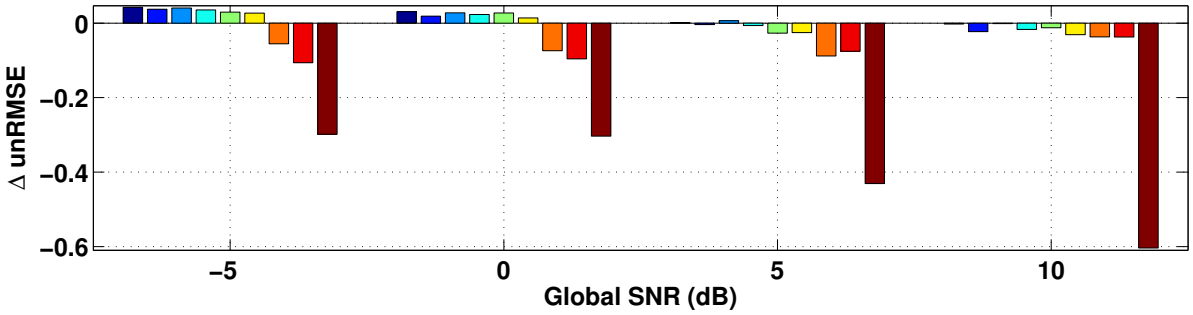

Figure 4.12: UnRMSE-scores for modulated pink noise

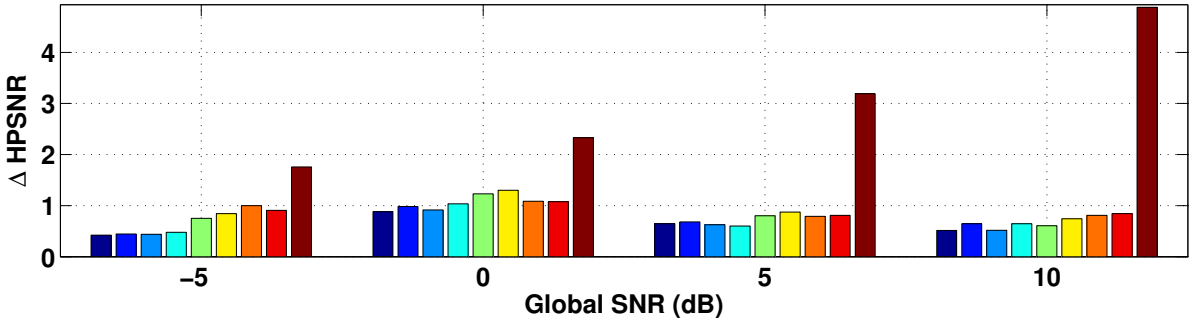

Figure 4.13: UnHPSNR-scores for modulated pink noise

4.4 Babble Noise

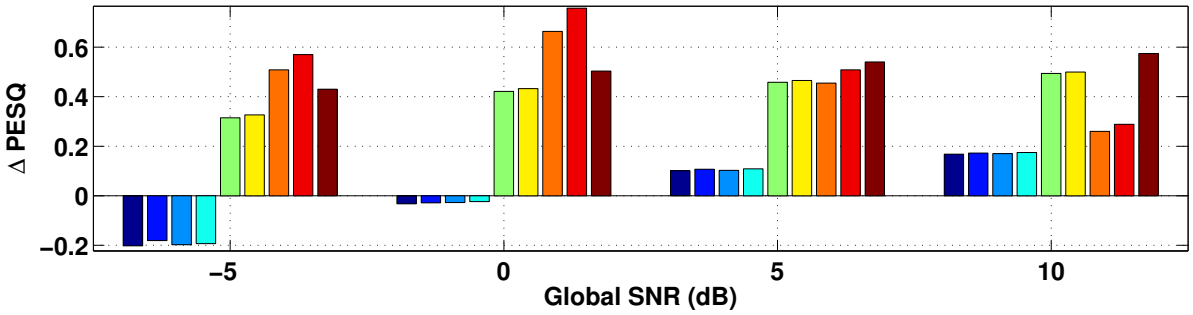

Figure 4.14: PESQ-scores for babble noise

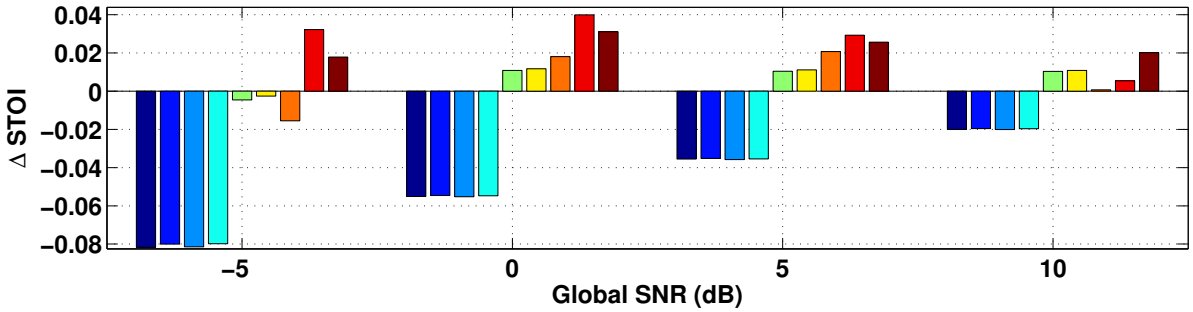

Figure 4.15: STOI-scores for babble noise

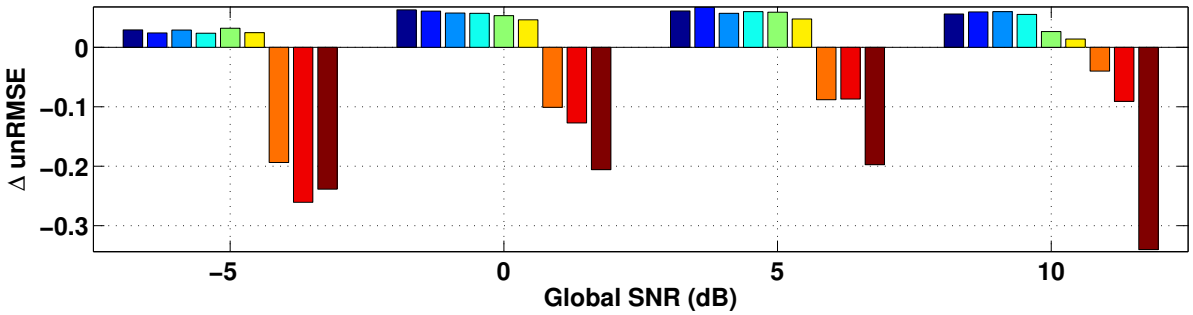

Figure 4.16: UnRMSE-scores for babble noise

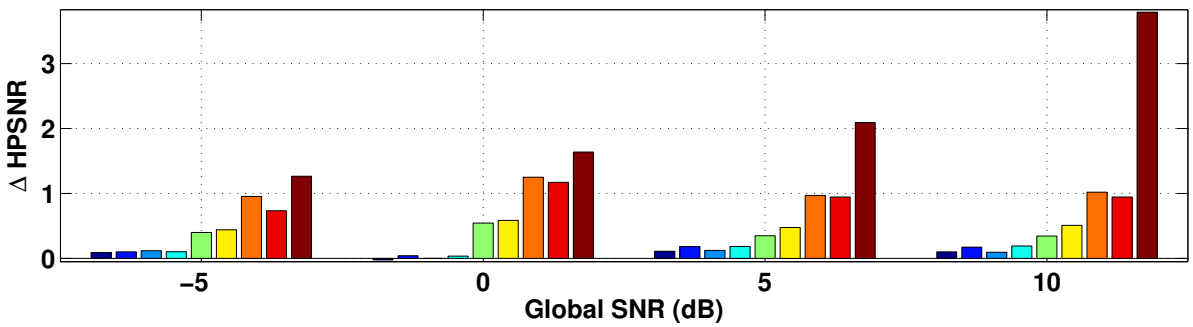

Figure 4.17: UnHPSNR-scores for babble noise

5 Conclusion

This thesis presented a joint MAP amplitude and phase estimator. In contrast to previous joint MAP estimators, the uniform prior distribution of phase was replaced by a von Mises distribution, yielding two interdependent estimators of phase and amplitude. Since a closed form solution for the estimator was not found, an iterative procedure has been proposed. To this end two different implementations have been examined, showing contradictory objective as well as subjective performance.

In the course of the analysis of the two estimators, the need of reliable information on the parameters of the phase prior distribution has been found to play a key role in the phase estimation procedure. Obtaining these parameters directly from the noisy observation appears to be an unfavourable approach, therefore future work may be directed towards the parameter estimation of the von Mises distribution. Nevertheless, the upper bounds of the proposed method very well pronounce the need of a reliable phase estimate, both for reconstruction and even more prominent for amplitude estimation.

As most amplitude-only STFT speech enhancement methods are only capable of improving the perceived quality, incorporating a phase estimate has the potential to improve intelligibility at the same time. This encourages to find new estimators that take the complex nature of speech DFT coefficients into account.

Furthermore, the *outer loop* implementation, which tends to over-harmonise the enhanced speech, reveals that state-of-the art objective quality measures are not capable of reliably predicting the perceived quality achieved by phase-aware speech enhancement. A strong buzzyness, surely degrading the quality (as confirmed by informal listening tests), helps to improve the PESQ score to a certain extent. This is emphasized by the fact that the PESQ upper-bound (clean phase for amplitude estimation and clean phase for reconstruction) is even exceeded by the *outer loop* implementation for certain SNRs and noise-types.

Bibliography

[1] D. L. Wang and J. S. Lim, "The Unimportance of Phase in Speech Enhancement", *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1982, vol. 30, no. 4, pp. 679-681.

[2] K. Paliwal, K. Wójcick and B. Shannon, "The importance of phase in speech enhancement", *Speech Communication*, 2011, vol. 53, pp. 465-494.

[3] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator", *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1984, vol. 32, no. 6, pp. 1109-1121.

[4] D. W. Griffin and J. S. Lim, "Signal Estimation from Modified Short-Time Fourier Transform", *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1984, vol. 32, no. 2, pp. 236-343.

[5] T. Lotter and P. Vary, "Speech Enhancement by MAP Spectral Amplitude Estimation Using a Super-Gaussian Speech Model", *EURASIP Journal on Applied Signal Processing*, 2005, no. 7, pp. 1110-1126.

[6] G. Degottex and D. Erro, "A uniform phase representation of the harmonic model in speech synthesis applications", *EURASIP Journal on Audio, Speech and Music Processing*, 2014, no. 38.

[7] A. Das and J. H. L. Hansen, "Phoneme Selective Speech Enhancement using Parametric Estimators and he Mixture Maximum Model: A Unifying Approach", *IEEE Transactions on Audio, Speech and Language Processing*, 2012, vol. 20, no. 8, pp. 2265-2279.

[8] P. J. Wolfe and S. J. Godsill, "Eficient Alternatives to the Ephraim and Malah Suppresion Rule for Audio Signal Enhancement", *EURASIP Journal on Applied Signal Processing*, 2003, no. 10, pp. 1043-1051.

[9] J. Kulmer and P. Mowlaee, "Harmonic Phase Estimation in Single-Channel Speech enhancement using Von Mises Distribution and Prior SNR", *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015.

[10] M. Krawczyk, R. Rehr and T. Gerkmann, "Phase-Sensitive Real-Time Capable Speech enhancement Under Voiced-Unvoiced Uncertainty", *Signal Processing Conference (EUSIPCO), IEEE*, 2014, pp. 1-5.

[11] T. Gerkmann, "Bayesian Estimation of Clean Speech Spectral Coefficients Given *a Priori* Knowledge of the Phase", *IEEE Transactions on Signal Processing*, 2014, vol. 62, no. 16, pp. 4199-4208.

[12] T. Gerkmann, "MMSE-Optimal Enhancement of Complex Speech Coefficients with uncertain prior knowledge of the clean speech phase", *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 4478-4482.

[13] T. Gerkmann and M. Krawczyk, "MMSE-Optimal Spectral Amplitude Estimation Given the STFT-Phase", *IEEE Signal Processing Letters*, 2013, vol. 20, no. 2, pp. 129-132.

[14] M. Krawczyk and T. Gerkmann, "STFT Phase Phase Improvement for Single Channel Speech Enhancement", *International Workshop on Acoustic Signal Enhancement*, 4-6 September,2012, Aachen.

[15] M. Krawczyk and T. Gerkmann, "STFT Phase Reconstruction in Voiced Speech for an Improved Single-Channel Speech Enhancement", *IEEE Transactions on Signal Processing*, 2014, vol. 22, no. 12, pp. 1931-1940.

[16] C. Breithaupt, T. Gerkmann and R. Martin, "A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing", *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 4897-4900.

[17] C. H. You, S. N. Koh and S. Rahardja, "β-Order MMSE Spectral Amplitude Estimation for Speech Enhancement", *IEEE Transactions on Speech and Audio Processing*, 2005, vol. 13, no. 4, pp. 475-486.

[18] R. C. Hendriks, T. Gerkmann and J. Jensen, "DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement - *A Survey of the State-of-the-Art*", *Morgan & Claypool Publishers*, 2013.

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

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

[21] J. E. V. Le Roux , Y. Mizuno, H. Kameoka, N. Ono and S. Sagayama, "Consistent Wiener filtering: generalized time-frequency masking respecting spectrogram consistency", *9th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2010, pp. 89-96.

[22] P. Vary, "Noise suppression by spectral magnitude estimation mechanism and theoretical limits", *Signal Procecss.* 1985, vol. 8, no. 4, pp. 387-400.

[23] P. Mowlaee and R. Saeidi, "On Phase Importance in Parameter Estimation in Single-Channel Speech Enhancement", *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7462-7466.

[24] G. Shi, M. M. Shanechi and P. Aarabi, "On the importance of Phase in human speech recognition", *IEEE Transactions on Audio, Speech and Language Processing*, 2006, vol.14, no. 5, pp. 1867-1874.

[25] J. Kulmer and P. Mowlaee, "Phase Estimation in Single Channel Speech Enhancement Using Phase Decomposition", *IEEE Signal Processing Letters*, 2015, vol. 22, no. 5, pp. 598-602.

[26] R. Martin, P. Vary, "Digital Speech Transmission - *Enhancement, Coding and Error Concealment*", *John Wiley & Sons, Ltd*, 2006.

[27] C. Breithaupt, M. Krawczyk and R.Martin, "Parametrized MMSE Spectral Magnitude Estimation for the Enhancement of Noisy Speech", *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 4037-4040.

[28] J. Taghia, "Bayesian Modeling of Directional Data with Acoustic and other Applications", *PhD Thesis, KTH Royal Institute of Technology*, 2014.

[29] P. Berens, "CircStat: A MATLAB Toolbox for Circular Statistics", *Journal of Statistiacal Software*, 2009, vol. 31, no. 10.

[30] A. Papoulis, S. U. Pillai, "Probability, Random Variables, and Stochastic Processes", Fourth Edition, *McGraw-Hill Higher Education*, 2002.

[31] R. J. McAulay and M. L. Malpass, "Speech Enhancement Using a Soft Decision Noise Suppresion Filter", *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1980, vol. 28, no. 2, pp. 137-145 .

[32] R. Martin, "Speech Enhancement Using MMSE Short Time Spectral Estimation with Gamma Distributed Speech Priors", *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2002, pp. 253-256.

[33] R. Martin and C. Breithaupt, "Speech Enhancement in the DFT Domain using Laplacian Speech Priors", *International Workshop on Acoustic Echo and Noise Control (IWAENC2003)*, 2003, pp. 87-90.

[34] J. S. Lim and A. V. Oppenheim, "All-Pole Modeling of Degraded Speech", *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1978, vol. 26, no. 3, pp. 197-210.

[35] M. R. Sambur and N. S. Jayant, "LPC Analysis/Synthesis from Speech Inputs Containing Quantizing Noise or Additive White Noise", *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1976, vol. 24, no. 6, pp. 488-494.

[36] J. H. L. Hansen and M. A. Clements, "Constrained Iterative Speech Enhancement with Application to Speech Recognition", *IEEE Transactions on Signal Processing*, 1991, vol. 39, no. 4, pp. 795-805.

[37] K. Funaki, Y. Miyanagy, K. Tochinai, "On a Time-Varying Complex Speech Analysis", *IX European Signal Processing Conference*, 1998.

[38] K. Funaki, "Speech Enhancement Based on Iterative Wiener Filter Using Complex LPC Speech Analysis", *Recent Advances in Signal Processing, Ashraf A Zaher (Ed.)*, 2009, ISBN: 978-953-307-002-5, InTech, Available from: http://www.intechopen.com/books/recent-advances-in-signal-processing/speech-enhancement- based-on-iterative-wiener-filter-using-complex-lpc-speech-analysis.

[39] P. Mowlaee, R. Saiedi and R. Martin, "Phase estimation for signal reconstruction in single-channel speech enhancement", *Proc. Int. Conference Spoken Language Processing*, 2012.

[40] A. Sugiyama, R. Miyahara, "Phase randomization - A new paradigm for single-channel signal enhancement", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7487-7491.

[41] 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*, 1985, vol. 33, no. 2, pp. 443-445.

[42] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-Based Noise Power Estimation With Low Complexity and Low Tracking Delay" *IEEE Transactions on Audio, Speech, Language Processing*, 2012, vol. 20, pp. 1383-1393.

[43] N. I. Fisher, "Statistical Analysis Of Circular Data." *Cambridge University Press, Cambridge,* UK 1995

[44] ITU-T, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs", P.862

[45] Y. Hu and P. C. Loizou, "Evaluation of Objective Quality Measures for Speech Enhancement", *IEEE Transactions on Audio, Speech and Language Processing*, 2008, vol.16, no. , pp. 229-238.

[46] A. Gaich and P. Mowlaee, "On Speech Quality Estimation of Phase-Aware Single-Channel Speech Enhancement", *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 216-220.

[47] A. Gaich and P. Mowlaee, "On Speech Intelligibility Estimation of Phase-Aware Single-Channel Speech Enhancement", *INTERSPEECH*, 2015.

[48] C. H. Taal, R. C. Hendriks, R. Heusdens and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech",

[49] B.J. Borgström and A. Alwan, "A Unified Framework for Designing Optimal STSA Estimators Assuming Maximum Likelihood Phase Equivalence of Speech and Noise", *INTERSPEECH*, 2011, vol. 19, no. 8 , pp. 2579-2590.

[50] P. Mowlaee and J. Kulmer, "Phase Estimation in Single Channel Speech Enhancement: Limits-Potential" , *IEEE Transactions on Audio, Speech and Language Processing*, 2015, vol. 23, no. 8, pp. 1283-1294.

[51] M. Cooke, J. Barker, S. Cunningham and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition", *Journal Acoustic Society of America*, 2006, vol. 120, no. 5.

[52] 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", *Technical Report, DRA Speech Research Unit*, 1992.

[53] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-Based Noise Power Estimation With Low Complexity and Low Tracking Delay" *IEEE Transactions on Audio, Speech, and Language Processing*, 2012, vol. 20, no. 4, pp. 1383-1393.