PHASE-SENSITIVE REAL-TIME CAPABLE SPEECH ENHANCEMENT UNDER VOICED-UNVOICED UNCERTAINTY

Martin Krawczyk, Robert Rehr, and Timo Gerkmann

Speech Signal Processing Group, Department of Medical Physics and Acoustics
University of Oldenburg, Germany
{martin.krawczyk, r.rehr, timo.gerkmann}@uni-oldenburg.de

ABSTRACT

In many short-time Fourier transform (STFT)-based single channel speech enhancement algorithms, the clean speech spectral amplitude is estimated from a noisy observation to suppress additive noise. For the estimation, only the noisy amplitudes and functions thereof, like the *a priori* or *a posteriori* signal-to-noise ratio (SNR), are utilized. Information about the clean speech spectral phase is mostly not employed. In this work we present a comprehensive speech enhancement setup that combines phase-sensitive and phase-insensitive amplitude estimation, improving the perceptual speech quality of the enhanced signal in terms of PESQ compared to phase-insensitive amplitude estimation alone. The proposed algorithm is real-time capable in the sense that it is implemented in a causal block-wise manner and the computational complexity is feasible.

***Index Terms*—** speech enhancement, noise reduction, phase estimation, amplitude estimation

1. INTRODUCTION

In STFT based single channel speech enhancement, a speech signal $s(n)$ corrupted by additive noise $v(n)$ is observed with only one microphone. The noisy observation $y(n) = s(n)+v(n)$ is chopped into segments of length N, with a segment shift of L samples, and multiplied by a window function $w(n)$. The spectral representation of the noisy observation is obtained by applying a discrete Fourier transform (DFT), giving

$$Y(k,\ell) = |Y(k,\ell)|\,\mathrm{e}^{\mathrm{j}\phi_{\mathrm{Y}}(k,\ell)} = S(k,\ell) + V(k,\ell)\,. \tag{1}$$

For brevity, we neglect the frequency index k and the segment index ℓ in the sequel whenever possible. The complex spectral coefficients of the clean speech, the noise, and the noisy mixture are denoted by $S = |S|\mathrm{e}^{\mathrm{j}\phi_{\mathrm{S}}}$, V, and Y, respectively. In speech enhancement, typically only the spectral amplitudes $|Y|$ are modified, while the noisy phase ϕ_{Y} is kept unchanged, e.g. [1, 2]. This is often motivated by the assumption that the enhancement of the spectral amplitude is perceptually more important than the enhancement of the spectral phase [3]. Further, in [1] it has been shown that the phase of the noisy signal in the STFT-domain ϕ_{Y} is the minimum mean-square error (MMSE) optimal estimator of the clean speech phase ϕ_{S} given that ϕ_{S} is uniformly distributed between $-\pi$ and π. However, in [4] Paliwal et al. conducted listening tests to investigate the importance of the spectral phase in speech enhancement and came to the conclusion that incorporating information about the phase can potentially improve the perceived signal quality.

One of the most well-known approaches to estimate the spectral phase of a speech signal has been proposed by Griffin and Lim [5]. It exploits correlations between overlapping signal segments to iteratively estimate the spectral phase from the spectral amplitude. Several improvements have been suggested over the years, leading to faster convergence and real-time capability, see e.g. [6] for an overview. However, these methods rely on the availability of the clean speech spectral amplitudes. As a result, the performance is limited when only estimated amplitudes are available [7], and artifacts such as echo, smearing and modulations may occur [8]. However, in the context of single-channel source separation the iterative approaches have been shown to outperform classical Wiener filter approaches [8][7].

Recently, we proposed to estimate the clean STFT-phase both on and between spectral harmonics from an estimate of the fundamental frequency in voiced speech [9]. This estimate was then applied to speech enhancement [10, 11]. The idea of applying fundamental frequency based phase estimates for speech enhancement was also investigated by Mehmetcik and Çiloğlu [12]. However, in contrast to [9], in [12] the noisy phase is only modified on the spectral harmonics, while the noisy phase between the speech spectral harmonics remains unchanged. This limits the achievable improvement in a speech enhancement framework which explains the differences in performance between [12] and [9].

Like many amplitude estimators the log-spectral amplitude estimator (LSA) is derived under the assumption that the

This work was supported in part by the Cluster of Excellence 1077 "Hearing4All", funded by the German Research Foundation (DFG).

clean speech phase ϕ_{S} is unknown and uniformly distributed. However, using [9] we obtain an estimate of ϕ_{S}. As a result, ϕ_{S} can not be assumed uniformly distributed anymore, meaning that the LSA is not the optimal amplitude estimator in this context. Thus, in [11] a novel, phase-sensitive clean speech spectral amplitude estimator has been proposed that incorporates knowledge of ϕ_{S} for an improved estimation of $|S|$. This estimator has been shown to outperform existing approaches in terms of speech quality as predicted by PESQ, for both, oracle phase as well as blind phase estimates obtained via [9].

The technique presented in [9] provides phase estimates only for voiced sounds. For unvoiced sounds, the employed deterministic model does not hold. Accordingly, phase information can only be incorporated during voiced speech, while unvoiced sounds have to be handled differently. In this paper, we present a real-time capable, self contained setup for single channel speech enhancement that combines clean speech phase and amplitude estimation.

We shortly present the clean speech phase estimation in Section 2, followed by a real-time capable version of the noise robust fundamental frequency estimation algorithm PEFAC [13] in Section 3, and the phase-sensitive amplitude estimator [11] in Section 4. In Section 5 we rate the reliability of the phase estimates based on the probability of a segment being voiced and combine phase-sensitive and phase-insensitive speech enhancement accordingly. After presenting the evaluation in Section 6 this paper is concluded in Section 7.

2. PHASE ESTIMATION

In [9] we showed that during voiced speech, one can find characteristic structures within the phase. These structures can be visualized best by demodulating each STFT frequency band into the baseband and taking the phase difference from one segment to the next. An example of such a phase spectrum is presented in [9]. Similar to the speech structures in the spectral amplitudes, the structures in the spectral phase are lost to a great extend when the clean speech signal is degraded by noise. The goal of the phase estimation in [9] is to reconstruct the clean speech phase based only on a noisy observation. For this, a harmonic signal model is employed for voiced speech, consisting of sinusoidal components at the fundamental frequency f_0 and its harmonics $f_h = (h+1)\, f_0$:

$$s(n) \approx \sum_{h=0}^{H-1} 2A_h \cos\left(2\pi f_h \frac{n}{f_{\mathrm{s}}} + \varphi_h\right), \tag{2}$$

with harmonic index h, number of harmonics H, amplitude A_h, time domain phase φ_h, and sampling frequency f_{s}. In contrast to [9], here we formulate the phase estimator without using a baseband transformation, as it simplifies the formulas.

It is assumed that each STFT frequency band k is dominated by the harmonic component closest to that band. This assumption is well fulfilled in case the spectral resolution of the STFT is fine enough to separate the harmonic components. We denote the frequency of the harmonic h that is closest to band k as f_h^k.

For a single frequency component the clean speech spectral phase can be computed analytically based on the phase of the previous segment as

$$\widehat{\phi_{\mathrm{S}}}(k,\ell) = \widehat{\phi_{\mathrm{S}}}(k,\ell{-}1) + 2\pi f_h^k \frac{L}{f_{\mathrm{s}}}, \tag{3}$$

where the *hat* symbol is used to denote estimated values. Thus, starting from an initial phase estimate at a voiced speech onset ℓ_0, we can recursively estimate the phase in consecutive segments given that f_0 is known. For bands directly containing a harmonic component we employ the noisy phase for the initialization of (3).

In bands between the spectral harmonics the noisy phase does not provide a decent initialization, since the speech energy is typically very low with respect to the noise energy. In [9] it has been shown that the spectral phase between the harmonic components is a function of the fundamental frequency and the phase response of the spectral analysis window $w\,(n)$. Based on a phase estimate in frequency band k, obtained with (3), the phases of the surrounding bands $k{+}i$ can be estimated by

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

where we account for the influence of the phase response ϕ_{W} of the spectral analysis window $w(n)$.

With the above method at hand, it is possible to estimate the spectral phase of voiced speech based on the fundamental frequency. In practice, the fundamental frequency is unknown and needs to be estimated, which will be discussed in the next section.

3. FUNDAMENTAL FREQUENCY ESTIMATION

To estimate the fundamental frequency we employ the PEFAC algorithm proposed in [13] and implemented in [14]. There, the complex STFT-coefficients are first squared and then mapped onto a logarithmic frequency axis ranging from 10 Hz to 4 kHz, yielding $R(q) = |Y(q)|^2$, with $q = \log(f)$ denoting the logarithmic frequency in ln(Hz). While on a linear frequency axis the distance between two harmonic components is given by f_0, the distances become independent of f_0 on a logarithmic axis, as

$$\ln(f_h) = \ln(h+1) + \ln(f_0)\,. \tag{5}$$

Thus, the relative distances between the harmonic peaks are the same for all fundamental frequencies and only the absolute position is shifted for different f_0. This allows for an efficient search for the best f_0 candidate, which can be formulated as a correlation with a single function, defined as [13]

$$g\,(q) = \rho - \frac{1}{\gamma - \cos(2\pi e^q)}. \qquad (6)$$

Here, the parameter γ controls the width of the peaks of $g(q)$, while the normalization term ρ is chosen such that the mean of $g(q)$ is zero.

To improve the robustness against narrow-band noises $R(q)$ is modified as [14]

$$R'(q) = R(q)\,\frac{\overline{S}(q)}{\overline{R}(q)}. \qquad (7)$$

For this, R is smoothed across time and frequency to obtain $\overline{R}(q)$, which is then compared to the long term average speech spectrum $\overline{S}(q)$ computed offline. For narrow-band noise components, $\overline{R}(q)$ locally increases with respect to $\overline{S}(q)$. Accordingly, the ratio of the two in (7) reduces, effectively suppressing the noise. An estimate of the fundamental frequency $\hat{f}_0$ is then found via

$$R''(q) = R'(q) * g(-q) \qquad (8)$$

$$\hat{f}_0 = \exp\left(\hat{q}_0\right) = \exp\left(\underset{q}{\operatorname{argmax}}\{R''(q)\}\right), \qquad (9)$$

where the asterisk denotes convolution. To enforce continuity of the fundamental frequency trajectories, we use the tracking algorithm of [14]. However, to keep processing latency low, only information from past segments is utilized, i.e. the tracking does not involve a look-ahead.

3.1. Voiced-Unvoiced Probability Estimation

Let $\mathcal{H}_1$ denote the hypothesis that a segment is voiced and $\mathcal{H}_0$ the hypothesis that it is unvoiced. We are now interested in computing the a posteriori probability that a signal segment is voiced for a given set of features $\boldsymbol{\mathcal{F}}$, i.e. $P\left(\mathcal{H}_1|\boldsymbol{\mathcal{F}}\right)$. With Bayes' theorem, we can obtain the posterior probability from the likelihoods and priors as

$$P\left(\mathcal{H}_1|\boldsymbol{\mathcal{F}}\right) = \frac{P(\mathcal{H}_1)\,p\left(\boldsymbol{\mathcal{F}}|\mathcal{H}_1\right)}{P(\mathcal{H}_1)\,p\left(\boldsymbol{\mathcal{F}}|\mathcal{H}_1\right) + P(\mathcal{H}_0)\,p\left(\boldsymbol{\mathcal{F}}|\mathcal{H}_0\right)}, \qquad (10)$$

where we assume equal priors for voiced and unvoiced speech, $P(\mathcal{H}_1) = P(\mathcal{H}_0) = 0.5$. As in [14] we derive two features from $R'(q)$ and $R''(q)$, while the likelihoods are modeled by a Gaussian Mixture Model (GMM) with six mixtures. Training of the GMMs is performed using utterances from the TIMIT training set degraded by additive noise taken from the NOISEX-92 database. We use babble, white, and car noise at SNRs ranging from -10 db to 20 dB in 5 dB steps.

The resulting probability $P\left(\mathcal{H}_1|\boldsymbol{\mathcal{F}}\right)$ will be used in Section 5 for a soft mix of phase sensitive and insensitive speech enhancement.

With $\hat{f}_0$ and (3), (4) we can now estimate the clean speech phase ϕ_S. In the next section we show how this information can be utilized for an improved spectral amplitude estimation.

4. PHASE-SENSITIVE AMPLITUDE ESTIMATION

In [11] we proposed a novel phase-sensitive MMSE-optimal estimator for the (compressed) clean speech spectral amplitudes, incorporating information about the clean speech phases for an improved clean speech amplitude estimation. In contrast to existing approaches like [1, 15, 2, 16], the clean speech phase ϕ_S is not assumed to be uniformly distributed between $-\pi$ and π. Instead, the novel estimator is derived under the assumption that ϕ_S is deterministic and known. Further, the probability density function (PDF) of the speech spectral amplitudes is modeled to follow a χ-distribution with shape parameter μ [17, 16]. The estimator results in [11]

$$\begin{aligned}\widehat{|S|} &= \left(\mathrm{E}\left(|S|^\beta \mid Y, \phi_\mathrm{S}\right)\right)^{\frac{1}{\beta}} \\ &= \sqrt{\tfrac{1}{2}\tfrac{\xi}{\mu+\xi}\sigma_\mathrm{V}^2}\left(\tfrac{\Gamma(2\mu+\beta)}{\Gamma(2\mu)}\,\tfrac{\mathrm{D}_{-(2\mu+\beta)}(\nu)}{\mathrm{D}_{-(2\mu)}(\nu)}\right)^{\frac{1}{\beta}}, \qquad (11)\end{aligned}$$

with statistical expectation $\mathrm{E}(\cdot)$, the compression parameter β, the gamma function $\Gamma(\cdot)$, the *a priori* SNR ξ, the parabolic cylinder function $\mathrm{D}_\cdot(\nu)$ [18, Eq. (9.24)], and its argument

$$\nu = -\sqrt{2\frac{\xi}{\mu+\xi}\zeta}\;\cos(\underbrace{\phi_\mathrm{Y} - \phi_\mathrm{S}}_{\Delta\phi}). \qquad (12)$$

Note that most spectral amplitude enhancement schemes in the literature only employ the noisy amplitude and measures derived from it, like the *a priori* SNR $\xi = \sigma_\mathrm{S}^2/\sigma_\mathrm{V}^2$ and the *a posteriori* SNR $\zeta = |Y|^2/\sigma_\mathrm{V}^2$ to obtain an estimate of the clean speech amplitude $|S|$. Here, $\sigma_\mathrm{V}^2 = \mathrm{E}\left(|V|^2\right)$ and $\sigma_\mathrm{S}^2 = \mathrm{E}\left(|S|^2\right)$ denote the power spectral densitys (PSDs) of the noise and the speech, respectively. In contrast, the estimator (11) proposed in [11] makes use of the phase difference $\Delta\phi$ as an additional measure to distinguish noise from speech. Large deviations of the noisy phase ϕ_Y from the clean phase ϕ_S are only possible in bins that are dominated by the noise component [19][11]. Accordingly, the larger the phase difference $\Delta\phi$, the more attenuation is applied. On the other hand, time-frequency points where $\Delta\phi$ is small, i.e. $\phi_\mathrm{Y} \approx \phi_\mathrm{S}$, are preserved better as compared to phase-insensitive estimators.

In practical scenarios the clean speech phase is unknown and only the phase of the noisy mixture ϕ_Y can be observed. With the phase estimation technique for voiced speech [9], we now have an estimate of the clean speech phase ϕ_S at hand. This phase estimate can then be used in (12) and (11) for an improved amplitude estimation. As we do not have an estimator for the phase in unvoiced speech, in the following section a combination of phase-sensitive and insensitive speech enhancement techniques is proposed.

5. SPEECH ENHANCEMENT UNDER VOICED-UNVOICED UNCERTAINTY

For the novel phase-sensitive amplitude estimator [11] a clean speech phase estimate is needed in (12). With [9] phase estimates can be obtained only in voiced speech, in unvoiced speech or speech absence we lack an estimate of ϕ_{S}. Accordingly, the phase-sensitive amplitude estimator cannot be applied in these regions. Therefore, outside of voiced regions we do not assume the phase to be deterministic and known, but rather uniformly distributed between $-\pi$ and π. Like for the phase-sensitive estimator [11], a χ-distribution for the speech spectral amplitudes with the same shape parameter μ as well as the same compression β is used. This leads to the phase-insensitive counterpart proposed in [16].

Both, phase-sensitive and insensitive amplitude estimators are run in parallel and are mixed based on the current segment's probability of being voiced $P(\mathcal{H}_1|\boldsymbol{\mathcal{F}})$,

$$\widetilde{|S|} = P(\mathcal{H}_1|\boldsymbol{\mathcal{F}})\,\widehat{|S|}_{[11]} + P(\mathcal{H}_0|\boldsymbol{\mathcal{F}})\,\widehat{|S|}_{[16]}. \qquad (13)$$

For segments that show a high probability of being voiced, $P(\mathcal{H}_1|\boldsymbol{\mathcal{F}}) \approx 1$, the phase estimation scheme [9] is applicable and the clean speech amplitude estimate $\widetilde{|S|}$ is obtained using [11]. The lower $P(\mathcal{H}_1|\boldsymbol{\mathcal{F}})$, the less confident we are in the phase estimates. Accordingly, the phase-insensitive amplitude estimate obtained by [16] dominates the mixture (13) in unvoiced regions.

The evaluation of the proposed phase-sensitive amplitude enhancement under voiced-unvoiced uncertainty is presented in the next section.

6. EVALUATION

For the evaluation of the proposed algorithm, we take 128 speech samples from the TIMIT database at a sampling frequency of 8 kHz, one half being uttered by male, the other half by female speakers. The speech is deteriorated by additive babble noise [20] at SNRs ranging from 0 dB to 20 dB. Neither the speech nor the noise used for evaluation have been used in the training phase of the GMMs in Section 3. The STFT segment length and shift are set to 32 ms and 4 ms, respectively. The fundamental frequency estimation is performed on 90 ms windows, centered around the STFT-segments of the enhancement framework, increasing the overall latency of the algorithm to a minimum of 45 ms. The unbiased MMSE noise power estimator [21] is employed together with the decision-directed approach [15] to estimate the *a priori* SNR. Further, $\beta = \mu = 0.5$ is used in (11) as well as for the phase-insensitive amplitude enhancement [16].

Instead of the absolute phases $\widehat{\phi_{\mathrm{S}}}$ and ϕ_{Y}, we employ the phase changes $\widehat{\Delta\phi_{\mathrm{S}}}(\ell) = \widehat{\phi_{\mathrm{S}}}(\ell) - \widehat{\phi_{\mathrm{S}}}(\ell-1)$ and $\Delta\phi_{\mathrm{Y}} = \phi_{\mathrm{Y}}(\ell) - \phi_{\mathrm{Y}}(\ell-1)$ in (12). This showed to be more robust to phase estimation errors and makes the initialization of (3) at voiced speech onsets unnecessary.

We now compare the novel mixture of phase-sensitive and insensitive speech enhancement to phase-insensitive enhancement alone in terms of speech quality. For this, PESQ is computed as implemented in [14] for the noisy input together with the results of the phase-insensitive enhancement [16] and of the proposed approach. The results are presented in Figure 1. In the upper graph, it can be seen that the proposed approach achieves an improvement of the PESQ mean opinion scores (MOS) with respect to [16] for all input SNRs considered here. For input SNRs of 10 dB or less an improvement of the PESQ-MOS by around 0.1 points is achieved.

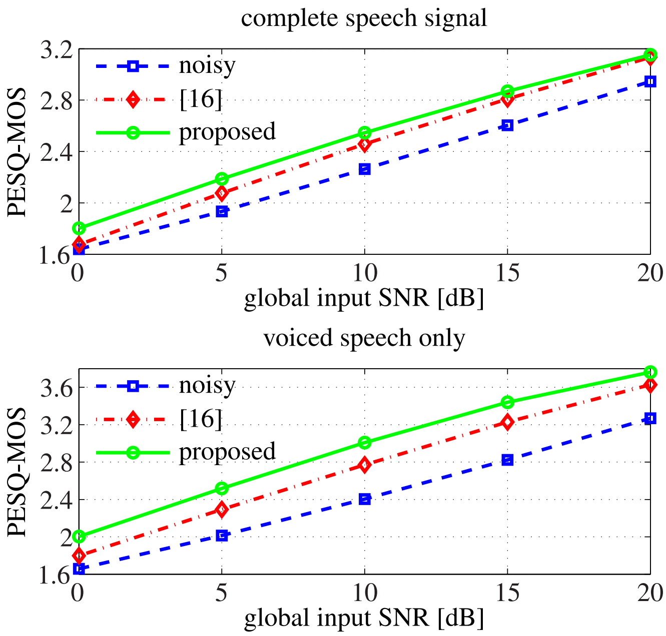

Fig. 1. PESQ-MOS evaluated on complete signals (top) and in voiced speech only (bottom) for the noisy input and the enhanced speech estimated with the proposed approach. As a reference, the phase-insensitive approach [16] is also presented, using the same parameter setting, $\beta = \mu = 0.5$.

Note that as in unvoiced speech no phase estimate is available, the performance of the proposed algorithm is virtually the same as for the phase-insensitive estimator [16]. The benefit of the proposed approach comes in voiced speech, where we have an estimate of the clean speech phase. This is depicted in the lower graph of Figure 1, where PESQ is evaluated in voiced segments only. For SNRs up to 15 dB an additional improvement of more than 0.2 points compared to the phase-insensitive estimator is achieved.

7. CONCLUSION

In this work we have presented a comprehensive speech enhancement setup that employs information about the clean speech spectral phase. In voiced speech the clean speech phase is estimated and utilized for an improved, phase-sensitive clean speech amplitude estimation. In unvoiced sounds, classical phase-insensitive amplitude estimation is used. A novel combination of the two estimators based on the probability of a signal segment being voiced has been presented. The proposed algorithm has been shown to outperform phase-insensitive amplitude enhancement alone in terms of speech quality as predicted by PESQ.

Acknowledgments

The authors would like to thank Sira Gonzalez for providing valuable insights into the fundamental frequency estimator PEFAC.

8. 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] 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.

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

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

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

[6] N. Sturmel and L. Daudet, "Signal reconstruction from STFT magnitude: a state of the art," in *International Conference on Digital Audio Effects (DAFx)*, Paris, France, Sep. 2011, pp. 375–386.

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

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

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

[10] T. Gerkmann, M. Krawczyk, and R. Rehr, "Phase estimation in speech enhancement — unimportant, important, or impossible?" in *IEEE Conv. Elect. Electron. Eng. Israel*, Eilat, Israel, Nov. 2012.

[11] 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.

[12] E. Mehmetcik and T. Çiloğlu, "Speech enhancement by maintaining phase continuity," in *Proc. of Meetings of the Acoustical Society of America*, vol. 18, no. 055002, Nov. 2012.

[13] S. Gonzalez and M. Brookes, "A pitch estimation filter robust to high levels of noise (PEFAC)," in *EURASIP Europ. Signal Process. Conf. (EUSIPCO)*, Barcelona, Spain, Sep. 2011, pp. 451–455.

[14] M. Brookes, "VOICEBOX: a speech processing toolbox for MATLAB." [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[15] 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.

[16] 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)*, Las Vegas, NV, USA, Apr. 2008, pp. 4037–4040.

[17] 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.

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

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

[20] H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses," *EURASIP J. Adv. Sig. Proc.*, 2009.

[21] 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, May 2012.