On MMSE-Based Estimation of Amplitude and Complex Speech Spectral Coefficients Under Phase-Uncertainty

Martin Krawczyk-Becker, *Student Member, IEEE*, and Timo Gerkmann, *Senior Member, IEEE*

***Abstract*—Among the most commonly used single-channel approaches for the enhancement of noise corrupted speech are Bayesian estimators of clean speech coefficients in the short-time Fourier transform domain. However, the vast majority of these approaches effectively only modifies the spectral amplitude and does not consider any information about the clean speech spectral phase. More recently, clean speech estimators that can utilize prior phase information have been proposed and shown to lead to improvements over the traditional, phase-blind approaches. In this work, we revisit phase-aware estimators of clean speech amplitudes and complex coefficients. To complete the existing set of estimators, we first derive a novel amplitude estimator given uncertain prior phase information. Second, we derive a closed-form solution for complex coefficients when the prior phase information is completely uncertain or not available. We put the novel estimators into the context of existing estimators and discuss their advantages and disadvantages.**

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

I. Introduction

In many everyday situations, we are confronted with acoustic noise. Severe acoustic noise not only complicates human-to-human communication, but also poses a problem to many technical devices, such as mobile phones or hearing aids. For such devices to enable successful communications even in challenging acoustic scenarios, algorithms for the reduction of acoustic noise are a key component. Here we consider single-channel speech enhancement approaches, which can either be applied directly to a noisy microphone signal or to the output of a spatial multi-microphone pre-processing stage. We further concentrate on Bayesian estimators of the clean speech, which estimate the clean speech based on statistical assumptions about the speech and the noise components. The majority of these algorithms is formulated in the short time discrete Fourier transform (STFT) domain due to its low computational complexity and intuitive interpretation. In this work, we differentiate between two classes of estimators: estimators of the complex-valued clean speech spectral coefficients S and estimators of the real-valued clean speech spectral amplitude $A = |S|$. For example, if the speech and the noise are independently circular-complex Gaussian distributed, the Wiener filter is the optimal estimator of S in the minimum mean squared error (MMSE) sense, while the short-time spectral amplitude estimator (STSA) [1] is the MMSE optimal estimator of A. Under the Gaussian assumption, it has further been shown that the clean speech spectral phase is uniformly distributed and that the noisy phase is the optimal Bayesian estimator [1]. Consequently, both approaches only modify the spectral amplitude, while the noisy phase is left unchanged. Over time, several more advanced estimators have been derived, which optimize for compressed amplitudes, e.g., [2], [3], incorporate heavy-tailed speech priors, e.g., [4]–[6], or both, [7], [8]. Optimizing for compressed amplitudes has been reported to be perceptually beneficial, e.g., [2], and can be considered as a simple model of the compressive behavior of the human auditory system. While in [2] and [8] the logarithm is used as the compressive function, in [3] and [7] a more general β-order compression has been proposed. Heavy-tailed, i.e., super-Gaussian, speech priors have been proposed, e.g., in [4]–[8], as they are reported to fit the histogram of clean speech better than a Gaussian prior [4], [5].

Recent years have seen a rising interest in the role of the spectral phase for speech enhancement. For instance, in [9], the general importance of the spectral phase for speech enhancement is highlighted by means of numerous instrumental and subjective experiments. It has furthermore been shown that considering the spectral phase in spectral subtraction can substantially reduce musical noise [10] and has the potential to improve automatic speech recognition performance [11] compared to conventional spectral subtraction. Also in the modulation frequency domain, separately processing the real and imaginary parts of the spectral coefficients instead of only their amplitudes, which effectively also modifies the spectral phase, leads to improvements in instrumental measures as well as subjective quality over magnitude only enhancement [12]. Additionally, different approaches for the estimation of the clean speech spectral phase have been proposed, e.g., [13]–[15]. While in [13] the iterative estimation of the spectral phase from the clean speech spectral magnitude is investigated, in [14], [15] methods that estimate the clean spectral phase from the noisy observation based on a harmonic signal model are proposed. Once an estimate of the clean speech spectral phase is available, there are different ways to utilize the additional information for an improved speech enhancement. A straight forward way is to simply exchange the noisy phase

Manuscript received December 15, 2015; revised June 20, 2016; accepted August 9, 2016. Date of publication August 24, 2016; date of current version September 19, 2016. This work was supported by the DFG Project GE2538/2-1. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yunxin Zhao.

The authors are with the Signal Processing Group Department of Informatics University of Hamburg, Hamburg 22527, Germany (e-mail: martin.krawczyk-becker@uni-hamburg.de; timo.gerkmann@uni-hamburg.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/TASLP.2016.2602549

TABLE I
CUP AND AUP TOGETHER WITH THEIR SPECIAL CASES, I.E., NEGLECTING THE PRIOR PHASE INFORMATION ($\kappa = 0$) AND ASSUMING THAT THE INITIAL PHASE ESTIMATE YIELDS EXACTLY THE TRUE CLEAN PHASE ($\kappa \rightarrow \infty$)

	estimator of complex coefficients	*estimator of amplitudes*
phase-blind $\kappa = 0$	**BECOCO (21)**	MOSIE (13)
uncertain phase $0 < \kappa < \infty$	CUP (15)	**AUP (10)**
certain phase $\kappa = \infty$	CDP (16)	ADP (12)

The estimators that are derived in this paper are highlighted in bold print.

with the estimated clean speech phase and reconstruct the time domain signal, e.g., [14], [15]. In a natural next step, we can combine the phase estimate with a spectral amplitude that we estimate with one of the approaches mentioned above. However, if an estimate of the clean speech spectral phase is available, the traditional phase-blind approaches, like the STSA [1], are not MMSE optimal anymore [16]. In [16], a phase-aware estimator of the clean speech spectral amplitude has been derived that is optimal in the MMSE sense if the true clean speech phase is given. In practice, however, typically only an estimate of the clean speech phase is available, e.g., obtained via the model-based approaches in [14], [15] or iteratively as proposed in [17]. In [18], the uncertainty in such a phase estimate is incorporated into an estimator of the (C)omplex spectral speech coefficients given (U)ncertain (P)hase information (CUP) by means of a prior distribution for the true clean speech phase. This estimator has been shown to improve the speech quality as well as the speech intelligibility as predicted by instrumental measures with respect to traditional phase-blind approaches. In [19], CUP has further been extended by using different, non-Gaussian distributions for the noise. For a more extensive overview of the history and recent advances in phase-aware speech processing, the interested reader is referred to [20] and [21].

In this paper, we revisit phase-aware estimators of clean speech amplitudes and complex coefficients. To complete the existing set of estimators, we first derive the novel estimator of the speech (A)mplitudes given (U)ncertain (P)hase information (AUP). Secondly, we derive a closed-form solution for complex coefficients when the initial phase is completely uncertain or not available, resulting in the novel phase-(B)lind (E)stimator of (CO)mplex (CO)efficients (BECOCO). We then put the novel estimators into the context of existing approaches, summarized in Table I, where we highlight the entries that have been blank before and have been filled as a contribution of this paper, i.e., AUP and BECOCO. We discuss their advantages and disadvantages based on a theoretical analysis and investigate how the quality of the initial phase information affects the final enhancement results. The presented analysis allows for a detailed assessment and comparison of the different phase-aware estimators and their sensitivity to errors in the initial phase estimate. Finally, the estimators are evaluated on noise corrupted speech.

In Section II, we introduce the basic concept of phase-aware clean speech estimation that is common to all estimators considered in this work. While in Section III we derive the novel phase-aware amplitude estimator AUP and discuss its special cases, in Section IV we revisit the complex estimator CUP, leading to the derivation of BECOCO. The estimators are then analyzed and compared in Section V. An instrumental evaluation on noise corrupted speech with respect to speech quality and intelligibility is presented and discussed in Section VI, before concluding the paper in Section VII.

II. Principles of Phase-Aware Clean Speech Estimation

In the STFT domain we denote the noise corrupted observation in each time-frequency point (ℓ, k) as

$$Y_{k,\ell} = S_{k,\ell} + V_{k,\ell}, \tag{1}$$

with mutually independent clean speech $S_{k,\ell}$ and additive noise $V_{k,\ell}$. In the remainder of this paper we neglect the segment index ℓ and frequency index k for notational convenience. We can express the complex-valued coefficients in terms of their amplitudes and phases, i.e., $Y = \mathrm{R}e^{j\Phi^Y}$, $S = \mathrm{A}e^{j\Phi^S}$, and $V = \mathrm{D}e^{j\Phi^V}$.

We further assume that some initial estimate $\widetilde{\Phi^S}$ of the clean speech phase Φ^S is available, which could for example be obtained with the phase reconstruction approach proposed in [14]. To incorporate this prior information into an estimator of the clean speech coefficients S – or functions $f(S)$ thereof – we search for the expected value $\mathrm{E}(\cdot)$ given the noisy observation and the initial phase estimate $\widetilde{\phi^S}$:

$$\begin{aligned}\widehat{f(S)} = \mathrm{E}\Big(f(S) \mid y, \widetilde{\phi^S}\Big) &= \int_0^\infty \int_0^{2\pi} f(S) \\ &\times p_{A,\Phi^S|r,\phi^Y,\widetilde{\phi^S}}\Big(a, \phi^S \mid r, \phi^Y, \widetilde{\phi^S}\Big)\, d\phi^S\, da,\end{aligned} \tag{2}$$

where we use the hat-symbol to distinguish estimated quantities from their true counterparts, e.g., $\widehat{X}$ is an estimate of X. Furthermore, lower-case letters denote realizations of the random variables in capital letters, e.g., a is a realization of A. This style of notation is used throughout this paper, but the subscripts of the probability density functions (PDFs) will be dropped for brevity, e.g., $p_A(a) = p(a)$. Note that the posterior $p(a, \phi^S \mid r, \phi^Y, \widetilde{\phi^S})$ is implicitly also conditioned on σ_S^2 and σ_V^2, which we do not state explicitly to achieve a compact notation. The resulting estimator is optimal in the sense that it minimizes the mean squared error (MSE) [22]

$$\mathrm{E}\Big(|\widehat{f(S)} - f(S)|^2 \mid y, \widetilde{\phi^S}\Big). \tag{3}$$

With Bayes' rule and assuming the speech prior $p(S)$ to be circular-symmetric in the complex plane, we can reformulate the posterior and the estimator (2) becomes (see [18] for details):

$$\boxed{\begin{aligned}\widehat{f(S)} &= \mathrm{E}\Big(f(S) \mid y, \widetilde{\phi^S}\Big) \\ &= \frac{\int_0^\infty \int_0^{2\pi} f(S) p\left(y|a, \phi^S\right) p(a) p\left(\phi^S|\widetilde{\phi^S}\right) d\phi^S da}{\int_0^\infty \int_0^{2\pi} p\left(y|a, \phi^S\right) p(a) p\left(\phi^S|\widetilde{\phi^S}\right) d\phi^S da}.\end{aligned}} \tag{4}$$

This formula yields the basis for all phase-aware estimators that we consider in this work.

To derive a specific clean speech estimator, we have to solve (4). For this, we first have to make some assumptions about the distributions of the speech and the noise. Note that here we use the same signal models as in [18]: the speech spectral coefficients S are assumed to follow a circular-symmetric heavy-tailed super-Gaussian distribution with variance σ_S^2. We therefore model the prior of the corresponding speech amplitudes A with a χ-distribution, i.e.,

$$p_A(a) = \frac{2}{\Gamma(\mu)} \left(\frac{\mu}{\sigma_S^2}\right)^{\mu} a^{2\mu-1} \exp\left(-\frac{\mu}{\sigma_S^2} a^2\right). \tag{5}$$

While $\mu = 1$ corresponds to a Gaussian distribution of the complex speech coefficients, to model more heavy-tailed speech priors we set $0 < \mu < 1$. Such heavy-tailed distributions have been shown to better fit the histograms of clean speech [4] and also to lead to better results in phase-blind clean speech estimators, e.g., [5]. Note that (5) corresponds to the generalized gamma distribution as used e.g., in [5] with $\gamma^{[5]} = 2$ and $\nu^{[5]} = \mu$. We further assume that the noise V is zero-mean circular symmetric complex Gaussian distributed with variance σ_V^2, which, in polar coordinates, results in the likelihood

$$p\left(r, \phi^Y \mid a, \phi^S\right) = \frac{r}{\pi\sigma_V^2} \exp\left(-\frac{|\mathrm{r}\,\mathrm{e}^{j\phi^Y} - \mathrm{a}\,\mathrm{e}^{j\phi^S}|^2}{\sigma_V^2}\right) \tag{6}$$

$$= \frac{r}{\pi\sigma_V^2} \exp\left(\frac{2ra\cos\left(\phi^S - \phi^Y\right) - r^2 - a^2}{\sigma_V^2}\right). \tag{7}$$

The only part of (4) that is still missing is $p(\phi^S|\widetilde{\phi^S})$, which is the PDF of the true clean speech phase Φ^S given the initial phase estimate $\widetilde{\phi^S}$. As proposed in [18], we model $p(\phi^S|\widetilde{\phi^S})$ using a *von Mises* distribution with mean direction $\widetilde{\phi^S}$,

$$p\left(\phi^S|\widetilde{\phi^S}\right) = \exp\left(\kappa\cos\left(\phi^S - \widetilde{\phi^S}\right)\right) / \left(2\pi I_0\left(\kappa\right)\right), \tag{8}$$

where κ is the concentration parameter and $I_n(\cdot)$ is the modified Bessel function of the first kind and n-th order. For an increasing concentration parameter κ, the circular variance of (8) decreases, while the mean direction $\widetilde{\phi^S}$ corresponds to the mode of the circularly symmetric von Mises distribution (8). The von Mises distribution hence allows us to effectively model the certainty of the available initial phase estimate $\widetilde{\phi^S}$ by adequately choosing κ. Illustrative examples of $p(\phi^S|\widetilde{\phi^S})$ for $\widetilde{\phi^S} = 0$ and three different values for the concentration parameter κ are presented in Fig. 1. For large values of κ, $p(\phi^S|\widetilde{\phi^S})$ is strongly concentrated around $\widetilde{\phi^S}$. Accordingly, the true clean speech phase ϕ^S is likely to be reasonably close to the initial phase estimate $\widetilde{\phi^S}$. In other words, $\widetilde{\phi^S}$ represents a reliable initial estimate of the true clean speech phase ϕ^S. For small values of κ on the other hand, $p(\phi^S|\widetilde{\phi^S})$ approaches a uniform distribution, i.e., the initial phase estimate $\widetilde{\phi^S}$ yields only little information about the true clean speech phase.

Now that we have models for all distributions in (4), we can derive estimators of the clean speech complex coefficients and of the clean speech amplitudes by choosing $f(S)$ accordingly. In the following sections, we derive novel estimators, but also revisit existing estimators to allow for a comprehensive comparison and discussion in a wider context. To highlight the estimators contributed in this paper, we put the resulting novel estimators into a box.

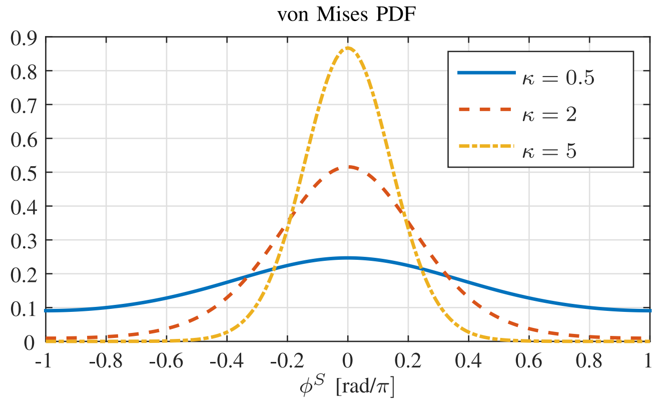

Fig. 1. Von Mises distribution for a mean direction of $\widetilde{\phi^S} = 0$ and three different values for the concentration parameter κ.

III. Phase-Aware Amplitude Estimation

The phase-aware estimator CUP, proposed in [18] and revisited in Section IV, estimates the compressed clean speech *complex coefficients*. However, for phase-blind approaches, estimators of the clean speech *amplitude* have been reported to yield less speech distortions than estimators of the complex coefficients, e.g., [1]. To investigate if this is also the case for phase-aware estimators, we now derive the novel estimator of the speech amplitude given uncertain phase information AUP.

A. Amplitude Estimation Given Uncertain Phase Information (AUP)

For the derivation of the novel estimator AUP, we define

$$f(S) = |S|^\beta = A^\beta \tag{9}$$

in (4), i.e., AUP estimates the compressed speech amplitudes. The parameter β introduces some flexibility with respect to the cost function (3). For example, setting $0 < \beta < 1$ results in a compression of spectral amplitudes, which has been reported to yield perceptually beneficial results in phase-blind amplitude estimation [3], [7].

To find AUP, we insert (5), (7), and (9) into (4) and then solve the integral over the speech amplitude using [23, (3.462.1)], leading to

$$\boxed{\widehat{A}^\beta = \left(\sqrt{\frac{1}{2}\frac{\xi}{\mu+\xi}\sigma_V^2}\right)^\beta \frac{\Gamma(2\mu+\beta)}{\Gamma(2\mu)} \times \frac{\int_0^{2\pi} e^{\nu^2/4}\mathrm{D}_{(-2\mu-\beta)}(\nu)\, p\left(\phi^S|\widetilde{\phi^S}\right) d\phi^S}{\int_0^{2\pi} e^{\nu^2/4}\mathrm{D}_{(-2\mu)}(\nu)\, p\left(\phi^S|\widetilde{\phi^S}\right) d\phi^S},} \tag{10}$$

with the a priori signal-to-noise ratio (SNR) $\xi = \frac{\sigma_S^2}{\sigma_V^2}$, the parabolic cylinder function $\mathrm{D}_\cdot(\cdot)$ [23, (9.241.2)], and the argument

$$\nu = -\frac{r}{\sigma_V}\sqrt{2\frac{\xi}{\mu+\xi}}\cos(\underbrace{\widetilde{\phi^S} - \phi^Y}_{\Delta\phi}). \qquad (11)$$

Here, $\Delta\phi$ denotes the difference between the observed phase ϕ^Y and the true clean speech phase ϕ^S. Finally, we can plug the von Mises phase prior (8) into (10) to obtain AUP. Unfortunately, there is no known closed-form solution to the phase integral for a von Mises phase prior $p(\phi^S|\widetilde{\phi^S})$. However, since the integral over the phase is limited to $-\pi$ and π, it can be solved numerically with high precision [18]. A look-up table can be computed off-line, reducing the computational complexity during runtime to a simple table look up. For a specific value of κ, here we use a three-dimensional table with a resolution of 1 dB for ξ and γ, and $\pi/100$ for $(\widetilde{\phi^S} - \phi^Y)$. For the synthesis of the enhanced time domain signal, the estimated compressed amplitude (10) is first expanded via $(\widehat{A}^\beta)^{1/\beta}$. The amplitude is then combined with the noisy phase, giving the final clean speech estimate $\widehat{S}_{\mathrm{AUP}} = \widehat{A}\exp(j\Phi^Y)$. Note that the amplitude estimator AUP thus only enhances the spectral amplitude and does not modify the noisy phase. One motivation for proposing AUP, i.e., for using the spectral phase only for amplitude estimation but not for phase improvement, is that artifacts known from phase modifications, see e.g., [13], [14], are impossible.

As will be presented in Section IV-A, the difference between the novel AUP and the complex estimator CUP [18], lies in the definition of $f(S)$ in (4). While for AUP only amplitudes are optimized for (9), in CUP both amplitudes and phases are included in the estimation as we will see in (14). Besides this difference, all statistical assumptions about the distributions in (4) are the same. In this sense, AUP represents the amplitude counterpart to CUP.

We now have a closer look at two special cases of AUP, namely $\kappa \to \infty$ and $\kappa = 0$ in (8). While for $\kappa \to \infty$ the uncertainty in the initial phase $\widetilde{\phi^S}$ is neglected, setting $\kappa = 0$ effectively leads to a phase-blind estimator. We show that both cases resemble known estimators of the clean speech amplitude, for which closed-form solutions exist.

B. Perfectly Known Speech Phase ($\kappa \to \infty$)

For a concentration parameter $\kappa \to \infty$, the von Mises distribution (8) approaches a delta function, $p(\phi^S|\widetilde{\phi^S}) \to \delta(\phi^S - \widetilde{\phi^S})$, which is only non-zero for $\phi^S = \widetilde{\phi^S}$. In this case, the initial phase estimate $\widetilde{\phi^S}$ is implicitly assumed to be deterministic and identical to the true clean speech phase. Inserting (8) into (10) for $\kappa \to \infty$, we can utilize the sifting property of the delta function to solve the integral over ϕ^S, yielding the speech (A)mplitude estimator given (D)eterministic (P)hase information (ADP) proposed in [16]

$$\widehat{A}_D^\beta = \mathrm{E}\big(A^\beta \mid y, \phi^S\big) = \left(\sqrt{\frac{1}{2}\frac{\xi}{\mu+\xi}\sigma_V^2}\right)^\beta \times \frac{\Gamma(2\mu+\beta)}{\Gamma(2\mu)}\frac{\mathrm{D}_{(-2\mu-\beta)}(\nu)}{\mathrm{D}_{(-2\mu)}(\nu)}, \qquad (12)$$

which does not incorporate any uncertainty in the prior phase information. In practice, however, the initial phase $\widetilde{\phi^S}$ yields only an estimate of the clean speech phase. By choosing $\kappa \to \infty$, the uncertainty of this estimate is neglected, potentially leading to suboptimal enhancement results for an unreliable initial phase $\widetilde{\phi^S}$. As AUP is defined as an estimator of spectral amplitudes only, for signal reconstruction we again use the noisy phase. In (12), we introduce the index d to denote estimators that assume that the clean speech phase is perfectly known and deterministic, i.e., $\kappa \to \infty$.

C. Phase-Blind ($\kappa = 0$)

For $\kappa = 0$, the von Mises distribution (8) reduces to a uniform distribution, which is $p(\phi^S|\widetilde{\phi^S}) = \frac{1}{2\pi}$ between $-\pi$ and π and zero elsewhere. Accordingly, the initial phase $\widetilde{\Phi^S}$ does not provide any useful information and the estimator (4) becomes phase-blind. Solving (4) for $f(S) = |S|^\beta = A^\beta$, we obtain the parametric amplitude estimator in [7]

$$\widehat{A}_B^\beta = \mathrm{E}\big(A^\beta \mid y\big) \qquad (13)$$

as a special case of AUP for total uncertainty in the a priori phase estimate. Here, the index B is used to denote phase-blind estimators. In accordance to [24], we denote the phase-blind amplitude estimator (13) as MOSIE, i.e., (M)MSE estimation with (O)ptimizable (S)peech model and (I)nhomogeneous (E)rror criterion.

IV. Phase-Aware Estimation of Complex Coefficients and Relations to Phase-Aware Amplitude Estimation

In this section, we revisit the phase-aware estimator of complex speech coefficients CUP [18], highlighting differences and similarities to the novel estimator AUP. After introducing the general formulation, similar to AUP, the special cases of $\kappa \to \infty$ and $\kappa = 0$ are presented and discussed. For the latter, we derive a novel phase-blind estimator of the compressed speech coefficients, which may be considered the complex counterpart to the phase-blind amplitude estimator (13).

A. Complex Estimation Given Uncertain Phase Information (CUP)

In [18] the phase-aware estimator CUP is derived by solving (4) for

$$f(S) = S^{(\beta)} = A^\beta e^{j\Phi^S}, \qquad (14)$$

i.e., CUP estimates the compressed complex-valued speech coefficients, rather than the compressed speech amplitudes as done for AUP (see (9)).

Again, equations (5), (7), and (14) are inserted into (4) and the integral over the speech amplitude is solved using [23, (3.462.1)], giving [18]

$$\widehat{S^{(\beta)}} = \left(\sqrt{\frac{1}{2}\frac{\xi}{\mu+\xi}\sigma_V^2}\right)^{\beta} \frac{\Gamma\left(2\mu+\beta\right)}{\Gamma\left(2\mu\right)} \times \frac{\int_0^{2\pi} e^{j\phi^S} e^{\nu^2/4}\mathrm{D}_{(-2\mu-\beta)}(\nu)p\left(\phi^S|\widetilde{\phi^S}\right)d\phi^S}{\int_0^{2\pi} e^{\nu^2/4}\mathrm{D}_{(-2\mu)}(\nu)p\left(\phi^S|\widetilde{\phi^S}\right)d\phi^S}, \quad (15)$$

which is notationally very similar to AUP (10), differing only in the exponential term $e^{j\phi^S}$ in the numerator of (15). Note that in general we expect that $\widehat{A^\beta} \neq |\widehat{S^{(\beta)}}|$, i.e., that the amplitude of the CUP estimate differs from the amplitude estimate obtained via AUP. As for AUP, also for CUP, no closed-form solution has been found for a von Mises phase prior [18]. Thus (10) is solved numerically and tabulated to allow for real-time processing. The final estimate is obtained via $\widehat{S}_{\mathrm{CUP}} = |\widehat{S^{(\beta)}}|^{1/\beta}\frac{\widehat{S^{(\beta)}}}{|\widehat{S^{(\beta)}}|}$. Note that in general, the phase of $\widehat{S}_{\mathrm{CUP}}$ is not the initial phase estimate $\widetilde{\phi^S}$.

B. *Perfectly Known Speech Phase ($\kappa \to \infty$)*

Analogous to the amplitude estimator ADP in (12), the complex estimator CUP for $\kappa \to \infty$ reduces to

$$\widehat{S}_D^{(\beta)} = \mathrm{E}\left(A^\beta e^{j\Phi^S} \mid y, \phi^S\right) = \mathrm{E}\left(A^\beta \mid y, \phi^S\right)e^{j\Phi^S} = \widehat{A}_D^\beta e^{j\Phi^S}, \quad (16)$$

which we denote as the estimator of (C)omplex spectral speech coefficients given (D)eterministic (P)hase information (CDP). Interestingly, comparing (12) and (16), for the case of full certainty in the initial phase, the estimator of the clean amplitude AUP yields exactly the amplitude of the complex estimator CUP, i.e., $|\widehat{S}_D| = \widehat{A}_D$. This is a major difference to traditional phase-blind approaches, where, for example, the amplitude of the Wiener filter is *not* the amplitude obtained with the STSA [1], even though both the Wiener filter and the STSA are based on the same complex Gaussian models for the speech and noise spectral coefficients. While CUP estimates the complex coefficients of clean speech, AUP only estimates the amplitudes. Thus, when the phase is perfectly known, i.e., $\kappa \to \infty$, the CUP spectral phase estimate corresponds to the clean speech phase (see (16)), while in AUP still the noisy phase is used for reconstruction.

C. *Phase-Blind ($\kappa = 0$)*

For the estimation of compressed *complex* coefficients, to the best of our knowledge, no phase-blind estimator of $f(S) = S^{(\beta)}$ for non-Gaussian speech priors has been proposed in the literature. To find a closed form solution for this novel estimator, we insert (5), (7), and the uniform phase prior ($\kappa = 0$) into (4), yielding

$$\widehat{S}_B^{(\beta)} = \frac{\int_0^\infty a^{2\mu-1+\beta}e^{-Ca^2}\int_0^{2\pi} e^{j\phi^S}\, e^{\frac{2ra}{\sigma_V^2}\cos\left(\phi^S-\phi^Y\right)}d\phi^S da}{\int_0^\infty a^{2\mu-1}e^{-Ca^2}\int_0^{2\pi} e^{\frac{2ra}{\sigma_V^2}\cos\left(\phi^S-\phi^Y\right)}d\phi^S da}, \quad (17)$$

with $C = \frac{\mu\sigma_V^2+\sigma_S^2}{\sigma_S^2\sigma_V^2}$. For solving the integral over ϕ^S in the numerator, we substitute ϕ^S by $\phi = \phi^S - \phi^Y$, which leads to

$$e^{j\phi^Y}\int_{-\phi^Y}^{2\pi-\phi^Y} \left(\cos\left(\phi\right)+j\sin\left(\phi\right)\right)\exp\left(\frac{2ra}{\sigma_V^2}\cos\left(\phi\right)\right)d\phi. \quad (18)$$

Since $\sin(\phi)$ is 2π-periodic and odd while the exponential is 2π-periodic and even on the same interval, the integral over the imaginary part is zero. The integral over the real part — as well as the integral in the denominator – can be solved using

$$I_n\left(p\right) = \frac{1}{2\pi}\int_0^{2\pi}\cos\left(nz\right)\exp\left(p\cos\left(z\right)\right)dz. \quad (19)$$

Accordingly, (17) becomes

$$\widehat{S}_B^{(\beta)} = \frac{\int_0^\infty a^{2\mu-1+\beta}\exp\left(-\frac{\mu\sigma_V^2+\sigma_S^2}{\sigma_S^2\sigma_V^2}a^2\right)I_1\left(\frac{2ra}{\sigma_V^2}\right)da}{\int_0^\infty a^{2\mu-1}\exp\left(-\frac{\mu\sigma_V^2+\sigma_S^2}{\sigma_S^2\sigma_V^2}a^2\right)I_0\left(\frac{2ra}{\sigma_V^2}\right)da}e^{j\phi^Y}. \quad (20)$$

Substituting $x = a^2$ (leading to $da = \frac{dx}{2a}$) and using [23, (6.643.2),(9.220.2)] we get

$$\boxed{\widehat{S}_B^{(\beta)} = \frac{\Gamma\left(\mu+\frac{\beta+1}{2}\right)}{\Gamma\left(\mu\right)}\frac{M\left(\mu+\frac{\beta+1}{2};2;\gamma\frac{\xi}{\mu+\xi}\right)}{M\left(\mu;1;\gamma\frac{\xi}{\mu+\xi}\right)} \times\left(\sigma_V^2\right)^{\frac{\beta-1}{2}}\left(\frac{\xi}{\mu+\xi}\right)^{\frac{\beta+1}{2}}Y,} \quad (21)$$

with the confluent hypergeometric function $M(\cdot;\cdot;\cdot)$ and the a posteriori SNR $\gamma = \frac{|Y|^2}{\sigma_V^2}$.

Again, we obtain the final estimator of the complex clean speech spectral coefficients by reversing the compression, i.e., $\widehat{S}_B = |\widehat{S}_B^{(\beta)}|^{1/\beta}\exp(j\angle\widehat{S}_B^{(\beta)}) = |\widehat{S}_B^{(\beta)}|^{1/\beta}\exp(j\Phi^Y)$. Note that in general, as opposed to $\kappa \to \infty$, in the phase-blind case we have $|\widehat{S}_B| \neq \widehat{A}_B$. We refer to this estimator as the phase-(B)lind (E)stimator of (CO)mplex (CO)efficients (BECOCO).

Besides being the special case of CUP for $\kappa = 0$, the estimator (21) can also be interpreted as the complex-valued counterpart to the phase-blind amplitude estimator MOSIE (13). It is further an extension to the super-Gaussian estimator of S in [6, eq.(17)], in the sense that it also incorporates a parameterized error function (3) in addition to the flexible prior for the speech amplitudes (5).

In Table I we provide an overview of the complex estimator CUP and the amplitude estimator AUP. Entries that have been blank before and have been filled as a contribution of this paper are highlighted, i.e., AUP (10) and the phase-blind estimator of the complex speech coefficients BECOCO (21).

V. Analysis

A. *Phase-Blind ($\kappa = 0$)*

We first have a closer look at the novel phase-blind estimator BECOCO (21) that arises as a special case of CUP (15) for $\kappa = 0$. In Fig. 2 we present its input-output characteristic (IOC) [25] for two choices of μ and β together with those of its amplitude

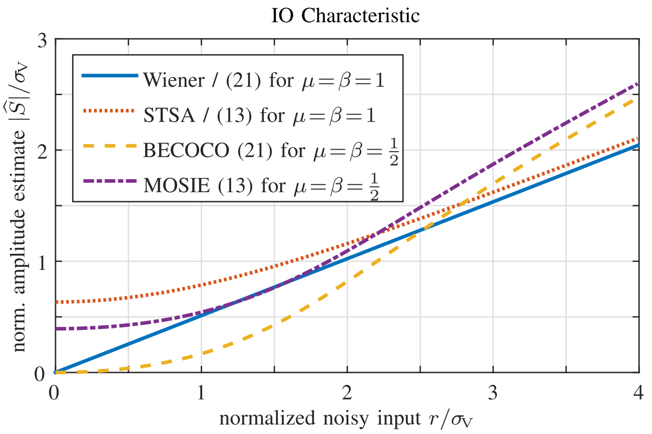

Fig. 2. Input-output characteristic of the phase-blind estimators for $\xi = 0.2$. For $\mu = \beta = 1$, (21) reduces to the Wiener filter, while (13) reduces to the STSA. We further present the curves for $\mu = \beta = 0.5$, where more suppression is applied at low normalized inputs and less suppression at large normalized inputs as compared to their Gaussian counterparts.

counterpart MOSIE (13). The IOC of an estimator presents the amplitude of the clean speech estimate that is obtained for the respective noisy input on the abscissa. To make the analysis independent of an absolute scaling, the input and the output are both normalized by σ_V .

It has been shown in [7] that the phase-blind amplitude estimator (13) reduces to the STSA [1] when using a Gaussian speech prior ($\mu = 1$) and optimizing for uncompressed amplitudes ($\beta = 1$). The novel complex estimator BECOCO (21) in turn reduces, when inserting $\mu = \beta = 1$ into (21) and using $M(\alpha; \alpha; z) = e^z$, to the Wiener filter, which is indeed the MMSE-optimal phase-blind estimator of S for a Gaussian speech prior. We also present the curves for $\mu = \beta = 0.5$, which has been reported in [7] to provide good perceptual results. Compared to their Gaussian counterparts, both estimators apply more suppression to low normalized inputs and less suppression to large normalized inputs. While low inputs r/σ_V are more likely in noise dominated time-frequency regions, large inputs are more likely if speech is present. Still, large inputs can also be caused by noise outliers. The reduced attenuation of such large inputs by MOSIE and BECOCO for $\mu = \beta = 0.5$ thus results in a better protection of the speech component at the price of an increased risk of musical noise. It is well known [1] that in the Gaussian and uncompressed case the complex estimator (Wiener) is more aggressive than the corresponding amplitude estimator (STSA). Based on the IOCs it can more generally be stated that the complex estimator BECOCO (21) is more aggressive than the amplitude estimator MOSIE (13), for all valid combinations of μ and β.

For specific values of μ and β, the amplitude estimator MOSIE (13) is known to resemble many other well-known solutions. See [24] for a detailed list. Accordingly, the complex estimator (21) now also yields complex counterparts to all these amplitude estimators, including the log-spectral amplitude estimator [2] for $\mu = 1$ and $\beta \to 0$ [3]. This highlights the generality of CUP and AUP, which do not only allow for phase-aware speech enhancement, but also yield very general phase-blind estimators for $\kappa = 0$.

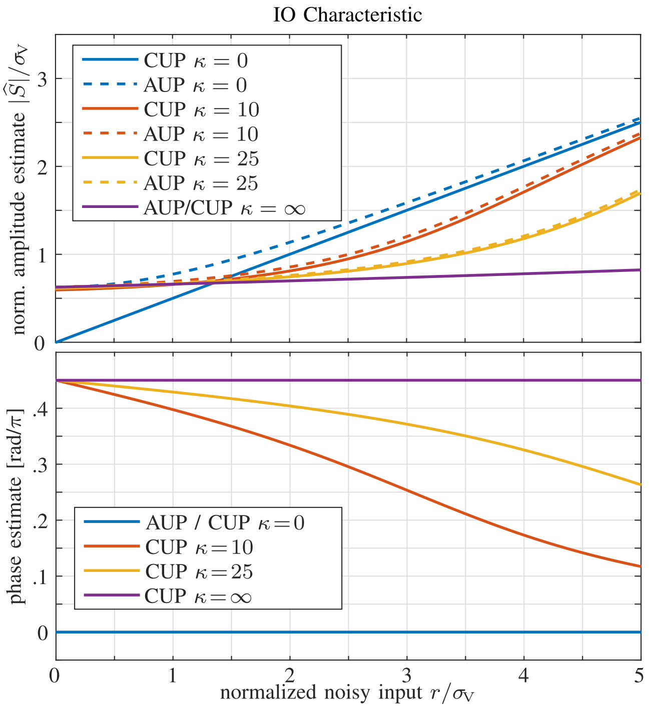

Fig. 3. IOCs and phases of AUP and CUP for $\mu = \beta = 1$, $\xi = 1$, a phase difference of $\Delta\widetilde{\phi} = 0.45\pi$, i.e., $\widetilde{\phi^S} = 0.45\pi$ and $\phi^Y = 0$, and various concentration parameters κ.

B. Phase-Aware ($\kappa \geq 0$)

We now consider the general and more interesting case of $\kappa \geq 0$, i.e., we have some certainty in the prior phase information and CUP and AUP are both truly phase-aware.

In Fig. 3 we investigate how the behavior of the estimators change for an increasing certainty κ. The initial phase is set to $\widetilde{\phi^S} = 0.45\pi$ and the observed noisy phase to $\phi^Y = 0$. We present both, the IOCs and the phase of the corresponding estimate. As argued in Section V-A, for $\kappa = 0$ the IOCs of CUP and AUP significantly differ (Fig. 2). For the other extreme, $\kappa \to \infty$, we know from comparing (16) and (12) that the amplitude-IOCs are the same, but also that CUP provides the initial phase estimate $\widetilde{\phi^S}$ while AUP combines its amplitude estimate with the noisy phase, independent of the value of κ. Accordingly, the differences between CUP and AUP are dominated by the different amplitude estimates for small κ and by the phase estimates for large κ. For intermediate κ, the two estimators yield both, different amplitude estimates and different phase estimates.

For low inputs r/σ_V , which are more likely in noise dominated time-frequency points, the observed phase ϕ^Y is likely to be heavily corrupted. Larger normalized inputs, or a posteriori SNRs $\gamma = r^2/\sigma_V^2$, are more likely to stem from speech and hence ϕ^Y is likely to be relatively close to the clean speech phase. Thus, the influence of the prior phase information reduces towards larger a posteriori SNRs γ (except for $\kappa \to \infty$) and CUP and AUP approach their phase-blind counterparts. The main improvement over phase-blind approaches hence comes at lower a posteriori SNRs, where the initial phase estimate is

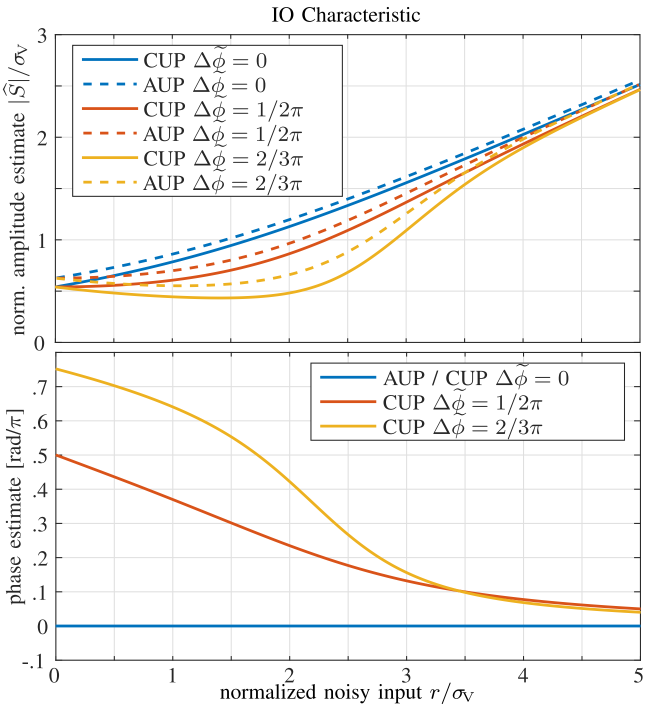

Fig. 4. IOCs and phases of AUP and CUP for $\mu = \beta = 1$, $\xi = 1$, $\kappa = 4$, and three different phase differences. For all curves, the noisy phase is $\phi^Y = 0$.

more reliable than the noisy phase. The same also holds for the phase estimated by CUP, which approaches the (increasingly less noisy) observed phase with increasing a posteriori SNRs for $\kappa < \infty$. For undisturbed clean speech, both, CUP and AUP, accordingly yield the observed clean speech phase.

In Fig. 4, we now present the IOCs and phases of CUP and AUP for different phase differences $\Delta\widetilde{\phi} = \widetilde{\phi^S} - \phi^Y$ and a fixed certainty of $\kappa = 4$. It can be stated that the general behavior of AUP and CUP in terms of their IOCs is similar: the more the observed phase differs from the initial phase estimate, the more suppression is applied. For large $\Delta\widetilde{\phi}$, where the noisy phase differs significantly from the reasonably certain initial phase estimate $\widetilde{\phi^S}$, it is more likely that the respective time-frequency point is dominated by the noise rather than speech and more suppression is applied. Hence, the initial phase yields valuable information to distinguish speech from noise, allowing for improvements in speech enhancement with respect to conventional phase-blind approaches, see e.g., [18].

In general, the IOC of AUP is less aggressive than that of CUP, independent of the phase difference $\Delta\widetilde{\phi}$. For CUP however, the effect of using the estimated phase to synthesize the final enhanced time domain signal is not covered by the IOC. For overlapping signal segments, a modified spectral phase results in a different superposition of neighboring time-frequency points. This can lead to both, a destructive superposition of noise components as well as a constructive superposition of the speech component, possibly achieving an increased noise reduction but also an improved speech preservation. See e.g., [14] for a more detailed discussion. At the same time, modifications of the spectral phase can be sensitive to errors, e.g., [13]. In practice, strong errors in the initial phase, when accompanied by an overestimated certainty κ, may also affect the final phase estimate of CUP and potentially introduce undesired artifacts. This is not the case for the amplitude estimator AUP, which uses the noisy phase for signal synthesis, making it more robust to estimation errors in κ and $\widetilde{\phi^S}$.

VI. Evalution

For the evaluation and comparison of the estimators we use 128 gender-balanced sentences from the TIMIT database [26] sampled at $f_s = 16$ kHz and add different noise types at SNRs ranging from -5 dB to 15 dB. We consider stationary pink noise, pink noise modulated at a frequency of 0.5 Hz, factory noise [27], and babble noise [27]. The results are averaged over all four noise types to allow for a compact and general comparison. The noise variance σ_V^2 is estimated with the speech presence probability based approach in [28], while the speech variance σ_S^2 is obtained using the decision-directed approach [2] with a smoothing factor of 0.96. In the original proposal [2], a smoothing factor of 0.98 is recommended, but here we lowered the smoothing factor to reduce speech distortions, especially at speech onsets, at the price of slightly more musical noise. We set the form parameter in (5) to $\mu = 0.5$, modeling a heavy-tailed distribution of amplitudes, which corresponds to a super-Gaussian distribution of S. In [7], this value has been reported to yield a good trade-off in terms of outliers and clarity of speech. To consider the compressive character of the human auditory system in the estimators, we further set the compression parameter to $\beta = 0.5$ as proposed in [7]. For analysis and synthesis, we use square-root Hann windows of 32 ms with an overlap of 75 %, corresponding to a segment length of $N = 512$ samples and a segment shift of $L = 128$ samples. To increase the perceptual quality of the enhanced signal, we further limit the maximum attenuation in each time-frequency point to -15 dB.

To study the spectral phase and the spectral amplitude of the different estimators in isolation, we employ three different measures. First, we evaluate the accuracy of the phase of the final clean speech estimate $\angle\widehat{S}_{k,\ell}$ by means of the phase SNR (PSNR) [15]

$$\mathrm{PSNR} = 10\log_{10}\left\{\frac{\sum_{k,\ell} A_{k,\ell}^2}{\sum_{k,\ell} A_{k,\ell}^2\left(1-\cos\left(\phi_{k,\ell}^S - \angle\widehat{S}_{k,\ell}\right)\right)}\right\}. \tag{22}$$

The closer the estimated phase resembles the true clean speech phase, the larger the phase SNR (PSNR). The amplitude weighting puts emphasize on the phase of speech components with relevant signal energy, where phase errors are arguably perceptually most relevant. Secondly, we evaluate the segmental noise reduction (NR) and the segmental speech SNR (SSNR) [29], which give an idea of how much noise is suppressed and how well the speech is preserved, respectively. All analyzed estimators can be expressed by a gain that is multiplicatively applied to the noisy input in each time-frequency point. NR is obtained by applying the absolute value of this spectral gain to the noise signal, whereas for SSNR it is applied to the clean speech signal. In the time-domain, the two signals are then compared to the clean speech signal or the noise signal as detailed in [29] to obtain the final SSNR and NR, respectively. To enable a separate analysis of amplitude and phase effects, for NR and SSNR we apply the absolute value of the gain functions. NR and SSNR thus only depend on the amplitudes of the speech estimates, while phase effects are evaluated using PSNR. Lastly, we also employ two measures that are commonly used in the context of speech enhancement, namely perceptual evaluation of speech quality (PESQ) [30] and short-time objective intelligibility measure (STOI) [31]. These measures now consider both, the enhancement of the spectral amplitude and of the spectral phase. We map the raw STOI output values to actual intelligibility scores by applying the mapping function that has been proposed for the IEEE database used in [31]. To improve the visualization and ease the comparison between the different algorithms, we do not plot the absolute values of PESQ and STOI, but rather the improvement over the noisy input.

A. Oracle $\widetilde{\phi^S}$ and κ

To facilitate the analysis and comparison of the different approaches on real speech data, here we artificially create initial phase estimates $\widetilde{\phi^S}$ that follow a von Mises distribution (8) with a given certainty κ, centered around the true clean speech phase ϕ^S. For this, we first draw one realization for each time-frequency point (k, ℓ) from a von Mises distributed random variable with a mean direction of 0 and the desired certainty κ. To obtain the final initial phase estimate $\widetilde{\phi^S_{k,\ell}}$, each realization is then shifted by the respective clean speech phase $\phi^S_{k,\ell}$ to obtain the desired mean direction $\widetilde{\phi^S_{k,\ell}} = \phi^S_{k,\ell}$. This gives us the necessary flexibility for a thorough evaluation and allows us to analyze the effects of incorporating phase information in detail, circumventing the limitations of current phase estimators like [14] or [15]. The exact knowledge of the distribution of $\widetilde{\Phi^S}$ is the only oracle information that is employed in these experiments. All other parameters, like σ_S^2 and σ_V^2, are still estimated from the noisy microphone signal Y. A completely blind approach is presented in the next section, where $\widetilde{\phi^S}$ is estimated from the noisy observation Y and κ is adapted as a function of the probability of a frame being voiced.

In Fig. 5, we present results for three different values of certainty κ, increasing from 0.1 to 2 to 100 from left to right. As a well-known reference, we also present the results for the Wiener filter. Regarding the phase-blind approaches, which are independent of κ, we can see that the estimator of the spectral amplitudes MOSIE (13) is less aggressive than the complex estimator BECOCO (21) and the Wiener filter, as it achieves a higher SSNR but also a lower NR. Furthermore, it can be stated that BECOCO, while achieving a similar SSNR, achieves slightly more noise reduction than the Wiener filter. Since all phase-blind estimators, as well as the phase-aware amplitude estimators AUP and ADP use the noisy phase for signal reconstruction, they all depict the same PSNR as the noisy input signal, which linearly increases from about 8 dB to 22 dB for increasing input SNRs. For $\kappa = 0.1$ at the left of Fig. 5, which reflects a rather unreliable initial phase estimate, $\widetilde{\phi^S}$ yields only very little information and the achievable benefit of phase-aware speech enhancement is limited. Thanks to the incorporation of this uncertainty of the initial phase in the estimators AUP and CUP, both approach their phase-blind counterparts, effectively neglecting the strongly corrupted initial phase information. The phase-aware estimators ADP (12) and CDP (16) on the other hand ignore the uncertainty and assume that the provided initial phase $\widetilde{\phi^S}$ yields the exact clean speech phase. Consequently, given an unreliable phase estimate, i.e., $\kappa = 0.1$, ADP and CDP yield the worst results, with a very aggressive amplitude suppression. This over-attenuation causes clearly perceptible speech distortions, which is also reflected in NR and SSNR. While both estimators provide the exact same amplitude, the complex estimator CDP (16) additionally uses the corrupted initial phase for signal synthesis, leading to a very low PSNR and also the lowest scores in PESQ and STOI. The results again highlight the importance of considering the uncertainty in $\widetilde{\phi^S}$.

With an increasingly certain initial phase, also the potential gain of phase-aware speech enhancement increases. For $\kappa = 2$, PESQ and STOI predict improvements in quality of up to 0.2 MOS and in intelligibility of 25 % over the phase-blind estimators at -5 dB SNR. These improvements are most pronounced for the complex estimators CUP (15) and CDP (16). While in high SNRs CUP yields the highest speech quality according to PESQ, neglecting the uncertainty of $\widetilde{\phi^S}$ in CDP, despite introducing artifacts, seems to benefit speech intelligibility in low SNRs according to STOI.

When a very reliable initial phase estimate is available, i.e., $\kappa = 100$ at the right of Fig. 5, the potential gain over phase-blind approaches is the largest. In this case, CUP and AUP approach CDP (16) and ADP (12), which assume $\kappa \rightarrow \infty$. The amplitude estimator AUP achieves an improvement of around 0.2 MOS in PESQ and 20 % in STOI over the Wiener filter at -5 dB SNR. Using the complex estimator CUP increases the gain over the phase-blind approaches further, to 0.4 MOS in PESQ and more than 35 % in STOI. These remarkable improvements again stress the relevance of utilizing phase information for speech enhancement. Based on informal listening, the benefit of the phase-aware estimators lies in an increased noise reduction, especially in non-stationary noises, while the speech component is preserved.

As the amplitudes of AUP and CUP are virtually the same for $\kappa = 100$, according to Fig. 3, the performance gain between the two is only due to the modification of the spectral phase. Interestingly, the phase of CUP is not only more accurate than that of AUP and the phase-blind approaches, but even more accurate than that of CDP (16), which uses the initial phase estimate for signal reconstruction. Using the very reliable phase estimate of CUP in the overlap-add synthesis stage leads to a constructive superposition of the speech and a destructive superposition of the residual noise of adjacent signal segments. The complex estimator CUP thus achieves the largest noise reduction. However, for very reliable initial phases, i.e., $\kappa = 100$, using the modified phase in time-frequency regions where

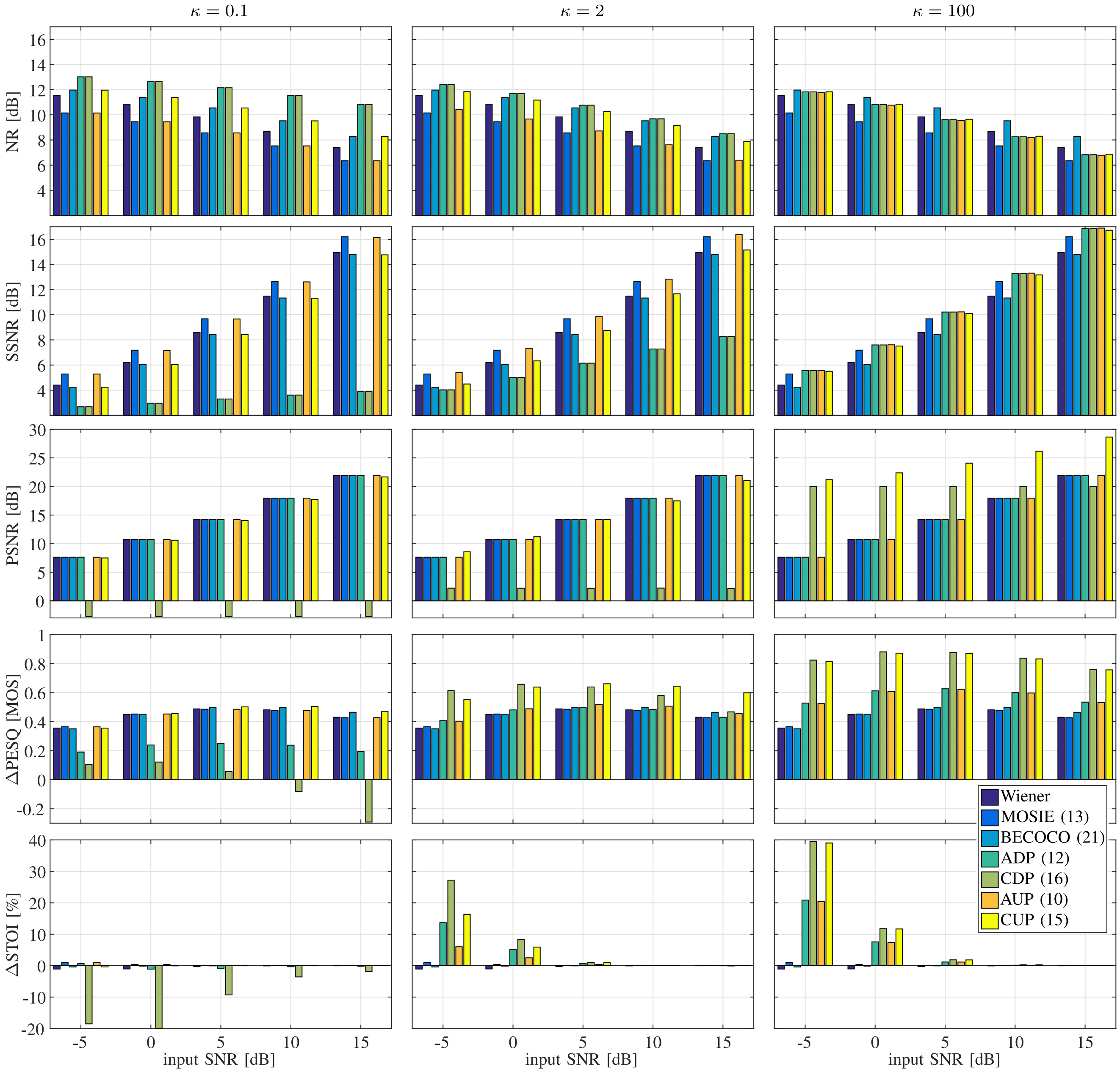

Fig. 5. NR, SSNR, PSNR, as well as PESQ improvements and STOI improvements relative to the noisy input signal, averaged over four noise types (pink, modulated pink, factory, babble). We set $\mu = \beta = 0.5$. From left to right, the quality of the initial phase estimate $\widetilde{\phi}^S$ increases, i.e., its concentration around the true clean speech phase increases from $\kappa = 0.1$ to $\kappa = 2$ and $\kappa = 100$.

the noise is dominant but not sufficiently suppressed can lead to artifacts in the enhanced signal. For more realistic situations with lower κ, this is however less problematic, since the phase of CUP is closer to the noisy phase. The phase-aware amplitude estimators AUP and ADP always use the noisy phase for signal synthesis and effectively avoid any phase-artifacts.

In practice, both the initial phase $\widetilde{\phi}^S$ and its certainty κ need to be estimated in order to compute CUP and AUP. If κ is overestimated, the estimators rely too much on the initial phase, which may result in signal degradations as observed for ADP (12) and CDP (16) on the left of Fig. 5. Underestimating κ on the other hand diminishes the performance of CUP and AUP that could be achieved with the available initial phase, eventually reducing to that of the respective phase-blind estimator.

The general trends observed in Fig. 5 are representative for each of the four evaluated noise types. The benefit of the phase-aware estimators, however, is the largest for non-stationary noises, especially in terms of PESQ, where the additional phase information allows for a better suppression of noise outliers, like babble bursts, as discussed in Section V-B.

We performed an analysis of variance (ANOVA) in conjunction with a post hoc Tukey's range test to analyze the results for statistical significant differences between the different algorithms at the $p < 0.05$ level. For $\kappa = 2$ and $\kappa = 100$, the improvements for the best performing phase-aware algorithm over all phase-blind approaches in PESQ are statistically significant for all SNRs and noise types. The same holds for the STOI improvements at -5 dB and 0 dB input SNR.

B. Blind Estimation of $\widetilde{\phi^S}$

In this section, we consider the more practical case that the initial phase is estimated from the noisy microphone signal. Specifically, the initial phase $\widetilde{\phi^S}$ is obtained using [14], which is based on a harmonic signal model for the clean speech. In [14], phase estimation along time, along frequency, and a combination of both has been proposed. For the application at hand, it showed that the estimation along frequency yields the most promising results. The fundamental frequency, which is needed to compute $\widetilde{\phi^S}$, is estimated with the noise-robust fundamental frequency estimator PEFAC [32] on the noisy observation. The simple harmonic model of [14] fits well for voiced sounds, where it may yield reliable initial phase estimates. However, it is less suited for other sounds like fricatives or even speech absence. We hence set the certainty κ used to compute CUP and AUP in each time-frequency point (k, ℓ) according to [18]

$$\kappa(k,\ell) = \begin{cases} 4\,P_V(\ell), & \text{for } kf_s/N < 4 \text{ kHz} \\ 2\,P_V(\ell), & \text{for } kf_s/N \geq 4 \text{ kHz}, \end{cases} \tag{23}$$

with the probability that the signal segment ℓ contains voiced speech $P_V(\ell)$, which is also estimated with PEFAC. The higher the probability that the underlying speech sound is voiced, the more we trust our initial phase estimate and increase κ. Furthermore, it is commonly assumed that the harmonic model – and thus also the phase estimates obtained with it – is less accurate for high frequencies than for low frequencies, partly due to fundamental frequency estimation errors that may accumulate over frequency. To take this into account, we reduce κ above 4 kHz in (23). The values 2 and 4 in (23) have been proposed in [18] and where chosen via informal listening such that a good subjective quality is achieved for CUP. To allow for a fair comparison, we employ the two phase-aware estimators ADP (12) and CDP (16) that neglect the uncertainty in the initial phase estimate only in signal segments that contain voiced speech sounds, which are detected using [32]. In the remaining segments, we use the respective phase-blind counterpart MOSIE (13) or BECOCO (21).

The results for the blind setup are presented in Fig. 6. On the left, only voiced speech is evaluated, for which the employed estimator of the initial phase [14] has originally been designed for, while on the right the complete signals are taken into account. As the phase-blind estimators (Wiener, Mosie (13), and BECOCO (21)) are independent of the initial phase estimate, the results are the same as in the oracle experiment in Fig. 5. The complex phase-aware estimator CUP is again more aggressive than the phase-aware amplitude estimator AUP in terms of NR and SSNR. The phase estimate of CUP further achieves a much higher PSNR than the complex phase-aware estimator CDP (16) that assumes $\kappa \rightarrow \infty$, but it is still lower than for the noisy phase used by the Wiener filter, MOSIE (13), BECOCO (21), ADP (12), and AUP (10). Nevertheless, phase modifications in low SNRs, which are hardly reflected in PSNR, still lead to some additional noise reduction after overlap-add. The complex estimator CUP consistently yields the highest PESQ scores, with an improvement of more than 0.1 MOS in PESQ over the Wiener filter for all SNRs in voiced speech. When evaluated over the complete signals, the gain reduces to some degree, especially towards higher input SNRs. In this setup, the amplitude estimator AUP yields only little improvements in PESQ over the phase-blind approaches.

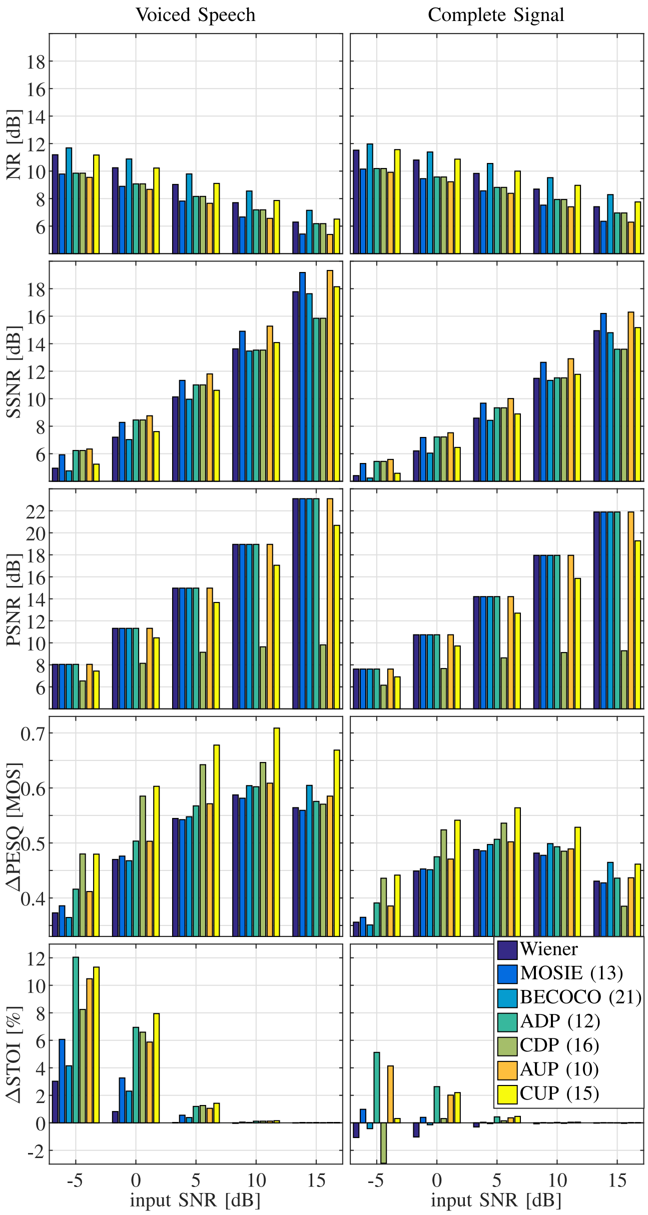

Fig. 6. NR, SSNR, PSNR, as well as ΔPESQ and ΔSTOI, averaged over four noise types (pink, modulated pink, factory, babble). On the left, only voiced speech has been evaluated, for which the employed estimator of the initial phase [14] has originally been designed for. On the right, the complete signals have been evaluated. Again we set $\mu = \beta = 0.5$. The initial phase $\widetilde{\phi^S}$ is blindly estimated on the noisy observation and the concentration parameter κ is obtained via (23).

On the bottom left of Fig. 6, for voiced speech STOI predicts an intelligibility improvement for all four phase-aware estimators over the conventional phase blind approaches at low SNRs. When evaluating the complete signals (bottom right), however, only the phase-aware amplitude estimators AUP and ADP also improve STOI at negative SNRs, where speech intelligibility improvement is most relevant. A reason for this is that at very low SNRs the accuracy of the blindly estimated initial phase and also of the voicing probability $P_V(\ell)$ decreases, corrupting the estimation of $\widetilde{\phi^S}$ and κ via (23). For instance, strong interfering speakers in babble noise can cause an overestimation of the voicing probability and thus also of κ (23) during unvoiced speech or speech absence. The drop in predicted intelligibility in negative SNRs for the complex estimators suggests that modifying the spectral phase of the enhanced signal is more sensitive to such erroneous phase information than phase-aware amplitude enhancement alone.

To investigate the statistical significance of these results, we use the same method as for the oracle experiments in the previous section. We found that the improvements in PESQ for CUP over the phase-blind approaches are statistically significant for SNRs lower or equal to 10 dB for voiced speech and for SNRs lower or equal to 5 dB for the complete signals, except for babble noise at -5 dB. The STOI improvements at -5 dB and 0 dB of the best performing phase-aware algorithm over all phase-blind estimators are also significant, except for babble noise at -5 dB for the complete signals.

Comparing the outcome of the blind experiments to the results of the oracle experiments in Fig. 5, we can state that the complex phase-aware enhancement of CUP has a better performance than AUP in the oracle case, but at the same time it is also less robust to errors in practical scenarios. The performance gap between the oracle experiments and the blind experiments further highlights the relevance of an accurate estimation of the initial phase and its uncertainty. Considering the renewed interest in the role of the spectral phase and the recent advances in phase estimation, e.g., [15], [20], [21], [33], [34], we believe that significant improvements in the estimation of the clean speech phase can be expected in the near future. AUP and CUP could both utilize such more accurate prior information, allowing for further improvements over the traditional speech enhancement approaches.

VII. Conclusions

In this paper, we presented two novel clean speech estimators that complete the existing set of phase-aware estimators: a novel amplitude estimator given uncertain prior phase information as well as a closed-form solution for complex coefficients when the prior phase information is completely uncertain or not available. We put the new estimators into the context of existing estimators and analyze the advantages and disadvantages, including their sensitivity to errors in the prior phase information, providing new insights into the matter of phase-aware speech enhancement.

References

[1] 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.

[2] 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.

[3] 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.

[4] R. Martin, “Speech enhancement based on minimum mean-square error estimation and super-Gaussian priors,” *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 845–856, Sep. 2005.

[5] 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.

[6] 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, Jan. 2008.

[7] C. Breithaupt, M. Krawczyk, and R. Martin, “Parameterized MMSE spectral magnitude estimation for the enhancement of noisy speech,” in *IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2008, pp. 4037–4040.

[8] R. C. Hendriks, R. Heusdens, and J. Jensen, “Log-spectral magnitude MMSE estimators under super-Gaussian densities,” in *Proc. 10th Annu. Conf. Int. Speech Commun. Assoc.*, Sep. 2009, pp. 1319–1322.

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

[10] Y. Lu and P. C. Loizou, “A geometric approach to spectral subtraction,” *Speech Commun.*, vol. 50, no. 6, pp. 453–466, 2008.

[11] T. Kleinschmidt, S. Sridharan, and M. Mason, “The use of phase in complex spectrum subtraction for robust speech recognition,” *Comput. Speech Lang.*, vol. 25, no. 3, pp. 585–600, 2011.

[12] Y. Zhang and Y. Zhao, “Real and imaginary modulation spectral subtraction for speech enhancement,” *Speech Commun.*, vol. 55, pp. 509–522, 2013.

[13] N. Sturmel and L. Daudet, “Signal reconstruction from STFT magnitude: A state of the art,” in *Int. Conf. Digit. Audio Effects*, Sep. 2011, pp. 375–386.

[14] M. Krawczyk and T. Gerkmann, “STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1931–1940, Dec. 2014.

[15] P. Mowlaee and J. Kulmer, “Harmonic phase estimation in single-channel speech enhancement using phase decomposition and SNR information,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 9, pp. 1521–1532, Sep. 2015.

[16] 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.

[17] P. Mowlaee and R. Saeidi, “Iterative closed-loop phase-aware single-channel speech enhancement,” *IEEE Signal Process. Lett.*, vol. 20, no. 12, pp. 1235–1239, Dec. 2013.

[18] T. Gerkmann, “Bayesian estimation of clean speech spectral coefficients given a priori knowledge of the phase,” *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4199–4208, Aug. 2014.

[19] S. Vanambathina and T. K. Kumar, “Speech enhancement by Bayesian estimation of clean speech modeled as super Gaussian given a priori knowledge of phase,” *Speech Commun.*, vol. 77, pp. 8–27, 2016.

[20] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, “Phase processing for single channel speech enhancement: History and recent advances,” *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 55–66, Mar. 2015.

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

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

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

[24] C. Breithaupt and R. Martin, “Analysis of the decision-directed SNR estimator for speech enhancement with respect to low-SNR and transient conditions,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 2, pp. 277–289, Feb. 2011.

[25] J. E. Porter and S. F. Boll, “Optimal estimators for spectral restoration of noisy speech,” in *IEEE Int. Conf. Acoust. Speech Signal Process.*, Mar. 1984, pp. 18A.2.1–18A.2.4.

[26] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, *DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CD-ROM. NIST Speech Disc 1-1.1*. Gaithersburg, MD, USA, 1993.

[27] 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,” *Speech Commun.*, vol. 12, pp. 247–251, Jul. 1993.

[28] 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.

[29] 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.

[30] ITU-T, “Perceptual evaluation of speech quality (PESQ),” *ITU-T Recommendation P.862*, 2001.

[31] 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.

[32] 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.

[33] P. Magron, R. Badeau, and B. David, “Phase reconstruction of spectrograms with linear unwrapping: Application to audio signal restoration,” in *Eur. Signal Process. Conf.*, Sep. 2015, pp. 1–5.

[34] S. M. Nørholm, M. Krawczyk-Becker, T. Gerkmann, S. van de Par, J. R. Jensen, and M. G. Christensen, “Least squares estimate of the initial phases in STFT based speech enhancement,” in *16th Annu. Conf. Int. Speech Commun. Assoc.*, Sep. 2015, pp. 1750–1754.

Martin Krawczyk-Becker (S’15) received the Dipl.-Ing. degree in electrical engineering and information sciences from the Ruhr-Universität Bochum, Bochum, Germany, in 2011, and the Dr.-Ing. degree from the Faculty of Medicine and Health Sciences, Universität Oldenburg, Oldenburg, Germany, in 2016. From January 2010 to July 2010, he was with Siemens Corporate Research, Princeton, NJ, USA. Currently, he is a Postdoctoral Researcher at the University of Hamburg, Hamburg, Germany. [...] His research interests include digital signal processing algorithms for speech and audio, with a focus on speech enhancement and noise reduction.

Timo Gerkmann (S’08–M’10–SM’15) studied electrical engineering and information sciences at the universities of Bremen and Bochum, Germany. He received his Dipl.-Ing. degree in 2004 and his Dr.-Ing. degree in 2010 both from the Faculty of Electrical Engineering and Information Sciences, Ruhr-Universität Bochum, Bochum, Germany. In 2005, he spent six months with Siemens Corporate Research, Princeton, NJ, USA. From 2010 to 2011, he was a Postdoctoral Researcher in the Sound and Image Processing Laboratory, Royal Institute of Technology (KTH), Stockholm, Sweden. From 2011 to 2015, he was a Professor of speech signal processing with the Universität Oldenburg, Oldenburg, Germany. From 2015 to 2016, he was the Principal Scientist in Audio & Acoustics, Technicolor Research & Innovation, Hanover, Germany. Since 2016, he has been a Professor of signal processing at the University of Hamburg, Hamburg, Germany. His research interests include digital signal processing algorithms for speech and audio applied to communication devices, hearing instruments, audio-visual media, and human–machine interfaces.