2012 IEEE 27-th Convention of Electrical and Electronics Engineers in Israel

Phase estimation in speech enhancement – unimportant, important, or impossible?

Timo Gerkmann, Martin Krawczyk, and Robert Rehr
Speech Signal Processing, Faculty V, University of Oldenburg, 26111 Oldenburg, Germany
{timo.gerkmann,martin.krawczyk,r.rehr}@uni-oldenburg.de

***Abstract*—In recent years, research in the field of single channel speech enhancement has focused on the enhancement of spectral amplitudes while the noisy spectral phase was left unchanged. In this paper we review the motivation for neglecting phase estimation in the past, and why recent publications imply that the estimation of the clean speech phase may be beneficial after all. Further, we present an algorithm for blindly estimating the clean speech spectral phase from the noisy observation and show that the application of this phase estimate improves the predicted speech quality.**

***Index Terms*—Speech enhancement, phase estimation, noise reduction, signal reconstruction.**

I. Introduction

Single channel speech enhancement describes the improvement of a corrupted speech signal captured with one microphone in a noisy environment, or at the output of a multichannel speech enhancement algorithm. Single channel speech enhancement is particularly difficult when the noise is nonstationary (such as traffic noise), or even speech like (as babble noise). As mobile speech communication devices are often employed in environments with nonstationary noise, recent research focuses on making the algorithms more robust in these noise conditions.

Speech enhancement algorithms usually involve a transformation of the noisy speech into a spectral domain to allow for an easier separation between speech and noise. A typical and efficient candidate is the short-time Fourier transform (STFT) domain. There, speech is segmented into short segments of approximately 10-30 ms, weighted with a tapered spectral analysis window and transformed to the Fourier domain.

We assume that in the STFT domain noisy speech is given by

$$Y(k,\ell) = S(k,\ell) + N(k,\ell)\,, \tag{1}$$

where the noisy speech $Y(k,\ell)$ is a superposition of clean speech $S(k,\ell)$ and noise $N(k,\ell)$. The frequency index is denoted by k while ℓ denotes the time-segment index.

As the STFT coefficients are complex valued, adding the complex noise coefficients $N(k,\ell)$ will distort both the amplitude as well as the phase of the clean speech signal. If we assume that speech and noise are complex-Gaussian distributed the well-known Wiener filter is the optimal estimator in the minimum mean-square error (MMSE) sense. However, the Wiener filter results in a real-valued gain-function which is multiplicatively applied to the noisy STFT coefficients. Thus the Wiener filter alters only the amplitude of the noisy speech while the noisy phase remains unchanged. The same holds for spectral subtraction methods, where an estimate of the noise amplitude is subtracted from the noisy spectral amplitudes (or functions thereof). Hence, also in spectral subtraction only the amplitude of noisy speech is modified, while the noisy phase is left unchanged [1].

As Wiener filtering and spectral subtraction only change the spectral amplitudes [2], the question arose whether speech spectral phase improvement would be a fruitful area of research. In an attempt to answer this question, Wang and Lim have done listening experiments to analyze the perceptual effects of an improved phase as compared to an improved amplitude. The results showed that enhancing spectral amplitudes has a much larger impact than enhancing the spectral phase. Their conclusion resulted in the paper entitled "The unimportance of phase in speech enhancement" [3]. Other researcher followed this line of thinking. Vary reported that in voiced speech distortions of the phase are only perceivable if the local signal-to-noise ratio (SNR) in a time-frequency point is lower than 6 dB [4]. Ephraim and Malah showed that under certain assumptions the noisy speech signal is the optimal estimate of the clean speech phase in the MMSE-sense.

From then on, the estimation of clean speech has focused to a large extent on deriving optimal estimators for the clean speech spectral amplitudes. Examples are estimators for clean speech spectral amplitudes (or their logarithm), assuming Rayleigh priors [5][6]. As speech priors were argued to be heavy tailed [7][8] (and hence not Rayleigh distributed), parameterizable priors were considered that allow to fit the prior models to empirical distributions [9][10][11]. Also parameterizable models for the compression were considered [12], as well as estimators that consider both parameterizable priors and parameterizable compressive functions [13].

While the vast majority of researchers aim at improving only spectral amplitudes, more recently Paliwal et al. have reconsidered the role of phase in speech enhancement by doing similar experiments as Wang and Lim. Interestingly, their conclusion points into the opposite direction as compared to Wang and Lim's work, resulting in a paper entitled "The importance of phase in speech enhancement" [14].

This paper is structured as follows. In Sec. II, we will discuss why some people argue that phase enhancement is not meaningful, and why others believe it is important. In

Sec. III we will discuss if an improvement of the noisy phase is possible at all. In Sec. IV we will show that a blind enhancement of the speech spectral phase from noisy speech is possible during voiced speech. Finally, a combination of phase and amplitude enhancement will be evaluated in Sec. V, an shown to increase the speech quality as predicted by PESQ as compared to amplitude enhancement alone.

II. Is phase estimation important or unimportant?

Noise added to the clean speech spectral coefficients as given in (1) will affect both the amplitude and the phase of the observation. Vary [4] discussed the effect of a disturbed phase for speech perception. For this he computed the STFT representation of a speech signal, and modified its phase before reconstructing the time domain signal. He observed that when the noisy phase is replaced by zeros, the resynthesized speech sounds completely voiced and monotonous, i.e. like having a constant pitch. If the phase is replaced by a random phase, uniformly distributed between $\pm\pi$, a rough, completely unvoiced speech is obtained. If noise is added to the clean speech phase, the speech will sound increasingly rough for a decreasing local SNR. Vary argued that if in voiced sounds and Gaussian noise the local SNR is larger than 6 dB, the resulting phase error is not perceivable [4]. From Vary's experiments we conclude that the phase can not be chosen arbitrarily, but that the noisy phase can be used as a reasonable estimate. However, we also conclude that phase estimation *is* beneficial whenever the local SNR is lower than 6 dB in voiced sounds. Note that this is often the case, e.g. for low power spectral harmonics or between speech spectral harmonics.

Wang and Lim [3] have done some listening experiments to evaluate how important the phase is for speech perception. For this, they generated two noisy speech signals at different SNRs. Then, they computed the STFT of the resulting noisy speech signals. Finally, for resynthesis they used the amplitude from one signal and the phase from the other to create a test stimulus (see Fig. 1). As a result, the degree of distortion was different for the amplitude as compared to the phase. Listeners were asked to compare the test stimulus to a noisy reference speech signal, and set the SNR of the reference such that the perceived quality is the same for the reference and the test stimulus. The result of this experiment was that the SNR gain obtained by mixing noisy amplitudes with the (almost) clean phase resulted in typical SNR improvements of 1 dB or less. Hence, Wang and Lim concluded that phase is unimportant in speech enhancement [3].

Paliwal et al. [14] have done similar experiments as Wang and Lim, but showed that employing the clean speech phase can significantly improve the quality of noisy speech if the segment overlap in the STFT is increased from 50% to 87.5% and zero-padding is applied. From their experiments they argue that "research into better phase spectrum estimation algorithms, while a challenging task, could be worthwhile" and, in contrast to Wang and Lim, entitled their paper "The importance of phase in speech enhancement".

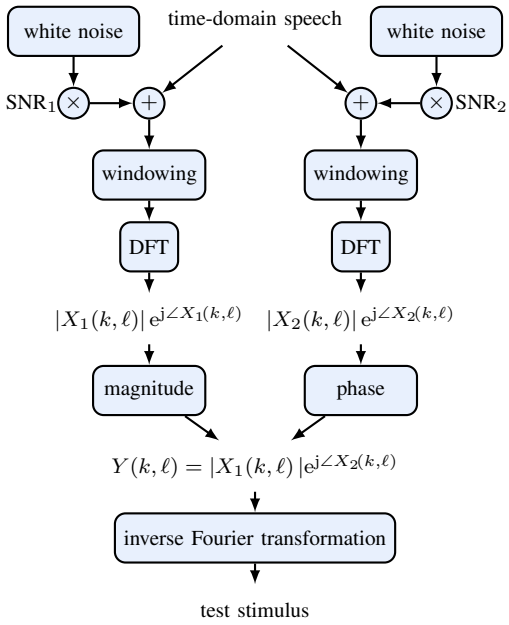

Fig. 1. The experiment of Wang and Lim [3].

While the new results from Paliwal et al. show that phase estimation is an interesting research topic, in the next section we address the question if an improvement of the noisy phase is possible at all.

III. Is phase estimation possible?

The fact that in most state-of-the-art speech enhancement algorithms no phase enhancement is employed, demonstrates that estimating the clean speech phase is a difficult task, and actually a lot more difficult than estimating the amplitude. This has also to do with the fact that the relationship between neighboring phase values in time-frequency space has to be correct. Neglecting these phase relations can lead to nonlinear phase distortions and dispersions [15]. Furthermore, even for noise that is additive in the time domain, phases are not additive, i.e. $\angle Y(k,\ell) \neq \angle S(k,\ell) + \angle N(k,\ell)$.

Already thirty years ago Quatieri, Hayes, Lim and Oppenheim [16][17] were considering ways to obtain the phase of a signal when only the amplitude is known (and vice versa). They showed that for minimum or maximum phase systems, the log-amplitude and the phase are related through the Hilbert transform. Further, Hayes et al. showed that most one-dimensional finite duration signals can be reconstructed from only the phase information [17] up to a scale factor. However, this method is very sensible and needs a very accurate phase estimate [18]. Also iterative methods were proposed to reconstruct a signal from only the phase information [16][17]. Griffin and Lim [19] proposed an iterative algorithm to reconstruct the phase of an STFT signal, when only the amplitude is known. For this, the time domain signal is reconstructed from the given amplitude. Then the signal is reanalyzed yielding a first estimate of the phase. It is shown

that after several iterations the true time-domain signal (and thus the true phase) can be obtained. However, up to 25-100 iterations are required [19], meaning that between 25 and 100 additional discrete Fourier transforms (DFTs) have to be computed. This makes the iterative approach unsuitable for most mobile applications. As observed by Vary [4], for local SNRs larger than 6 dB, the noisy phase is a reasonable estimate of the clean phase. Therefore, the number of iterations of Griffin and Lim's approach can be reduced to around 10 by only estimating the phase when the SNR is low [20]. However, the phase estimation algorithms based on [16][17][19] require knowledge of the clean speech spectral amplitude. It has been observed that in practice estimates of the speech spectral amplitudes do not represent the true amplitudes well enough to converge towards an optimal solution [20]. Further, the iterative algorithms may yield audible artifacts, such as echo, smearing and modulations [20].

As Paliwal et al. [14] have observed that the role of the phase is increasingly important when spectral analysis windows with a reduced dynamic range are employed, they propose to use different spectral analysis windows to obtain the spectral amplitude and phase, respectively. While they use a tapered spectral analysis, e.g. a Hamming window, to estimate amplitudes, the phase is obtained by a Chebyshev window, where the dynamic range can be controlled by an additional parameter. They showed that employing this mixed windowing can increase the quality of noisy speech. However, by applying this modification of the spectral analysis-synthesis scheme the perfect reconstruction property is lost, thus necessarily resulting in signal distortions. Furthermore, while the methods proposed in [14] modify the noisy phase, they are not capable of estimating the clean speech phase directly.

From a statistical point of view, if histograms are computed from STFT-bins that exhibit a similar estimated speech power spectral density, it has been shown that the phase is uniformly distributed and independent of the amplitude [9], [11]. Under these assumptions, it has been shown by Ephraim and Malah, that the MMSE-optimal estimate for the clean speech phase is the noisy phase. This observation tells us that when considering only a certain time-frequency point, the best estimate of the clean speech phase is the noisy phase. When looking at an image of the phase of an STFT domain speech signal, not much structure can be observed in the clean speech phase, which seems to agree with the statement that the noisy phase is the best estimator available. In practice however, we also have access to the phase values of the past, as well as of surrounding frequency bins. In Fig. 2, instead of plotting the phase directly, we have plotted the phase difference between the current frame and the previous frame. Furthermore, we transformed each frequency band into the baseband by multiplying a factor of $\exp(-\mathrm{j}2\pi k\ell L/N)$ to each band, where N is the DFT length, and L is the segment shift. Note that the original phase can still be reconstructed after these modifications. However, by applying these modifications we avoid phase wrapping. As a result, after applying the modification, in the phase representation in the lower left of Fig. 2 clear structures of the phase can be observed that follow nicely the structure of the clean speech spectral amplitudes in the top left of Fig. 2. From these observations we conclude that the noisy phase is only MMSE-optimal when we consider time-frequency points as being independent, and that this assumption may limit the performance of state-of-the-art speech enhancement frameworks.

Motivated by this observation, in [21] we derived an algorithm that is capable of blindly determining the clean speech phase in a direct way, when only noisy speech is given. Furthermore, when the proposed phase estimate is employed for resynthesis of noisy speech, it yields an improved speech quality. This method is outlined in the next section.

IV. Blind estimation of the clean speech phase

If the clean speech signal is deteriorated by additive noise, the afore mentioned structures inherited in the phase during voiced speech are lost to a large extent, as can be seen in the second column of Fig. 2. To blindly reconstruct these characteristic structures based on the noisy observation, a harmonic speech signal model is employed in voiced speech segments, given by

$$s(n) \approx \sum_{h=0}^{H-1} 2A_h \cos(\Omega_h n + \varphi_h), \tag{2}$$

with time index n, harmonic index h, amplitude A_h, time domain phase φ_h, and the number of harmonics H. The normalized angular frequencies are multiples of the fundamental frequency f_0, i.e. $\Omega_h = (h+1)\,2\pi f_0/f_\mathrm{s}$, where f_s denotes the sampling frequency. Assuming that in each STFT-band only the closest harmonic component is relevant, the expected phase shift from one segment to the next is directly related to the harmonic frequency and the segment shift L. This relationship can then be used to recursively reconstruct the clean speech phase, $\phi_\mathrm{S} = \angle S$, along time:

$$\widehat{\phi_\mathrm{S}}(k,\ell) = \widehat{\phi_\mathrm{S}}(k,\ell-1) + \Omega_h^k L, \tag{3}$$

where Ω_h^k is the angular frequency of the harmonic component dominant in band k. Here, in contrast to [21], a transformation of the STFT bands into the respective baseband is omitted to simplify the formulas.

In general, if the fundamental frequency is known, (3) allows for a reconstruction of the STFT-phase during voiced speech. However, initialization at the beginning of a voiced segment remains an issue. For bands directly containing a harmonic component the noisy phase yields a decent initialization, since the local SNR in those bands is likely to be high. In between these bands, the signal energy is typically very low, so the phase is heavily disturbed by the noise. Thus, simply initializing (3) with the noisy phase in all bands might lead to inter-band inconsistencies of the phase. Therefore, the phase is initialized by the noisy phase and reconstructed along time only in bands containing harmonics. Based on these phase

estimates, the remaining bands are then reconstructed across frequency in every segment separately via

$$\widehat{\phi_{\mathrm{S}}}(k+i) = \widehat{\phi_{\mathrm{S}}}(k) - \phi_{\mathrm{W}}\left(k - \frac{\Omega_h^k N}{2\pi}\right) + \phi_{\mathrm{W}}\left(k+i-\frac{\Omega_h^k N}{2\pi}\right), \tag{4}$$

where we neglect the segment index ℓ. With phase $\widehat{\phi_{\mathrm{S}}}(k,\ell)$ obtained along time via (3), the phase in neighboring bands $k+i$, with integer $i \in \left\{ \lceil -\frac{f_0/2}{f_{\mathrm{s}}} N \rceil, \ldots, \lceil \frac{f_0/2}{f_{\mathrm{s}}} N \rceil \right\}$ and $\lceil \cdot \rceil$ rounding up to the next largest integer, is gained by accounting for the phase shift introduced by the analysis window, i.e. ϕ_{W}. Note that $\Omega_h^k \frac{N}{2\pi}$ is a real-valued non-integer number between 0 and N. With (3), (4), and an estimate of the fundamental frequency at hand, it is now possible to blindly reconstruct the clean speech phase during voiced speech. In non-voiced segments however, the noisy phase is not modified.

We now exchange the noisy phase by the reconstructed one and synthesize the resulting time domain signal. This signal is then reanalyzed and presented on the right of Fig. 2. We see that the structures of the clean speech phase are well reconstructed (bottom right of Fig. 2). It is interesting to note that enhancing the spectral phase also results in an enhanced spectral amplitude after reanalysis (top right of Fig. 2).

Besides the stand-alone performance of this algorithm, which was evaluated in [21], it can easily be combined with any state-of-the-art amplitude estimation scheme. In the paper at hand, phase and amplitude enhancement are performed independently and the results, $\widehat{|S|}$ and $\widehat{\phi_{\mathrm{S}}}$, are combined prior to synthesis of the enhanced time domain signal via

$$\widehat{S} = \widehat{|S|}\exp\left(\mathrm{j}\widehat{\phi_{\mathrm{S}}}\right). \tag{5}$$

In the next section, we investigate if phase enhancement can improve existing speech enhancement algorithms further.

V. Evalutation

For the evaluation of combined phase and amplitude enhancement, a randomly chosen subset of the TIMIT database is deteriorated by additive babble noise at global SNRs ranging from -5 dB to 15 dB in steps of 5 dB. A segment length of 32 ms and a segment shift of 4 ms is used, at a sampling frequency of 8 kHz. The unbiased MMSE-based noise power estimator proposed in [22] is employed together with the decision-directed approach for the estimation of the a priori SNR [5].

For the estimation of the fundamental frequency, which yields the basis for the phase reconstruction, YIN [23] is used. Compared to [23], the segment shift is adjusted to 4 ms and the threshold for minimum selection is increased to 0.2, which leads to a slightly higher detection rate in low SNR conditions.

Now, we combine the proposed phase estimation scheme with the log-spectral amplitude (LSA) estimator from [6]. For the evaluation, we employ PESQ, as implemented in [24]. The results for babble noise are presented in Fig. 3, where the curve for the noisy input signal is given as a reference. Since the clean phase is reconstructed only in voiced signal segments,

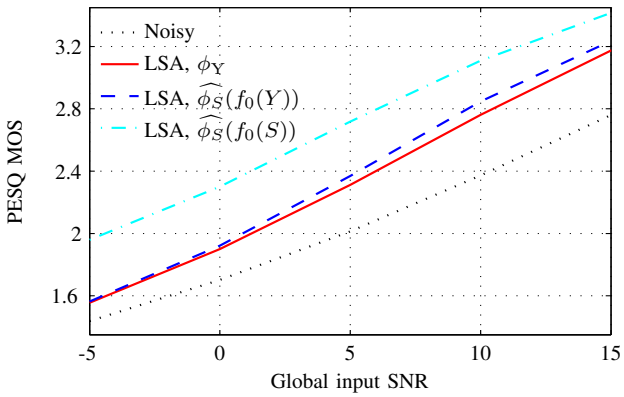

Fig. 3. PESQ MOS in babble noise during voiced speech for the noisy input signal together with signals enhanced via combinations of LSA and different STFT-phases: noisy ϕ_{Y}, blindly estimated $\widehat{\phi_S}(f_0(Y))$, and $\widehat{\phi_S}(f_0(S))$, estimated based on an f_0 estimate on the clean signal.

differences of the combined enhancement scheme and the well-known amplitude enhancement can only be observed in voiced regions. Thus, PESQ is only computed on voiced segments as detected by YIN on the clean speech signal.

It can be seen that the blind phase enhancement consistently improves the amplitude enhancement scheme, with improvements of up to 0.1 PESQ MOS.

Besides the blind phase estimation, we also present the results for the case that the fundamental frequency is estimated not on the noisy, but on the clean speech, to investigate the importance of a noise-robust fundamental frequency estimation. As expected, the 'clean' pitch estimate results in an improved phase reconstruction, especially for low SNR situations, where the detection rate of YIN is strongly reduced by the noise. The maximum improvement over the LSA alone is about 0.4 PESQ MOS.

VI. Conclusions

In single channel speech enhancement it is commonly believed that the spectral phase is *unimportant*, and that the noisy phase is the best estimate of the clean speech phase available. In contrast to this, in [21] we have shown that a blind estimation of the spectral phase is *possible* and increases the frequency weighted SNR of noisy speech by up to 1.8 dB. In this contribution we show that phase estimation can push the limits of single channel speech enhancement further and results in even higher PESQ scores than amplitude estimation alone. At the same time, the full potential of employing an improved phase is not utilized yet. Thus, we believe that research on phase improvement can take an *important* role in speech enhancement.

References

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

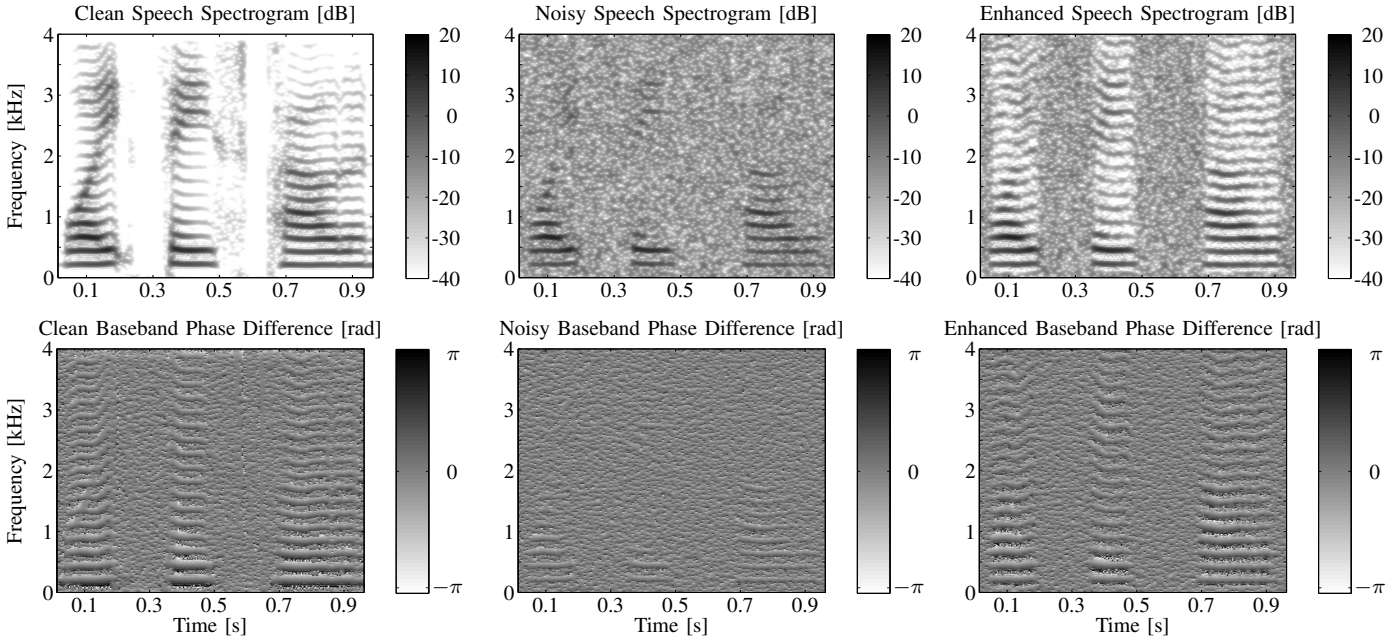

Fig. 2. Amplitude spectra of clean (left), noisy (middle), and enhanced (right) speech signals are presented in the upper row, together with the corresponding baseband phase difference from segment to segment, $\phi(k,l) - \phi(k,l-1)$, in the lower row. The speech signal is degraded by white noise at a global SNR of 0 dB. Note that the noise reduction between the harmonics – visible on the top right – is achieved by phase enhancement alone, no amplitude enhancement scheme is applied.

[2] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. of the IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.

[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, 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] 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.

[6] ——, "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.

[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] R. Martin, "Speech enhancement using MMSE short time spectral estimation with Gamma distributed speech priors," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2002, pp. 253–256.

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

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

[11] 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, Language Process.*, vol. 15, no. 6, pp. 1741–1752, Aug. 2007.

[12] 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.

[13] 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. (ICASSP)*, Apr. 2008, pp. 4037–4040.

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

[15] P. Hannon and M. Krini, "Dynamic spectro-temporal features for excitation signal quantization in a model-based speech reconstruction system," Kiel, Germany, Sep. 2011.

[16] T. F. Quatieri, "Phase estimation with application to speech analysis-synthesis," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA 02139, 1979.

[17] M. H. Hayes, J. S. Lim, and A. V. Oppenheim, "Signal reconstruction from phase or magnitude," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 6, pp. 672–680, Dec. 1980.

[18] C. Y. Espy and J. S. Lim, "Effects of additive noise on signal reconstruction from Fourier transform phase," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 31, no. 4, pp. 894–898, Aug. 1983.

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

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

[21] M. Krawczyk and T. Gerkmann, "STFT phase improvement for single channel speech enhancement," *Int. Workshop Acoustic Echo, Noise Control (IWAENC)*, Sep. 2012.

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

[23] A. d. Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Amer.*, vol. 111, no. 4, pp. 1917–1930, Apr. 2002.

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