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

Timo Gerkmann, *Member, IEEE*

***Abstract*—While most short-time discrete Fourier transform-based single-channel speech enhancement algorithms only modify the noisy spectral amplitude, in recent years the interest in phase processing has increased in the field. The goal of this paper is twofold. First, we derive Bayesian probability density functions and estimators for the clean speech phase when different amounts of prior knowledge about the speech and noise amplitudes is given. Second, we derive a joint Bayesian estimator of the clean speech amplitudes and phases, when uncertain *a priori* knowledge on the phase is available. Instrumental measures predict that by incorporating uncertain prior information of the phase, the quality and intelligibility of processed speech can be improved both over traditional phase insensitive approaches, and approaches that treat prior information on the phase as deterministic.**

***Index Terms*—Noise reduction, phase estimation, signal reconstruction, speech enhancement.**

I. Introduction

THE enhancement of corrupted speech is a very important part of today's mobile communication devices, such as hearing aids or cell phones. This is because additive noise degrades the quality and also the intelligibility of speech. The goal of speech enhancement algorithms is to reduce the noise while preserving the speech signal. Especially when only one microphone signal is available this is a difficult task, and many proposals and improvements for single channel speech enhancement algorithms arose in the past decades. In this paper, we derive statistically optimal estimators for the clean speech spectral coefficients and clean speech spectral phases for single channel speech enhancement in the short time discrete Fourier transform (STFT)-domain. Examples for statistical estimation schemes are maximum likelihood (ML) estimation as well as the Bayesian maximum a posteriori (MAP) and minimum mean squared error (MMSE) estimators (see e.g., [1, sections 5.12 and 11.4]).

Manuscript received November 23, 2013; revised March 24, 2014; accepted June 27, 2014. Date of publication July 08, 2014; date of current version July 18, 2014. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Minyue Fu. This work has been funded by the DFG Research Grant GE 2538/2-1.

The author is with the Speech Signal Processing Group, Department of Medical Physics and Acoustics, Cluster of Excellence "Hearing4all," Universität Oldenburg, 26111 Oldenburg, Germany (e-mail: timo.gerkmann@uni-oldenburg.de; website: http://www.speech.uni-oldenburg.de).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2014.2336615

The role of the phase has been widely discussed, both in the signal model [2], and also for single channel speech enhancement [3], [4]. Lotter and Vary [5] as well as Erkelens *et al.* [6] showed that complex speech coefficients can be well modeled by a circular symmetric probability density function (PDF). The assumption of a circular symmetric PDF implies that the phase is uniformly distributed, from which it follows that the MMSE-optimal estimate of the clean phase is in fact the noisy phase [7]–[9]. Consequently, under the assumption of a circular symmetric distribution, MMSE estimators of complex speech coefficients alter only the magnitude of the noisy spectral coefficients while the noisy phase remains unchanged. The most prominent example is the Wiener filter (e.g., [1, Section 11.4.3]). Furthermore, in perceptual experiments conducted by Wang and Lim [3], it was shown that improving the phase does not improve the performance of single channel speech enhancement algorithms. Also for these reasons, most subsequent research addressed an improved estimation of spectral amplitudes, while leaving the phase of the noisy signal unchanged.

More recently, Paliwal *et al.* again conducted experiments where the noisy phase is exchanged by the clean speech phase in speech enhancement algorithms [10]. However, as opposed to [3], they used a more redundant spectral representation in the STFT-domain (by means of a larger segment overlap and zero-padding) and could now show that employing the clean speech phase improves noise reduction algorithms significantly. From this it follows that if we have an estimate of the clean phase, potentially, we can also improve speech enhancement algorithms. However, blind estimators of the clean speech phase are rare. When the clean speech amplitude is known, Griffin and Lim showed that the clean speech phase can be reconstructed by iterative STFT analysis and synthesis [11]. Many variants and improvements for iterative phase estimation have been proposed over the years. A nice overview is given in [12]. Recent improvements are given for instance in [13], [14]. In [15], [16] we follow a different approach and propose to reconstruct the clean speech phase on and between speech spectral harmonics based on a sinusoidal model of voiced speech. The basic idea of [15] is that in the STFT-domain the difference between the harmonic frequencies of voiced speech and the STFT-band center-frequencies is captured by the phase. Thus, with the algorithm proposed in [15] the STFT-phase is reconstructed from only the noisy observation and an estimate of the speech fundamental frequency. A general problem with both iterative and sinusoidal model based phase enhancement is that erroneous phase estimates may yield annoying artifacts in the synthesized speech signal [13], [15]. However, in [17] we showed that the phase estimate can also be employed only to improve Bayesian

1053-587X © 2014 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

amplitude estimation, which is less prone to artifacts. Therefore, in [17], [18] we propose to employ the phase estimate only for an improved amplitude estimation but to still use the phase of the noisy observation when going from the STFT-domain back to time-domain.

Of course, it is desirable to obtain a phase estimate that can be robustly combined with the enhanced amplitudes without introducing artifacts. To achieve this, the key idea of this paper is to incorporate uncertain prior knowledge about the phase by means of a Bayesian phase enhancement framework. The prior information about the phase can for instance be obtained using the sinusoidal model based phase reconstruction algorithm [15], [16].

This paper is structured as follows. We start by discussing phase estimation when prior knowledge on the spectral amplitudes of speech and noise are given (Section II). In the remainder, we drop this requirement and treat both speech and noise as unknown random variables. While in Section III we derive ML and MAP estimators of the clean speech phase, in Section IV, we derive the joint MMSE estimator of the clean speech amplitude and phase when uncertain prior knowledge of the clean speech spectral phase is given. In Section V, we evaluate the clean speech estimator proposed in Section IV for different signal to noise ratios (SNRs) and noise types.

II. Phase Estimation With Known Spectral Amplitudes

This section starts by defining the signal model and notation. Then, we discuss the case of phase estimation when both the speech and noise spectral amplitudes are known. Afterwards, we derive the phase posterior under a Gaussian noise model for known speech spectral amplitudes.

A. Signal Model and Notation

We assume that the complex STFT coefficients of the noisy speech Y are given by an additive superposition of uncorrelated zero-mean speech and noise coefficients, S and N, as

$$Y_k(\ell) = S_k(\ell) + N_k(\ell), \tag{1}$$

where ℓ is the segment index and k is the frequency index. In the sequel, we omit the time index ℓ and frequency index k unless needed. The complex coefficients can be represented by their amplitudes and phases denoted as

$$Y = R\mathrm{e}^{\mathrm{j}\Phi_Y}; \qquad S = A\mathrm{e}^{\mathrm{j}\Phi_S}; \qquad N = V\mathrm{e}^{\mathrm{j}\Phi_N}. \tag{2}$$

Furthermore, we denote random variables by capital letters, e.g., S, A, Φ_S, and their realizations by the corresponding lower case letters, e.g., s, a, ϕ_S.

B. Phase Estimation Given the Noisy Speech as Well as the Speech and Noise Amplitudes

It is interesting to note that even in the idealistic case of known speech and noise amplitudes, the clean speech phase can not be uniquely obtained. In fact, there exist two possible solutions for the phase. This is illustrated in Fig. 1, where the possible realizations of the complex clean speech and noise coefficients are shown. As the corresponding amplitudes are assumed to be known, possible realizations of the complex speech and noise coefficients are given by circles. As furthermore it must hold that $y = s + n$, two solutions for the speech and noise phases are valid. These two solutions are given by the intersections of the two circles in Fig. 1. This phase ambiguity can be resolved for instance by minimizing the group delay [19]. However, assuming that the speech and noise spectral amplitudes are known is a rather limiting assumption, as in practice only estimates are available.

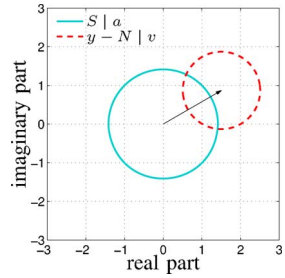


Fig. 1. Possible realizations of the complex speech and noise coefficients, S and N, when the speech and noise amplitudes a and v are known while their phases ϕ_S, ϕ_N are unknown (Section II-B). In this example we chose $a = \sqrt{2}$ and $v = 1$. The solid (blue) circle represents possible locations for the complex speech coefficients, the black arrow represents the complex noisy observation $y = \sqrt{3}\exp(\mathrm{j}30°)$, and the dashed (red) circle represents possible locations for the complex noise-related coefficients $y - N$. As per definition $Y = S + N$, two possible realizations for the speech and noise complex coefficients exist, given by the two intersections of the circles, i.e., when $S = y - N$.

C. Phase Posterior for a Known Speech Amplitude

Now, we start relaxing the requirements on prior amplitude knowledge and model the noise N as a complex Gaussian random variable with given variance, while the speech amplitudes are still treated as being known. We show that under these assumptions the speech phase posterior is given by a *von Mises* distribution. In Fig. 2 we illustrate the considered scenario of a known spectral amplitude a but complex Gaussian distributed noise coefficients N. Again, the black arrow represents the noisy observation, and the solid circle represents the possible realizations of the complex speech coefficients $(S \mid a)$, i.e., when the speech amplitude a is known, but the speech phase is unknown. The scatter plot represents $(y - N)$ where the complex noise coefficients N are sampled from a circular complex Gaussian distribution. Valid representations of N are those that fulfill $S = y - N$, i.e., those dots of the scatter plot that lie on the solid circle. Tracking the solid circle by eye, one may already see that most of the scattered dots on the solid circle are in the direction of the noisy observation. Thus, we expect that also the phase distribution $p_{\Phi_S|a,y}(\phi_S \mid a, y)$ will not be uniform. In this section, we show that $p_{\Phi_S|a,y}(\phi_S \mid a, y)$ results in a *von Mises* distribution with mean-direction ϕ_Y.

To derive the PDF of the speech phase when the clean speech amplitude and the noisy observation are given, we employ Bayes' theorem as

$$p_{\Phi_S|a,y} = \frac{p_{A,\Phi_S,Y}}{p_{A,Y}} = \frac{p_{A,\Phi_S|y}}{p_{A|y}}. \tag{3}$$

Assuming a complex Gaussian distribution for both the speech and noise spectral coefficients, the numerator and denominator of the right hand side of (3) are well known: the numerator is

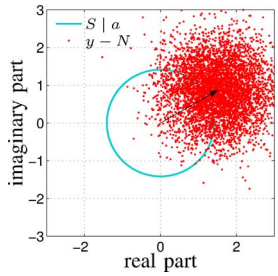


Fig. 2. Possible realizations of the complex speech and noise coefficients, S and N, when the speech amplitude a is known while its phase ϕ_S is unknown and N is modeled by a complex Gaussian (Section II-C). In this example we chose $a = \sqrt{2}$ while the noise variance is $\sigma_{\mathrm{N}}^2 = 1$. The solid (blue) circle represents possible locations for the complex speech coefficients, the black arrow represents the complex noisy observation $y = \sqrt{3}\exp(\mathrm{j}30°)$, and the scatter plot represents possible locations for the complex noise-related coefficients $y - N$.

recognized as the Gaussian posterior function $p_{S|y}(s \mid y)$ transformed to the polar domain, while the denominator is recognized as a *Rice* distribution. To see this, we start with the speech posterior $p_{S|y}$ and Bayes' theorem

$$p_{S|y}(s \mid y) = \frac{p_{Y|s}(y \mid s)\, p_S(s)}{p_Y(y)}. \tag{4}$$

To model the PDFs in (4), under the Gaussian assumption with noise power spectral density (PSD) $\sigma_{\mathrm{N}}^2 = \mathrm{E}\big(|N|^2\big)$, the likelihood of complex speech coefficients can be written as (e.g., [1])

$$p_{Y|s}(y \mid s) = \frac{1}{\pi\sigma_{\mathrm{N}}^2} \exp\left(-\frac{|y-s|^2}{\sigma_{\mathrm{N}}^2}\right), \tag{5}$$

and with $\sigma_S^2 = \mathrm{E}\big(|S|^2\big)$ the speech prior distribution can be written as

$$p_S(s) = \frac{1}{\pi\sigma_S^2} \exp\left(-\frac{|s|^2}{\sigma_S^2}\right). \tag{6}$$

With these assumptions, also the evidence is complex Gaussian distributed, as

$$p_Y(y) = \frac{1}{\pi\sigma_Y^2} \exp\left(-\frac{|y|^2}{\sigma_Y^2}\right), \tag{7}$$

where we assume that speech and noise are uncorrelated, such that $\sigma_Y^2 = \mathrm{E}\big(|Y|^2\big) = \sigma_S^2 + \sigma_{\mathrm{N}}^2$. Using (5)–(7) in (4), after some basic algebraic computations, we also obtain a complex Gaussian distribution for the posterior of the clean speech complex coefficients (e.g., [20]), as

$$p_{S|y}(s \mid y) = \frac{1}{\pi\lambda} \exp\left(-\frac{|s - G_{\mathrm{W}} y|^2}{\lambda}\right), \tag{8}$$

with mean $\mathrm{E}(S \mid y) = G_{\mathrm{W}} y$, Wiener's gain-function

$$G_{\mathrm{W}} = \frac{\sigma_S^2}{\sigma_S^2 + \sigma_{\mathrm{N}}^2}, \tag{9}$$

and variance

$$\lambda = \frac{\sigma_S^2 \sigma_{\mathrm{N}}^2}{\sigma_S^2 + \sigma_{\mathrm{N}}^2}. \tag{10}$$

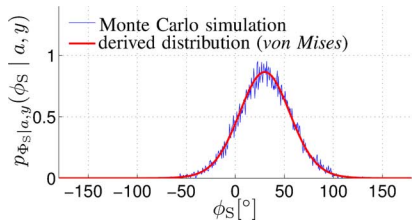


Fig. 3. Monte Carlo simulation of $p_{\Phi_S|a,y}(\phi_S \mid a, y)$ and the derived *von Mises* distribution (13). An excellent fit can be observed. As in Fig. 2 $\phi_Y = 30°$, $a = \sqrt{2}$, $\sigma_{\mathrm{N}}^2 = 1$, and we set $\sigma_S^2 = 2$.

After transforming (8) to the polar domain, i.e., $p_{A,\Phi_S|y} = A\, p_{S|y}$ we obtain our model for the numerator of (3)

$$p_{A,\Phi_S|y}(a, \phi_S \mid r, \phi_Y) = \\ = \frac{a}{\pi\lambda} \exp\left(-\frac{a^2 + G_{\mathrm{W}}^2 r^2 - 2aG_{\mathrm{W}} r\cos(\phi_S - \phi_Y)}{\lambda}\right). \tag{11}$$

The denominator of (3), $p_{A|y}$, can be obtained by integrating (11) over the phase ϕ_S. Then, with [21, eq. (3.339)] we obtain the well-known *Rice* distribution

$$p_{A|y}(a \mid r) = \frac{2a}{\lambda} \exp\left(-\frac{a^2 + G_{\mathrm{W}}^2 r^2}{\lambda}\right) I_0\left(\frac{2aG_{\mathrm{W}} r}{\lambda}\right) \tag{12}$$

with $I_\nu(\cdot)$ the modified Bessel function of the first kind [21, eq. (8.445)].

Finally, dividing (11) by (12), we can solve (3), and obtain the phase posterior for a known speech amplitude:

$$\boxed{p_{\Phi_S|a,y}(\phi_S \mid a, r, \phi_Y) = \\ = \tfrac{1}{2\pi} \exp\Big(\tfrac{2ar}{\sigma_{\mathrm{N}}^2}\cos(\phi_S - \phi_Y)\Big) \,/\, I_0\big(2ar/\sigma_{\mathrm{N}}^2\big)}. \tag{13}$$

Thus, for a known speech amplitude, the clean speech phase posterior (13) is recognized as a *von Mises* distribution [22]

$$p_{\Phi_S|\widetilde{\phi},\kappa}\Big(\phi_S \mid \widetilde{\phi}, \kappa\Big) = \exp\Big(\kappa\cos(\phi_S - \widetilde{\phi})\Big) \,/\, (2\pi I_0(\kappa)) \tag{14}$$

with mean-direction $\widetilde{\phi} = \phi_Y$ and the concentration parameter $\kappa = 2ar/\sigma_{\mathrm{N}}^2$, where the circular variance decreases with increasing concentration κ. In Fig. 3 we demonstrate the validity of our derivations by comparing the derived posterior to a Monte Carlo simulation. The parameters are the same as in Fig. 2. Note that the *von Mises* distribution is symmetric around its mean-direction, and the mean-direction is also the mode of the *von Mises* distribution. Thus, for a known speech amplitude both the MAP-optimal estimate of the clean speech phase as well as the MMSE-optimal estimate of the clean speech phase are given by the noisy phase.

III. Phase Estimation for Unknown Speech and Noise Amplitudes

As in practice, the assumption of known speech spectral amplitudes is often not fulfilled, in the remainder of this paper our focus is on statistical estimators that treat both speech and noise amplitudes as unknown random variables. This situation is illustrated in Fig. 4. In this section, we first derive the ML estimator

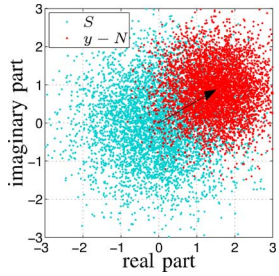


Fig. 4. Illustration of the situation that the spectral amplitudes a and v are unknown (Sections III, IV–V). The light (blue) scatter plot represents realizations of the complex Gaussian random variable S with variance $\sigma_S^2 = 2$, while the darker (red) scatter plot represents realizations of the complex Gaussian random variable $y - N$ with variance $\sigma_N^2 = 1$. The arrow indicates the noisy observation $y = \sqrt{3}\exp(\mathrm{j}30^\circ)$.

of the clean speech phase and then combine the derived likelihood with uncertain prior knowledge on the clean speech phase to obtain a MAP-optimal phase estimator.

A. Maximum Likelihood Phase Estimation

To derive the phase likelihood $p_{Y|\phi_S}(y \mid \phi_S)$, we again employ Bayes' theorem, as

$$p_{Y|\phi_S}(y \mid \phi_S) = \frac{p_{Y,\Phi_S}(y, \phi_S)}{p_{\Phi_S}(\phi_S)} = \frac{\int_0^\infty p_{Y|s}(y \mid \phi_S, a)\, p_{\Phi_S,A}(\phi_S, a)\mathrm{d}a}{p_{\Phi_S}(\phi_S)}. \tag{15}$$

As empirically shown in [5], [6], the PDF of isolated complex speech coefficients is circular symmetric, meaning that amplitudes and phases are independent, i.e., $p_{\Phi_S,A}(\phi_S, a) = p_{\Phi_S}(\phi_S)\, p_A(a)$. Then (15) simplifies to:

$$p_{Y|\phi_S}(y \mid \phi_S) = \int_0^\infty p_{Y|s}(y \mid \phi_S, a)\, p_A(a)\mathrm{d}a. \tag{16}$$

Assuming a complex Gaussian distribution for the noise spectral coefficients, the PDF $p_{Y|a,\phi_S} = p_{Y|s}$ is also Gaussian and given in (5). As in [23], [24] we propose to model the speech spectral amplitudes by a χ-distribution

$$p_A(a) = \frac{2}{\Gamma(\mu)} \left(\frac{\mu}{\sigma_S^2}\right)^\mu a^{2\mu-1} \exp\left(-\frac{\mu}{\sigma_S^2} a^2\right), \tag{17}$$

with the Gamma function $\Gamma(\cdot)$ [21, Eq. (8.31)], and shape parameter μ. Setting $\mu < 1$ allows us to model heavy-tailed (i.e., super-Gaussian) speech priors. Note that for complex Gaussian distributed speech coefficients, the speech amplitudes are Rayleigh distributed, which is the special case of the χ-distribution when $\mu = 1$. Using the χ-distribution (17) and the Gaussian distribution (5) in (16), we obtain

$$p_{Y|\phi_S}(r, \phi_Y \mid \phi_S) = \int_0^\infty \frac{2}{\Gamma(\mu)} \frac{1}{\pi\sigma_N^2} \left(\frac{\mu}{\sigma_S^2}\right)^\mu a^{2\mu-1} \exp\left(-\frac{\sigma_S^2 r^2 + (\sigma_S^2 + \mu\sigma_N^2)a^2 - 2ar\sigma_S^2\cos(\Delta\phi))}{\sigma_S^2\sigma_N^2}\right) \mathrm{d}a \tag{18}$$

with $\Delta\phi = \phi_S - \phi_Y$. Note that in our notation $p_{Y|\phi_S}(r, \phi_Y \mid \phi_S)$ in (18) resembles the conditioned joint PDF of the real and imaginary parts of Y, written as a function of the amplitudes r and phases ϕ_Y. The integral in (18) can be solved with [21, eq. (3.462.1)] yielding the likelihood

$$\boxed{p_{Y|\phi_S}(r, \phi_Y \mid \phi_S) = \frac{2^{1-\mu}}{\pi\sigma_N^2} \left(\frac{\mu}{\mu+\xi}\right)^\mu \frac{\Gamma(2\mu)}{\Gamma(\mu)} \exp\left(-\frac{r^2}{\sigma_N^2}\right) \exp\left(\frac{\nu^2}{4}\right) D_{(-2\mu)}(\nu)} \tag{19}$$

with

$$\nu = -\frac{r}{\sigma_N}\sqrt{2\frac{\xi}{\mu+\xi}}\cos(\underbrace{\phi_S - \phi_Y}_{\Delta\phi}), \tag{20}$$

the parabolic cylinder function [21, Eq. (9.241.2)]

$$\mathrm{D}_{(-2\mu)}(\nu) = \frac{\exp\left(-\nu^2/4\right)}{\Gamma(2\mu)} \int_0^\infty \mathrm{e}^{-x\nu - \frac{x^2}{2}} x^{2\mu-1}\mathrm{d}x, \tag{21}$$

and the *a priori* SNR $\xi = \sigma_S^2/\sigma_N^2$. Note that the argument ν contains the phase difference $\phi_S - \phi_Y$.

In Fig. 5, to demonstrate the validity of our derivation, we compare the derived distribution to a Monte Carlo simulation; an excellent fit can be observed. It can be seen that for large SNRs the maximum of the likelihood is in the direction of the clean speech phase. However clearly, for low SNRs, the likelihood contains less information about the clean speech phase and asymptotically approaches a circular distribution. With the likelihood at hand, we can formulate the ML estimator of the clean speech phase

$$\widehat{\phi_S}^{\mathrm{ML}} = \underset{\phi_S}{\operatorname{argmax}} \exp\left(\nu^2/4\right) D_{(-2\mu)}(\nu), \tag{22}$$

where we dropped all factors in (19) that are independent of ϕ_S. For $\mu > 0.1025$ the parabolic cylinder function $D_{(-2\mu)}(\nu)$ is a positive monotonically decreasing function of ν [25]. The factor $\exp\left(\nu^2/4\right)$ is also positive and increases exponentially with ν^2. Thus, the ML solution is given by the lowest negative ν, i.e., when the cosine in (20) is maximized. From this it follows that the ML-optimal estimator is the phase of the noisy signal ϕ_Y. As the factor $\exp\left(\nu^2/4\right)$ increases rapidly in ν, we can also relax the restriction $\mu > 0.1025$ meaning that

$$\widehat{\phi_S}^{\mathrm{ML}} = \phi_Y \tag{23}$$

holds for any $\mu > 0$.

B. Maximum a Posteriori Phase Estimation

In this section, we employ the derived likelihood function (19) to formulate a posterior distribution that incorporates a priori knowledge on the speech spectral phase. With this posterior we can then obtain a MAP estimate of the clean speech spectral phase which yields a trade-off between the noisy phase ϕ_Y and the mean-direction of the phase prior distribution, denoted by $\widetilde{\phi_S}$. The mean-direction $\widetilde{\phi_S}$ can be obtained from blind phase reconstruction algorithms such as [15], [16]. Another option could be to first obtain a phase estimate using consistent Wiener filtering [14], and to then employ it in the proposed scheme as uncertain prior phase information. However, the conceptual advantage of using [15] in this context is that we do not need an

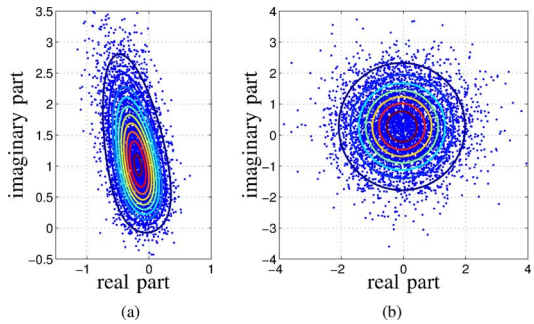


Fig. 5. The scatter plot results from a Monte Carlo simulation of the noisy observation when the clean speech phase is given. The contour plots represent the derived distribution for the likelihood $p_{Y|\phi_S}$ in (19) with $\mu = 1$. (a) $\sigma_N^2 = 0.1, \sigma_S^2 = 2, \phi_S = 100°$. (b) $\sigma_N^2 = 2, \sigma_S^2 = 0.1, \phi_S = 100°$.

estimate of the speech amplitudes to obtain the prior phase information. Instead the prior phase information is obtained from only the noisy observation and an estimate of the speech fundamental frequency, which can be obtained, e.g., using [26].

Given prior information $\widetilde{\phi_S}$ on the phase ϕ_S, with Bayes' theorem the posterior distribution can be written as

$$\begin{aligned} &p_{\Phi_S|\widetilde{\phi_S},y}\left(\phi_S \mid \widetilde{\phi_S}, y\right) \\ &\quad = \frac{p_{\Phi_S,\widetilde{\Phi_S},Y}\left(\phi_S, \widetilde{\phi_S}, y\right)}{\int_0^{2\pi} p_{\Phi_S,\widetilde{\Phi_S},Y}\left(\phi_S, \widetilde{\phi_S}, y\right) \mathrm{d}\phi_S} \\ &\quad = \frac{p_{Y|\phi_S}(y \mid \phi_S)\, p_{\Phi_S|\widetilde{\phi_S}}\left(\phi_S \mid \widetilde{\phi_S}\right)}{\int_0^{2\pi} p_{Y|\phi_S}(y \mid \phi_S)\, p_{\Phi_S|\widetilde{\phi_S}}\left(\phi_S \mid \widetilde{\phi_S}\right) \mathrm{d}\phi_S}. \end{aligned} \quad (24)$$

As, after the integration over ϕ_S, the denominator of (24) is not a function of ϕ_S anymore, we only need to maximize the numerator of (24) to obtain the MAP-optimal phase estimate. For this, we need a model for the phase prior $p_{\Phi_S|\widetilde{\phi_S}}\left(\phi_S \mid \widetilde{\phi_S}\right)$. We propose to model this prior by a *von Mises* distribution which, for a given angular mean and concentration, is the maximum entropy distribution [22, Section 3.5.4]. Leaving out all quantities that are independent of ϕ_S, we obtain the MAP estimator

$$\begin{aligned} \widehat{\phi_S}^{\mathrm{MAP}} &= \underset{\phi_S}{\operatorname{argmax}}\, p_{\Phi_S|\widetilde{\phi_S},y}\left(\phi_S \mid \widetilde{\phi_S}, r, \phi_Y\right) \\ &= \arg\max_{\phi_S} \exp\left(\frac{\nu^2}{4}\right) D_{(-2\mu)}(\nu) \exp\left(\kappa \cos(\phi_S - \widetilde{\phi_S})\right) \end{aligned} \quad (25)$$

with ν given in (20). The concentration parameter κ in (14) can be used to incorporate the uncertainty of the prior phase information $\widetilde{\phi_S}$. In this context, $\kappa \to \infty$ implies a large certainty about the prior phase information, and as a consequence, the MAP estimator yields $\widehat{\phi_S}^{\mathrm{MAP}} = \widetilde{\phi_S}$. On the contrary, $\kappa = 0$ implies a large uncertainty about the prior phase information and the MAP estimator yields the noisy phase $\widehat{\phi_S}^{\mathrm{MAP}} = \widehat{\phi_S}^{\mathrm{ML}} = \phi_Y$. For any other $0 < \kappa < \infty$ the MAP-estimator yields values between the phase of the noisy observation $\phi_Y = \widehat{\phi_S}^{\mathrm{ML}}$ and the prior phase information $\widetilde{\phi_S}$. Furthermore, for a fixed κ, the influence of the phase prior is SNR dependent. As can be seen in Fig. 5, for lower SNRs the phase likelihood is less informative than for higher SNRs. As a result, for low SNRs, the phase prior has a stronger influence on the MAP estimator than for higher SNRs. While analytically solving (25) for $\widehat{\phi_S}^{\mathrm{MAP}}$ is difficult, we can find the maximum for instance by a brute-force search over a set of candidate phases. The price, however, is an increased computational complexity.

IV. Joint Amplitude and Phase Estimation

While in the previous sections we looked at phase estimation independently of amplitude estimation, in this section we aim at estimating the clean speech amplitude and phase jointly in order to get an estimator of the (C)omplex spectral speech coefficients given (U)ncertain (P)hase information (CUP). While the basic idea of the CUP estimator has been published in [27], here we present a more detailed analysis, derivation, and evaluation.

A CUP estimator can be obtained by solving $\mathrm{E}\left(S \mid Y, \widetilde{\phi_S}\right)$, where $\widetilde{\phi_S}$ denotes a priori information of the clean speech phase. Again, this prior phase information can be obtained using the phase reconstruction algorithm [15], [16]. Instead of finding the MMSE estimate of the complex speech coefficients, Ephraim and Malah [28] argued that estimating logarithmically compressed spectral amplitudes is perceptually advantageous. You *et al.* [29] generalized the logarithmic amplitude compression [28] by employing a compression parameter β. As in [17], [24] we adopt this idea, but now derive a joint estimator of compressed speech amplitudes and the clean speech phase, as

$$\begin{aligned} \widehat{S}^{(\beta)} &= \mathrm{E}\left(A^\beta \mathrm{e}^{\mathrm{j}\Phi_S} \mid y, \widetilde{\phi_S}\right) \\ &= \int_0^\infty \int_0^{2\pi} a^\beta \mathrm{e}^{\mathrm{j}\phi_S}\, p_{A,\Phi_S|y,\widetilde{\phi_S}}\left(a, \phi_S \mid y, \widetilde{\phi_S}\right) \mathrm{d}\phi_S \mathrm{d}a. \end{aligned} \quad (26)$$

In order to solve (26), we need to model the joint posterior of the amplitude and phase given the prior phase information $\widetilde{\phi_S}$. This joint posterior $p_{A,\Phi_S|y,\widetilde{\phi_S}}$ can be rewritten with Bayes' theorem, as

$$\begin{aligned} &p_{A,\Phi_S|y,\widetilde{\phi_S}}\left(a, \phi_S \mid y, \widetilde{\phi_S}\right) \\ &= \frac{p_{Y,A,\Phi_S,\widetilde{\Phi_S}}\left(y, a, \phi_S, \widetilde{\phi_S}\right)}{p_{Y,\widetilde{\Phi_S}}\left(y, \widetilde{\phi_S}\right)} \\ &= \frac{p_{Y|a,\phi_S,\widetilde{\phi_S}}\left(y \mid a, \phi_S, \widetilde{\phi_S}\right) p_{A,\Phi_S,\widetilde{\Phi_S}}\left(a, \phi_S, \widetilde{\phi_S}\right)}{\iint p_{Y|a,\phi_S,\widetilde{\phi_S}}\left(y \mid a, \phi_S, \widetilde{\phi_S}\right) p_{A,\Phi_S,\widetilde{\Phi_S}}\left(a, \phi_S, \widetilde{\phi_S}\right) \mathrm{d}a \mathrm{d}\phi_S} \end{aligned} \quad (27)$$

Thus, to find a model for $p_{A,\Phi_S|y,\widetilde{\phi_S}}$, we need models for $p_{Y|a,\phi_S,\widetilde{\phi_S}}$ and $p_{A,\Phi_S,\widetilde{\Phi_S}}$. To find a model for $p_{Y|a,\phi_S,\widetilde{\phi_S}}$, we assume that if the clean speech phase ϕ_S is given, the speech phase prior $\widetilde{\phi_S}$ gives no further information on Y, i.e.,

$$p_{Y|a,\phi_S,\widetilde{\phi_S}} = p_{Y|a,\phi_S} = p_{Y|s}. \quad (28)$$

As the noise coefficients are assumed to be complex Gaussian distributed, the PDF $p_{Y|s}$ is also Gaussian and given in (5). From the observation that complex speech coefficients exhibit

a circular symmetric distribution, it follows that amplitudes and phases are independent [6]. Then, the joint PDF of the clean speech amplitude, phase, and the phase estimate can be rewritten as

$$\begin{aligned} p_{A,\Phi_S,\widetilde{\Phi_S}}\Big(a,\phi_S,\widetilde{\phi_S}\Big) &= p_A(a)\, p_{\Phi_S,\widetilde{\Phi_S}}\Big(\phi_S,\widetilde{\phi_S}\Big) \\ &= p_A(a)\, p_{\Phi_S|\widetilde{\phi_S}}\Big(\phi_S \mid \widetilde{\phi_S}\Big)\, p_{\widetilde{\Phi_S}}\Big(\widetilde{\phi_S}\Big). \end{aligned} \tag{29}$$

Using (28) and (29) in (27) the posterior results in

$$\begin{aligned} &p_{A,\Phi_S|y,\widetilde{\phi_S}}\Big(a,\phi_S \mid y,\widetilde{\phi_S}\Big) \\ &= \frac{p_{Y|s}(y \mid a,\phi_S)\, p_A(a)\, p_{\Phi_S|\widetilde{\phi_S}}\Big(\phi_S \mid \widetilde{\phi_S}\Big)\, p_{\widetilde{\Phi_S}}\Big(\widetilde{\phi_S}\Big)}{\int\int\, p_{Y|s}(y \mid a,\phi_S)\, p_A(a)\, p_{\Phi_S|\widetilde{\phi_S}}(\phi_S)\, p_{\widetilde{\Phi_S}}\Big(\widetilde{\phi_S}\Big)\mathrm{d}a\mathrm{d}\phi_S} \\ &= \frac{p_{Y|s}(y \mid a,\phi_S)\, p_A(a)\, p_{\Phi_S|\widetilde{\phi_S}}\Big(\phi_S \mid \widetilde{\phi_S}\Big)}{\int\int\, p_{Y|s}(y \mid a,\phi_S)\, p_A(a)\, p_{\Phi_S|\widetilde{\phi_S}}\Big(\phi_S \mid \widetilde{\phi_S}\Big)\mathrm{d}a\mathrm{d}\phi_S} \end{aligned} \tag{30}$$

where $p_{\widetilde{\Phi_S}}\Big(\widetilde{\phi_S}\Big)$ is canceled out, as it is not part of the integral. Using (30) in (26) results in

$$\mathrm{E}\Big(A^{\beta}\mathrm{e}^{\mathrm{j}\Phi_S} \mid y,\widetilde{\phi_S}\Big) = \frac{\int_0^{2\pi}\mathrm{e}^{\mathrm{j}\phi_S}\int_0^{\infty} a^{\beta}\, p_{Y|a,\phi_S}\, p_A \mathrm{d}a\, p_{\Phi_S|\widetilde{\phi_S}}\mathrm{d}\phi_S}{\int_0^{2\pi}\int_0^{\infty}\, p_{Y|a,\phi_S}\, p_A \mathrm{d}a\, p_{\Phi_S|\widetilde{\phi_S}}\mathrm{d}\phi_S}. \tag{31}$$

As in [17], [23], [24] and Section III-B we model the speech amplitudes $p_A(a)$ to be χ-distributed (17) as this allows us to model heavy-tailed speech priors by setting $\mu < 1$. As in Section III-B, the distribution of the clean speech phase around the prior phase information $\widetilde{\phi_S}$ can be modeled by the *von Mises* distribution (14), where the concentration parameter κ in (14) allows incorporating the uncertainty about the prior phase information $\widetilde{\phi_S}$.

With (5), (17), and [21, Eq. (3.462.1)], similar to [17] the integral over the amplitude can be solved and we get the CUP:

$$\boxed{\begin{aligned} &\widehat{S}^{(\beta)} = \mathrm{E}\Big(A^{\beta}\mathrm{e}^{\mathrm{j}\Phi_S} \mid y,\widetilde{\phi_S}\Big) \\ &= \left(\sqrt{\tfrac{1}{2}\tfrac{\xi}{\mu+\xi}}\sigma_{\mathrm{N}}^2\right)^{\beta}\frac{\Gamma(2\mu+\beta)}{\Gamma(2\mu)} \\ &\times\frac{\int_0^{2\pi}\mathrm{e}^{\mathrm{j}\phi_S}\,\exp\!\big(\nu^2/4\big)\mathrm{D}_{(-2\mu-\beta)}(\nu)\; p_{\Phi_S|\widetilde{\phi_S}}\mathrm{d}\phi_S}{\int_0^{2\pi}\exp(\nu^2/4)\mathrm{D}_{(-2\mu)}(\nu)\, p_{\Phi_S|\widetilde{\phi_S}}\mathrm{d}\phi_S} \end{aligned}}, \tag{32}$$

where ν is defined as in (20). As in (19), $\mathrm{D}_{(\cdot)}(\nu)$ is the parabolic cylinder function, $\xi = \sigma_S^2/\sigma_{\mathrm{N}}^2$ is the *a priori* SNR, and the argument ν contains the phase difference $\phi_Y - \phi_S = \Delta\phi$. The speech-estimate is then obtained as

$$\widehat{S} = \Big|\widehat{S}^{(\beta)}\Big|^{1/\beta}\,\frac{\widehat{S}^{(\beta)}}{\Big|\widehat{S}^{(\beta)}\Big|} = \widehat{A}\mathrm{e}^{\mathrm{j}\widehat{\Phi_S}}. \tag{33}$$

A. Implementation of the Proposed CUP Estimator

Solving the integral over the speech spectral phase ϕ_S in (32) is quite difficult, as it involves the integration over the parabolic cylinder function. However, as the phase has a limited span between $0 \leq \phi_S < 2\pi$, the integral in (32) can be solved numerically with high precision. Furthermore, in practice, speech enhancement gain functions that involve computationally complex special functions are commonly precomputed and tabulated anyways. Thus, we propose to solve the integral in (32) numerically and store the result in a table. For a given shape parameter μ and compression parameter β, this table has four dimensions, the *a priori* SNR ξ, the *a posteriori* SNR $r^2/\sigma_{\mathrm{N}}^2$, the concentration parameter κ, and the phase difference $\Delta\phi$. During runtime, the computational complexity is thus very low and given by a table look-up. In addition, as we solve the integral (32) numerically, we are also flexible with respect to the choice of the phase prior distribution $p_{\Phi_S|\widetilde{\phi_S}}$. In this work, we employ the *von Mises* distribution (14) with concentration parameter κ. The mean-direction of the prior distribution is given by $\widetilde{\phi_S}$, which can be obtained, e.g., using [15]. As in Section III, the concentration parameter κ controls the influence of the prior on the final result. For $\kappa = 0$, the prior has no influence on the result, while for $\kappa \to \infty$ the prior dominates the result.

B. Analysis of the Proposed CUP Estimator

In Fig. 6 we plot the corresponding input-output curves parameterized by the concentration parameter κ of the phase prior as a function of the noisy input y with phase $\phi_Y = 0$ for $\beta = \mu = 1$.

Three cases are interesting to observe, $\kappa = 0$, $\kappa \to \infty$, and $0 < \kappa < \infty$. For $\kappa = 0$ the phase prior distribution is uniform and, as a consequence, the prior phase information $\widetilde{\phi_S}$ is largely uncertain and therefore does not have an influence on the estimation of the clean speech phase and amplitude. Thus, for $\beta = \mu = 1$, when no amplitude compression is employed ($\beta = 1$) and Gaussian speech priors are implied ($\mu = 1$), the behavior of the proposed estimator resembles the Wiener filter: recall that for the Wiener filter (9), the complex clean speech coefficients are estimated as $\widehat{S}_{\mathrm{Wiener}} = Y\sigma_S^2/(\sigma_S^2 + \sigma_{\mathrm{N}}^2)$, i.e., we have a linear relation between the noisy magnitude $r = |y|$ and the estimated amplitude $\widehat{a} = |\widehat{S}_{\mathrm{Wiener}}|$. At the same time the estimated phase is the noisy phase, i.e., $\widehat{\phi_S} = \angle\widehat{S}_{\mathrm{Wiener}} = \phi_Y$. The same behavior is visible for $\kappa = 0$ and $\mu = \beta = 1$ in Fig. 6. In contrast, for $\kappa \to \infty$ the *von Mises* prior distribution (14) approaches a delta function with its peak at the mean-direction $\widetilde{\phi_S}$. As a result, for $\kappa \to \infty$ the amplitude of the estimated complex coefficients is identical to the estimator proposed in [17], while the estimated phase equals the mean-direction of the *von Mises* prior distribution, i.e., $\widehat{\phi_S} = \widetilde{\phi_S}$.

The advantage of the proposed approach is that we can now compromise a deterministic phase prior ($\kappa \to \infty$) and a uniformly distributed phase prior ($\kappa = 0$), by setting $0 < \kappa < \infty$. The proposed estimator thus allows to incorporate prior information about the speech spectral phase denoted by $\widetilde{\phi_S}$, but also allows us to incorporate an uncertainty about the prior phase information which can be controlled via κ. Small values $\kappa \to 0$

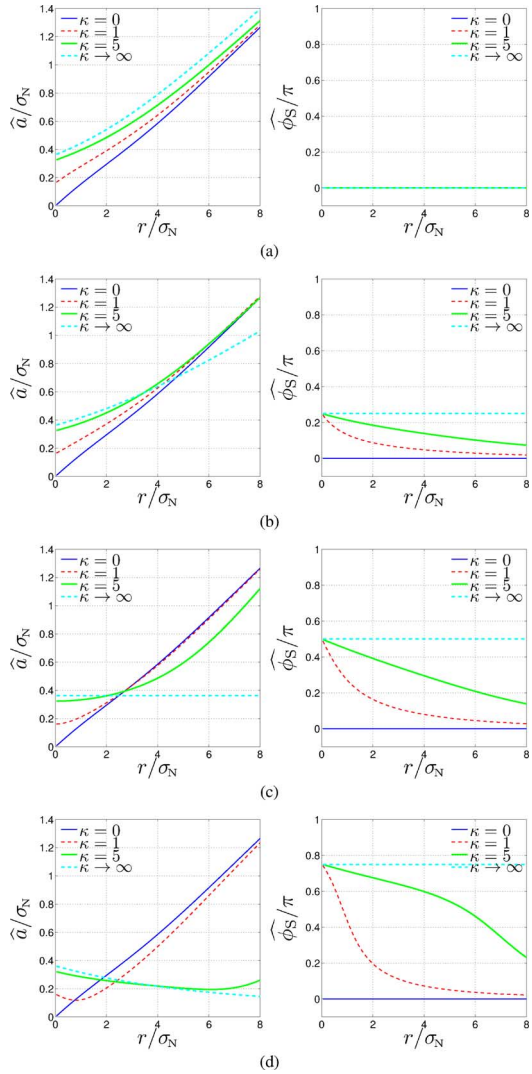


Fig. 6. Amplitude and phase responses of the CUP estimator (32) for $\xi = \sigma_S^2/\sigma_N^2 = 0.2$ and $\mu = \beta = 1$ for different values of the concentration parameter κ in (14). For $\kappa = 0$ the amplitude estimate approaches the behavior of a Wiener filter (left) and the phase estimate results in $\widehat{\phi_S} = \phi_Y$ (right). For $\kappa \to \infty$ the amplitude estimate approaches the result in [17] (left) and the phase estimate results in $\widehat{\phi_S} = \widetilde{\phi_S}$ (right). Amplitude and phase responses for $\phi_Y = 0$ and (a) $\widetilde{\phi_S} = 0$, (b) $\widetilde{\phi_S} = \pi/4$, (c) $\widetilde{\phi_S} = \pi/2$, (d) $\widetilde{\phi_S} = \frac{3}{4}\pi$.

reflect a large uncertainty about the prior phase information $\widetilde{\phi_S}$, while the opposite is true for $\kappa \to \infty$.

V. Experimental Results

In this section, we evaluate the performance of the " proposed CUP estimator (32)" for complex speech coefficients at a sampling rate of $f_s = 16\,\text{kHz}$. For this we process 100 sentences from female speakers and 100 sentences from male speakers taken from dialect region 6 of the TIMIT database [30]. These 200 sentences are corrupted by pink noise, modulated pink noise, nonstationary factory noise, and babble noise at different SNRs. The modulated pink noise is obtained by first creating Gaussian distributed stationary pink noise, and then multiplying it with the amplitude modulation function $(1 + 0.5\sin(2\pi n f_{\text{mod}}/f_s))$. Here, n is the time-domain sample index, and we set the modulation frequency to $f_{\text{mod}} = 0.5$ Hz. The remaining noises are taken from the Noisex-92 database [31]. As Paliwal *et al.* [10] indicated that improving the phase is particularly beneficial if the segment overlap is larger than 50%, for the spectral analysis and synthesis we use square-root Hann windows with 7/8th overlap. For lower frame overlaps, also the performance gain obtained by modifying the spectral phase will be lower. The chosen segment length is 32 ms without zero-padding, resulting in $N = f_s \cdot 32\text{ ms} = 512$ discrete Fourier coefficients, which is sufficiently large to resolve the spectral harmonics of male and female speakers. While we use the increased overlap for phase estimation and signal reconstruction, the noise PSD and speech PSD estimates are only updated every 16 ms to be able to use standard approaches. We estimate the noise PSD σ_N^2 based on a speech presence probability estimate with fixed priors [32], and the *a priori* SNR $\xi = \sigma_S^2/\sigma_N^2$ using the decision-directed approach [8] with the smoothing factor 0.96. While in [8] a factor of 0.98 was proposed, we only use a factor of 0.96 to reduce speech distortions. To limit speech distortions further, all applied gain functions are limited to be larger than $-15\,\text{dB}$. To model the heavy-tailed distribution of speech amplitudes, we set the shape parameter of the amplitude PDF (17) to $\mu = 0.5$. This value was proposed in [24] as it yields a good trade-off between outliers and clarity of speech. Further, to incorporate the compressive character of the auditory system, we set the compression parameter in (32) to $\beta = 0.5$. This value was proposed in [24] as it yields a good trade-off between noise reduction and speech distortions. The settings of $\mu = \beta = 0.5$ are also known as the super-Gaussian amplitude root (SUGAR) estimator [24]. For the estimation of the prior phase information $\widetilde{\phi_S}$ we employ the sinusoidal model based approach [15], where here we only employ the phase reconstruction along frequency. This phase estimator relies on an estimate of the fundamental frequency in voiced speech which we estimate using the PEFAC algorithm [26].

The "proposed CUP estimator (32)" allows us to incorporate the uncertainty about our prior phase information $\widetilde{\phi_S}$ controlled by the concentration parameter κ. Large values for κ reflect a high certainty about the prior phase information, while $\kappa \to 0$ means that we are uncertain about the prior. The certainty of the phase estimate obtained with [15] depends on the certainty about the frequency of each spectral harmonic in voiced speech. These harmonic frequencies are obtained based on multiples of the estimate of the fundamental frequency, meaning that any error in the fundamental frequency estimation is multiplied by the harmonic number. Thus, at high frequencies also the phase estimate is more prone to errors. To reflect this increased uncertainty, we choose κ to be larger for low frequencies than for high frequencies. As the phase estimator [15] does not yield reasonable phase estimates in unvoiced speech, we adapt the value of κ using the probability that a signal frame contains voiced speech $P_{H_V}(\ell)$, which we also obtain using PEFAC [26]. Thus, to account for the increased uncertainty of the prior phase information $\widetilde{\phi_S}$ in high frequencies and unvoiced speech, the uncertainty parameter κ is set to

$$\kappa(k,\ell) = \begin{cases} 4P_{H_V}(\ell), & kf_s/N < 4000\text{ Hz} \\ 2P_{H_V}(\ell), & kf_s/N \geq 4000\text{ Hz} \end{cases}. \qquad (34)$$

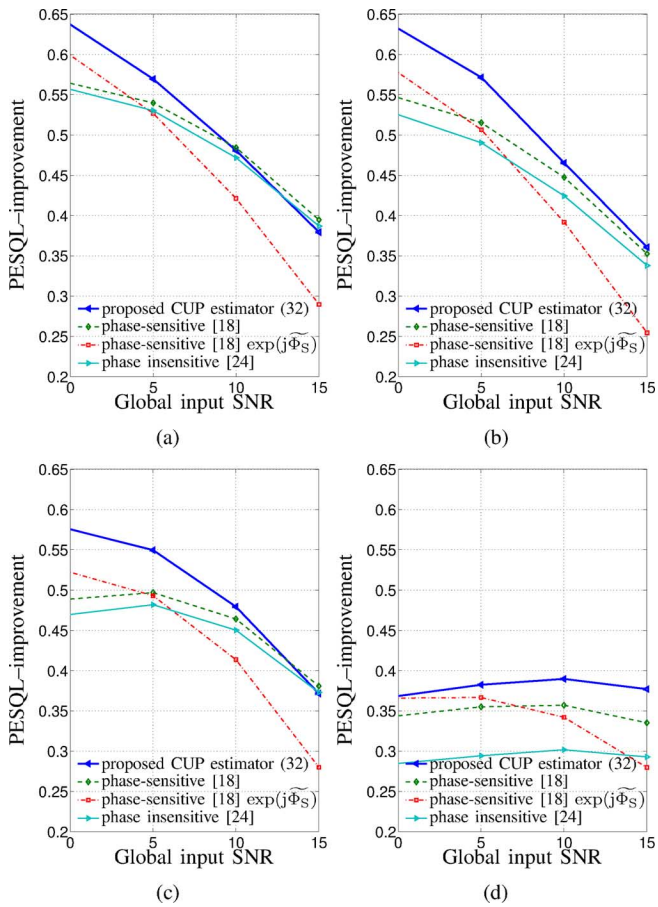


Fig. 7. Instrumental evaluation of the perceptual speech quality for different input SNRs and noise types averaged over 200 sentences (100 spoken by female speakers and 100 spoken by male speakers) from the TIMIT database at $f_s = 16\,\mathrm{kHz}$. We show the improvement in PESQL over the unprocessed noisy speech. (a) Pink noise. (b) Modulated pink Gaussian noise. (c) Nonstationary factory noise. (d) Babble noise.

The values of 4 and 2 are chosen to yield a good trade-off between artifacts and noise reduction. For this, values in the range $\kappa = 0 \ldots 500$ were tested.

We compare the "proposed CUP estimator (32)" to three competing methods. The first competing method "phase insensitive [24]" is a state-of-the-art amplitude enhancement scheme with the same distributional assumptions on the noise coefficients and the speech amplitudes as employed in the "proposed CUP estimator (32)". However, in contrast to the "proposed CUP estimator (32)" the speech phase is considered to be uniformly distributed. As a consequence, the phase of the noisy signal is not employed for amplitude estimation and also not modified in the STFT-domain enhancement.

The second competing algorithm is the phase-sensitive amplitude estimation scheme derived in [17], [18] which we denote as "phase-sensitive [18]". In this method, just as in the "proposed CUP estimator (32)", the phase reconstruction algorithm [15] is employed to obtain a phase estimate. However, in contrast to the "proposed CUP estimator (32)", the prior phase information $\widetilde{\phi_S}$ is treated as being deterministic. While the phase-sensitive gain function is derived in [17], in [18] we combine the phase-sensitive gain function in voiced speech with the phase-insensitive gain function [24] in unvoiced speech based on the probability that a segment is voiced $P_{H_\mathrm{V}}(\ell)$, estimated using PEFAC [26]. Note that in [18] the prior phase information $\widetilde{\phi_S}$ is only employed to improve amplitude estimation, while still the noisy phase is used for signal synthesis, as $\widehat{S} = \widehat{A}\exp(\mathrm{j}\Phi_Y)$.

The third method we use for comparison is also based on the phase-sensitive amplitude estimation scheme [18]. However, while in [18] the prior phase information $\widetilde{\phi_S}$ is only employed to improve amplitude estimation, in the third method, whenever PEFAC signals voiced speech, we use the clean speech phase prior information $\widetilde{\Phi_S}$ also for signal synthesis, as $\widehat{S} = \widehat{A}\exp(\mathrm{j}\widetilde{\Phi_S})$. This method is denoted as "phase-sensitive [18] $\exp(\mathrm{j}\widetilde{\Phi_S})$" and is equivalent to the CUP estimator when setting $\kappa \to \infty$.

The performance is evaluated using the perceptual evaluation of speech quality as provided in Loizou's book [33], which we denote by PESQL, and the short-time objective intelligibility (STOI) measure [34]. We show the improvements in PESQL (Fig. 7) and STOI (Fig. 8) relative to the unprocessed noisy speech. It can be seen that incorporating phase information for an improved amplitude estimation using the "phase-sensitive [18]" approach outperforms the "phase insensitive [24]" approach both in terms of predicted quality and intelligibility. The PESQL benefit of the "phase-sensitive [18]" estimator is largest for babble noise. At this point, it is interesting to note that the phase-sensitive amplitude estimator can only yield a benefit to the phase-insensitive counterpart in voiced speech. Hence, as voiced speech has most of its energy in low frequencies, at a sampling rate of $f_s = 8\,\mathrm{kHz}$ and in voiced speech the benefit of the phase-sensitive approaches is even more pronounced (see e.g., [17]) as compared to the results shown here. If the clean speech phase prior information $\widetilde{\phi_S}$ is treated as being deterministic and is also used for signal synthesis using the "phase-sensitive [18] $\exp(\mathrm{j}\widetilde{\Phi_S})$" approach, in low SNRs the PESQL improvement of the phase-sensitive approaches can be improved even further. However, in high SNRs, employing the deterministically treated clean speech phase estimate for signal synthesis degrades speech quality. As reported in [15], [17], informal listening confirms that artifacts may occur when the estimated phase is directly employed for signal synthesis instead of the phase of the noisy observation. Furthermore, STOI always predicts a decreased intelligibility for the "phase-sensitive [18] $\exp(\mathrm{j}\widetilde{\Phi_S})$" approach.

However, with the "proposed CUP estimator (32)", where we also consider the uncertainty about the prior phase information $\widetilde{\phi_S}$, even larger quality improvements are predicted in low SNRs and the quality degradation in high SNRs is avoided. Furthermore, STOI also predicts the largest improvements in intelligibility for the CUP estimator. Informal listening confirms that artifacts are reduced by taking the uncertainty of the phase estimate into account. Audio examples are provided at http://speech.uniol.de/cupestimator.html. Finally, it is interesting to note that the negative STOI scores in babble noise can be avoided by estimating the speech PSD using temporal

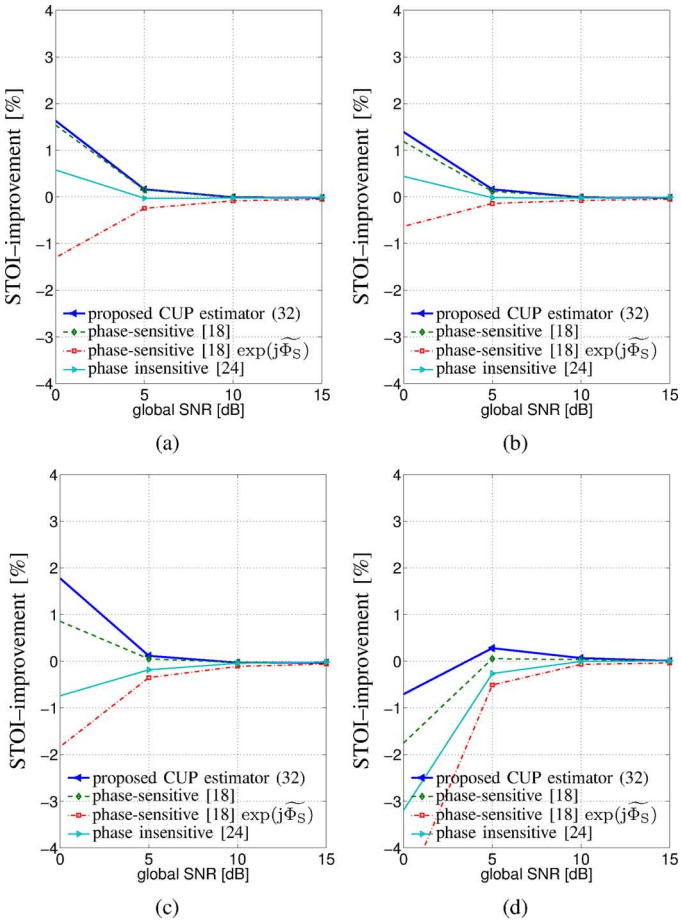


Fig. 8. Evaluation using the short-time objective intelligibility (STOI) measure [34] for the same setup as in Fig. 7 (a) Pink noise. (b) Modulated pink Gaussian noise. (c) Nonstationary factory noise. (d) Babble noise.

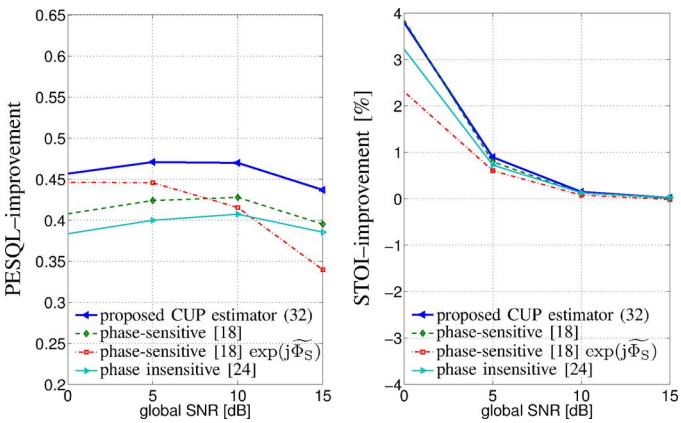


Fig. 9. PESQL and STOI for babble noise when temporal cepstrum smoothing [35], [36] is employed for speech PSD estimation.

cepstrum smoothing [35], [36], instead of the decision directed approach [8]. These results are given in Fig. 9.

VI. Summary and Conclusions

While most algorithms for STFT-domain speech enhancement modify only the spectral amplitude, more attention is drawn recently towards phase processing. In this paper we addressed the problem of phase estimation and phase-sensitive speech enhancement from a Bayesian perspective. We analyzed the phase estimation problem for different degrees of prior knowledge about the speech and noise coefficients. First, we considered the scenario that besides the noisy observation both the speech and noise spectral amplitudes are perfectly known. It is interesting to note that even in this idealistic case, the phase estimation problem is still ambiguous. This ambiguity can be mended, for instance by incorporating the group delay [19]. Next, we considered the case where only perfect knowledge of the clean speech amplitudes is given while the noisy coefficients follow a complex Gaussian distribution. We show that the resulting posterior density of the clean speech phase is given by a *von Mises* distribution with its mean-direction given by the phase of the noisy signal. Subsequently, we analyzed the most relevant scenario for speech enhancement, where both the speech and noise amplitudes are unknown random variables. For this scenario we derived the ML and MAP estimators for the clean speech phase when the complex noise coefficients are Gaussian distributed and the speech amplitudes are χ-distributed. While the ML-optimal estimate is shown to be the phase of the noisy observation, the MAP estimator allows for incorporating prior knowledge of the phase. This prior clean speech phase information can be obtained for instance using the phase reconstruction algorithm [15], [16]. The MAP-optimal phase estimate then results in a trade-off between the phase of the noisy observation and the prior information on the phase, controlled by the uncertainty of this prior information. Finally, we derived a joint MMSE-optimal estimator of the clean speech amplitude and phase, when uncertain prior knowledge of the clean speech phase is given. While combining a deterministic clean speech phase estimate with enhanced amplitudes may yield artifacts in the reconstructed speech, incorporating the uncertainty of the prior phase estimate with the proposed Bayesian estimators reduces these artifacts. Furthermore, for a large range of SNRs and noise types, we showed that this "proposed CUP estimator (32)" improves the instrumentally predicted speech quality and speech intelligibility further.

It is interesting to note that the proposed estimators can also be employed if multiple microphones are present. This is because under a Gaussian noise model, the output of an minimum variance distortionless response (MVDR) beamformer provides *sufficient statistics* for functions of the clean speech spectral coefficients [37]. Thus, while the estimators in this paper are derived for a single channel observation, if multiple microphones are available it is also statistically optimal to apply the derived single channel estimators at the output of an MVDR beamformer. Another interesting extension of the proposed estimators is a scenario where multiple speakers are present. In this scenario, perspectively, the proposed estimators can still be used when a multi-pitch tracker (e.g., [38]) is employed to estimate the prior phase information using [15].

Acknowledgment

The author would like to thank M. Krawczyk for valuable comments and discussions.

REFERENCES

[1] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding And Error Concealment*. Chichester, West Sussex, U.K.: Wiley, 2006.

[2] L. Deng, J. Droppo, and A. Acero, "Enhancement of log mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 2, pp. 133–143, Mar. 2004.

[3] D. L. Wang and J. S. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoust., Speech, Signal Process.*, no. 4, pp. 679–681, Aug. 1982.

[4] P. Vary, "Noise suppression by spectral magnitude estimation – mechanism and theoretical limits," *ELSEVIER Signal Process.*, vol. 8, pp. 387–400, May 1985.

[5] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP J. Appl. Signal Process.*, vol. 2005, no. 7, pp. 1110–1126, Jan. 2005.

[6] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 6, pp. 1741–1752, Aug. 2007.

[7] J. E. Porter and S. F. Boll, "Optimal estimators for spectral restoration of noisy speech," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, San Diego, CA, USA, Mar. 1984, pp. 18A.2.1–18A.2.4.

[8] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[9] J. S. Erkelens, R. C. Hendriks, and R. Heusdens, "On the estimation of complex speech DFT coefficients without assuming independent real and imaginary parts," *IEEE Signal Process. Lett.*, vol. 15, pp. 213–216, 2008.

[10] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *ELSEVIER Speech Commun.*, vol. 53, no. 4, pp. 465–494, Apr. 2011.

[11] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 2, pp. 236–243, Apr. 1984.

[12] N. Sturmel and L. Daudet, "Signal reconstruction from STFT magnitude: A state of the art," in *Proc. Int. Conf. Digit. Audio Effects (DAFx)*, Paris, France, Sep. 2011, pp. 375–386.

[13] N. Sturmel and L. Daudet, "Iterative phase reconstruction of Wiener filtered signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Kyoto, Japan, Mar. 2012, pp. 101–104.

[14] J. Le Roux and E. Vincent, "Consistent Wiener filtering for audio source separation," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 217–220, Mar. 2013.

[15] M. Krawczyk and T. Gerkmann, "STFT phase improvement for single channel speech enhancement," presented at the Int. Workshop Acoust. Echo, Noise Control (IWAENC), Aachen, Germany, Sep. 2012.

[16] M. Krawczyk and T. Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, 2014, to be published.

[17] T. Gerkmann and M. Krawczyk, "MMSE-optimal spectral amplitude estimation given the STFT-phase," *IEEE Signal Process. Lett.*, vol. 20, no. 2, pp. 129–132, Feb. 2013.

[18] M. Krawczyk, R. Rehr, and T. Gerkmann, "Phase-sensitive real-time capable speech enhancement under voiced-unvoiced uncertainty," presented at the EURASIP Eur. Signal Process. Conf. (EUSIPCO), Marrakech, Morocco, Sep. 2013.

[19] P. Mowlaee, R. Saeidi, and R. Martin, "Phase estimation for signal reconstruction in single-channel speech separation," presented at the ISCA Interspeech, Portland, OR, USA, Sep. 2012.

[20] R. Astudillo, "Integration of short-time Fourier domain speech enhancement and observation uncertainty techniques for robust automatic speech recognition," Ph.D. dissertation, TU Berlin, Berlin, Germany, 2010.

[21] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals Series and Products*, 7th ed. San Diego, CA, USA: Academic, 2007.

[22] K. V. Mardia and P. E. Jupp, *Directional Statistics*. Chichester, U.K.: Wiley, 2000.

[23] I. Andrianakis and P. R. White, "MMSE speech spectral amplitude estimators with Chi and Gamma speech priors," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Toulouse, France, May 2006, pp. 1068–1071.

[24] C. Breithaupt, M. Krawczyk, and R. Martin, "Parameterized MMSE spectral magnitude estimation for the enhancement of noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Las Vegas, NV, USA, Apr. 2008, pp. 4037–4040.

[25] K. Oldham, J. Myland, and J. Spanier, *An Atlas of Function*, 2nd ed. New York, NY, USA: Springer, 2009.

[26] S. Gonzalez and M. Brookes, "PEFAC – A pitch estimation algorithm robust to high levels of noise," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 2, pp. 518–530, Feb. 2014.

[27] T. Gerkmann, "MMSE-optimal enhancement of complex speech coefficients with uncertain prior knowledge of the clean speech phase," presented at the IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Florence, Italy, May 2014.

[28] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[29] C. H. You, S. N. Koh, and S. Rahardja, "β-order MMSE spectral amplitude estimation for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 4, pp. 475–486, Jul. 2005.

[30] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," Nat. Inst. of Standards and Technol. (NIST), Gaithersburg, MD, USA, 1988.

[31] A. Varga and H. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *ELSEVIER Speech Commun.*, vol. 12, pp. 247–251, Jul. 1993.

[32] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[33] P. C. Loizou, *Speech Enhancement – Theory and Practice*. Boca Raton, FL, USA: CRC Press/Taylor & Francis Group, 2007.

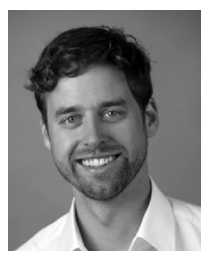
[34] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[35] C. Breithaupt, T. Gerkmann, and R. Martin, "A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Las Vegas, NV, USA, Apr. 2008, pp. 4897–4900.

[36] T. Gerkmann and R. Martin, "On the statistics of spectral amplitudes after variance reduction by temporal cepstrum smoothing and cepstral nulling," *IEEE Trans. Signal Process.*, vol. 57, no. 11, pp. 4165–4174, Nov. 2009.

[37] R. Balan and J. Rosca, "Microphone array speech enhancement by Bayesian estimation of spectral amplitude and phase," in *Proc. IEEE Sensor Array Multichannel Signal Process. Workshop (SAM)*, Rosslyn, VA, USA, Aug. 2002, pp. 209–213.

[38] M. G. Christensen and A. Jakobsson, *Multi-Pitch Estimation*. San Rafael, CA, USA: Morgan & Claypool, 2009.



Timo Gerkmann (S'08–M'10) studied electrical engineering at the University of Bremen and the University of Bochum, Germany. He received the Dipl.-Ing. degree in 2004 and the Dr.-Ing. degree in 2010, both at the Institute of Communication Acoustics (IKA) at the Ruhr-Universität Bochum, Bochum, Germany. In 2005, he spent six months with Siemens Corporate Research in Princeton, NJ, USA. During 2010 to 2011, he was a postdoctoral researcher at the Sound and Image Processing Lab at the Royal Institute of Technology (KTH), Stockholm, Sweden. Since 2011, he has been a professor for Speech Signal Processing at the Universität Oldenburg, Oldenburg, Germany. His main research interests are digital signal processing for speech and audio, including speech enhancement, modeling of speech signals, and hearing devices.