STFT PHASE IMPROVEMENT FOR SINGLE CHANNEL SPEECH ENHANCEMENT

Martin Krawczyk, Timo Gerkmann

Speech Signal Processing Group, Institute of Physics, University of Oldenburg, Germany
{martin.krawczyk, timo.gerkmann}@uni-oldenburg.de

ABSTRACT

In state-of-the-art single channel short-time Fourier transform (STFT) based speech enhancement algorithms only the amplitude of the noisy speech signal is improved, but its phase is left unchanged. It is commonly assumed that the noisy phase is the best estimate of the clean phase available. While using the noisy phase is indeed optimal under certain statistical assumptions, in this paper we show that blindly improving the noisy phase is possible when these, potentially limiting, assumptions are dropped. Without modifying the amplitude, the proposed algorithm leads to frequency weighted SNR improvements of up to 1.8 dB. Further, the presented phase enhancement scheme is real-time capable and can be combined with any off-the-shelf STFT-based amplitude estimator.

Index Terms— speech enhancement, phase estimation, noise reduction, signal reconstruction

1. INTRODUCTION

Algorithms for the enhancement of single-channel noisy speech are commonly defined in the short-time Fourier transform (STFT) domain due to, among other reasons, the fact that it allows for perfect reconstruction and that computationally efficient implementations exist. During the past 30 years, huge efforts have been expended in deriving and developing effective STFT-based speech enhancement algorithms, with quite some success. A well-known example of such algorithms is the amplitude estimator presented by Ephraim et al. [1]. Like many popular approaches, the algorithm aims at enhancing only the amplitude of the complex spectrum of a speech signal degraded by additive noise. The phase - which clearly is degraded as well - is, however, not modified. This is often motivated by the work of Vary [2] as well as Wang and Lim [3], where it is stated that the possible gain achieved by STFT-phase enhancement is small compared to the one possible through amplitude enhancement. In contrast to [3], Paliwal et al. [4] have shown that using the clean phase can indeed result in an increased performance of single channel speech enhancement algorithms. Further, they propose to use different spectral analysis windows for computing amplitude and phase, respectively [4]. While it is interesting to see that this results in a reduction of noise, with these modifications the perfect reconstruction property of the STFT is lost, meaning that their approaches necessarily also result in speech and signal distortions.

In this work, given noisy speech, we blindly reconstruct the phase of voiced clean speech directly in the STFT-domain. 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 [5]. Under these assumptions, the MMSE-optimal estimate for the clean phase is known to be the noisy phase [1] — as long as no further information is utilized. However, in this work we point out that in voiced sounds neighboring phase values are in fact highly correlated and also that phase trajectories are highly correlated with spectral amplitudes. We present an algorithm that is capable of exploiting these correlations to blindly reconstruct the clean phase and show that this results in an improved speech enhancement performance. From these results we conclude that using the noisy phase is only optimal under the limiting assumptions of independence and a uniform phase distribution. By dropping these assumptions we can indeed improve speech enhancement algorithms further. The proposed phase estimation scheme is designed in the STFT domain, such that it can be easily combined with state-of-the-art spectral amplitude estimators.

2. SIGNAL MODEL

At each sample n the noisy observation $y(n)$ is given by an additive superposition of speech $s(n)$ and noise $v(n)$, i.e. $y(n) = s(n) + v(n)$. The noisy signal is split into segments of length N, overlapping by $N-L$ samples. To each segment, the window $w(n)$ is applied prior to a Discrete Fourier Transform (DFT), yielding its STFT representation

$$\begin{aligned} Y(k,l) &= \sum_{n=0}^{N-1} y(lL+n)w(n)\mathrm{e}^{-\mathrm{j}\Omega_k n} \\ &= S(k,l) + V(k,l) \\ &= |Y(k,l)|\mathrm{e}^{\mathrm{j}\phi_Y(k,l)}, \end{aligned} \tag{1}$$

with frequency index k, segment index l, and normalized angular frequency $\Omega_k = 2\pi k/N$, $k = 0, 1, ..., N-1$. Note that (1) is the standard definition of the STFT-representation, which typically yields the basis to spectral amplitude enhancement algorithms. Any off-the-shelf amplitude enhancement algorithm can be applied to the noisy amplitude $|Y(k,l)|$, yielding an estimate for the clean speech amplitude, $|\hat{S}(k,l)|$, which is finally combined with the noisy phase for signal reconstruction. In this work, we put amplitude estimation aside and show that a significant noise reduction can be obtained only by improving the phase in a controlled way. This algorithm can then be combined with any amplitude estimator.

First, let us recall that the STFT representation $Y(k,l)$ in (1) can be interpreted as the output of a band-pass filter bank with N bands, where the window function $w(n)$ defines the prototype low-pass [2]. We now transform each band k of the noisy STFT $Y(k,l)$ into its baseband. Multiplication of both sides of (1) yields

$$\begin{aligned} Y_{\mathrm{B}}(k,l) &= Y(k,l)\mathrm{e}^{-\mathrm{j}\Omega_k lL} \\ &= \sum_{n=0}^{N-1} w(n)y(n+lL)\mathrm{e}^{-\mathrm{j}\Omega_k(n+lL)} \end{aligned} \tag{2}$$

where $Y_{\mathrm{B}}(k,l)$ can now be thought of as the output of a complex demodulator followed by a low-pass filter defined by window $w(n)$, subsampled by a factor L. In this context, B is used to distinguish the "baseband" representation from its STFT counterpart. It can be shown that some clear structure is inherited in the baseband phase of clean speech, $\angle Y_{\mathrm{B}}(k,l) = \phi_{Y_{\mathrm{B}}}(k,l)$. This is depicted in the lower left of Fig. 3, where not the phase directly, but the phase difference from one segment to the next, $\phi_{Y_{\mathrm{B}}}(k,l) - \phi_{Y_{\mathrm{B}}}(k,l-1)$, is presented for a clean speech example, i.e. $v(n) = 0$, $\forall n$. The corresponding amplitude spectrogram is presented in the upper left panel of Fig. 3. Indeed, the structures in the amplitude and phase spectra are quite related and also appear to be correlated along time, especially during voiced speech segments. A significant portion of these structures is lost, both, in the amplitude and the phase spectra if noise is added to the speech signal, exemplarily shown for white noise and a global SNR of 0 dB in the middle section of Fig. 3. We now aim at reconstructing the structures of the phase spectrum from the noisy signal during voiced speech segments.

A voiced speech sound can be modeled as a weighted superposition of H sinusoids, leading to the harmonic signal model

$$\tilde{s}(n) = \sum_{h=0}^{H-1} 2A_h \cos(\Omega_h n + \varphi_h), \tag{3}$$

with the real-valued amplitude $2A_h$, time-domain phase φ_h, and normalized angular frequency

$$\Omega_h = 2\pi f_h / f_{\mathrm{S}} = 2\pi(h+1) f_0 / f_{\mathrm{S}}. \tag{4}$$

Here, f_{S}, f_0, and f_h denote the sampling, fundamental and harmonic frequency, respectively.

With (2) and (3) we obtain the baseband STFT representation of the voiced speech model $\tilde{s}(n)$:

$$\begin{aligned}
&\tilde{S}_{\mathrm{B}}(k,l) \\
&\quad = \sum_{n=0}^{N-1} w(n) \left(\tilde{s}(n+lL) \mathrm{e}^{-\mathrm{j}\Omega_k (n+lL)} \right) \\
&\quad = \sum_{n=0}^{N-1} w(n) \sum_{h=0}^{H-1} A_h \big(\mathrm{e}^{\mathrm{j}((\Omega_h - \Omega_k)(n+lL) + \varphi_h)} \\
&\qquad\qquad + \mathrm{e}^{-\mathrm{j}((\Omega_h + \Omega_k)(n+lL) + \varphi_h)} \big).
\end{aligned} \tag{5}$$

In case the segment length N is chosen large enough and the low-pass defined by $w(n)$ is sufficiently narrow and steep to effectively separate the spectral harmonics, (5) can be simplified by assuming that each STFT bin k is dominated only by the closest complex exponential, denoted by

$$\Omega_h^k = \underset{\Omega_h}{\operatorname{argmin}}(|\Omega_k - \Omega_h|). \tag{6}$$

Please note that $\Omega_k N/(2\pi) = k$ is an integer, while $\Omega_h^k N/(2\pi) \in \mathbb{R}$, i.e. the harmonic frequency is not necessarily identical to one of the center frequencies of the DFT.

Equation (5) now reduces to

$$\tilde{S}_{\mathrm{B}}(k,l) \approx A_h \sum_{n=0}^{N-1} w(n) \mathrm{e}^{\mathrm{j}\left((\Omega_h^k - \Omega_k)(n+lL) + \varphi_h \right)}. \tag{7}$$

This simplification is symbolically depicted in Fig. 1 for $H = 2$ harmonics and band k, where in this example band k is the STFT-band with the center frequency closest to $\Omega_{h=0}$. All but the harmonic closest to Ω_k are canceled out by the low-pass filter $W(k)$ introduced by the analysis window $w(n)$.

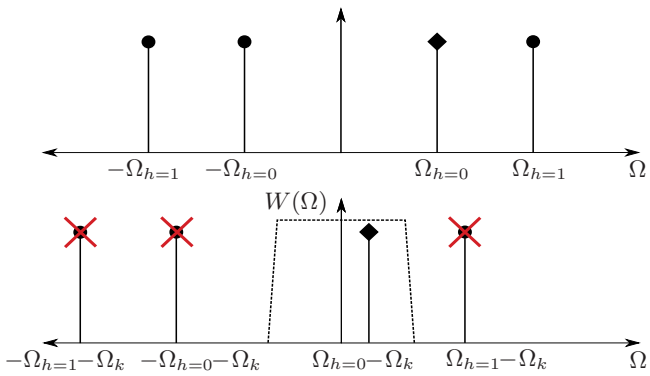

Fig. 1. Symbolic spectrum of a harmonic signal according to (3) (top) and its baseband version for band k (bottom), with $H=2$ harmonics. The low-pass filter $W(k)$, introduced by the time-domain window $w(n)$, effectively suppresses all components but the one closest to the frequency bin of interest, k.

3. PHASE RECONSTRUCTION ALONG TIME

With the simplifications made above it is possible to derive a formula for recursive segment-to-segment computation of the baseband STFT-phase, $\phi_{\tilde{S}_{\mathrm{B}}}(k,l)$, of the harmonic signal model introduced in (3). To this end we reformulate (7), giving

$$\begin{aligned}
\tilde{S}_{\mathrm{B}}(k,l) &\approx \mathrm{e}^{\mathrm{j}(\Omega_h^k - \Omega_k) lL} \mathrm{e}^{\mathrm{j}\varphi_h} A_h \sum_{n=0}^{N-1} \left(w(n) \mathrm{e}^{\mathrm{j}\Omega_h^k n} \right) \mathrm{e}^{-\mathrm{j}\Omega_k n} \\
&= \mathrm{e}^{\mathrm{j}(\Omega_h^k - \Omega_k) lL} \mathrm{e}^{\mathrm{j}\varphi_h} A_h W(k - \Omega_h^k \frac{N}{2\pi}) \\
&= |\tilde{S}_{\mathrm{B}}(k,l)| \mathrm{e}^{\mathrm{j}\phi_{\tilde{S}_{\mathrm{B}}}(k,l)},
\end{aligned} \tag{8}$$

where $W(k)$ is the spectral representation of $w(n)$. Note that $\mathrm{e}^{\mathrm{j}(\Omega_h^k - \Omega_k) lL}$ is the only part of (8) that depends on the segment index l. Therefore, given the harmonic frequency Ω_h^k, we can compute the phase shift from segment to segment analytically as

$$\begin{aligned}
\Delta\phi_{\tilde{S}_{\mathrm{B}}}(k,l) &= \phi_{\tilde{S}_{\mathrm{B}}}(k,l) - \phi_{\tilde{S}_{\mathrm{B}}}(k,l-1) \\
&= \left(\Omega_h^k - \Omega_k \right) L.
\end{aligned} \tag{9}$$

Straightforward reformulation of the above equation leads to the recursive formula

$$\phi_{\tilde{S}_{\mathrm{B}}}(k,l) = \phi_{\tilde{S}_{\mathrm{B}}}(k,l-1) + \left(\Omega_h^k - \Omega_k \right) L, \tag{10}$$

stating that the change of the baseband phase from one segment to the next depends only on the difference between the normalized frequency of the closest harmonic Ω_h^k and STFT center-frequency Ω_k in combination with the segment shift L.

Now we transfer the statements made above to the enhancement of noisy speech. As stated in the previous section, voiced speech can be modeled via the harmonic model given in (3). If an initial estimate of the clean speech baseband phase is available at the beginning of a voiced sound, i.e. at segment $l = l_0$, then $\phi_{\tilde{S}_{\mathrm{B}}}(k, l_0 + 1)$ can be computed based on $\phi_{\tilde{S}_{\mathrm{B}}}(k, l_0)$ and the current harmonic frequency Ω_h via (10) and (6). However, in practice the clean phase will always be disturbed by noise.

The harmonic frequencies Ω_h are directly related to the fundamental frequency f_0 as given in (4). For the estimation of the fundamental frequency some robust algorithms are available, e.g. [6].

Unfortunately, this is not the case for the problem of estimating the initial clean phase at the beginning of a voiced sound. However, at harmonic frequencies the energy of the speech signal exhibits local maxima. Thus, in the corresponding STFT bands, denoted by k', the instantaneous SNR is likely to show local maxima as well. Therefore, in bands directly containing a harmonic speech component, the noisy phase is considered to be a decent estimate of the clean speech phase, $\phi_{\tilde{S}_\mathrm{B}}(k', l_0) \approx \phi_{Y_\mathrm{B}}(k', l_0)$, see also [2]. Starting from $\phi_{\tilde{S}_\mathrm{B}}(k', l_0)$, the clean phase is then reconstructed via (10) from segment to segment.

In STFT bands in between the harmonics the SNR is typically much lower, and estimation of the initial clean speech phase is hardly possible. Hence, instead of trying to reconstruct the phase in these bands along time, the phase is estimated along frequency for every segment separately, based on the bands k'. This method is presented in the following section.

4. PHASE RECONSTRUCTION ALONG FREQUENCY

Besides the phase reconstruction along time, which makes use of segment-to-segment correlation, we now propose a technique to enhance the phase for each segment separately along frequency bin k. Again, the algorithm is defined in the baseband STFT domain and uses the harmonic speech model in (3) as well as the simplifications in (7). In Section 3 we have obtained the clean phase of the STFT-bands k' dominated by spectral harmonics. With this clean speech phase estimate at hand, we estimate the phases in neighboring bands, for which the local SNR is probably poor compared to the reference. First, we rewrite the second equation of (8):

$$\tilde{S}_\mathrm{B}(k', l) \approx A_h \mathrm{e}^{\mathrm{j}(\Omega_h^{k'} lL + \varphi_h)} \underbrace{\mathrm{e}^{-\mathrm{j}\Omega_{k'} lL} W(k' - \Omega_h^k \frac{N}{2\pi})}_{f(k')}, \quad (11)$$

where only $f(k')$ depends on the frequency bin k'. The equation above not only holds for band k', but for all STFT-bands for which $\Omega_h^{k'}$ is the closest and hence dominant harmonic, i.e. bands $k' + i$, with $i \in [\lceil -\frac{f_0/2}{f_\mathrm{S}} N \rceil, \ldots, \lceil \frac{f_0/2}{f_\mathrm{S}} N \rceil]$ and $\lceil \cdot \rceil$ rounds up to the next largest integer. Thus, we can estimate the phase $\phi_{\tilde{S}_\mathrm{B}}(k'+i, l)$ based on the reference $\phi_{\tilde{S}_\mathrm{B}}(k', l)$ via

$$\begin{aligned} &\phi_{\tilde{S}_\mathrm{B}}(k' + i, l) \\ &= \phi_{\tilde{S}_\mathrm{B}}(k', l) - i\frac{2\pi}{N} lL + \phi_W(k' + i - \frac{\Omega_h^k N}{2\pi}) - \phi_W(k' - \frac{\Omega_h^k N}{2\pi}). \end{aligned} \quad (12)$$

Note that $\phi_W(\cdot)$ is the phase of the spectral representation of the analysis window, which can be computed offline for any analysis window $w(n)$. In combination with the segment-to-segment phase reconstruction for the bands k' presented in the previous section, it is now possible to estimate the clean speech baseband-phase in every time-frequency point of a voiced speech signal. The reconstructed phase is combined with the noisy amplitude and then demodulated by multiplying by $\mathrm{e}^{\mathrm{j}\Omega_k lL}$, which yields the clean speech estimate

$$\hat{S}(k, l) = \left(|Y_\mathrm{B}(k, l)| \mathrm{e}^{\mathrm{j}\phi_{\tilde{S}_\mathrm{B}}(k,l)} \right) \mathrm{e}^{\mathrm{j}\Omega_k lL}. \quad (13)$$

In unvoiced segments, however, the noisy phase is employed directly. Note that in (13) it is easily possible to incorporate any amplitude estimator for the enhancement of the noisy amplitude, $|Y_\mathrm{B}(k, l)|$. Finally, the enhanced time domain signal is obtained by applying an inverse DFT followed by an overlap-add procedure. The performance of the proposed algorithm is presented and discussed in the following.

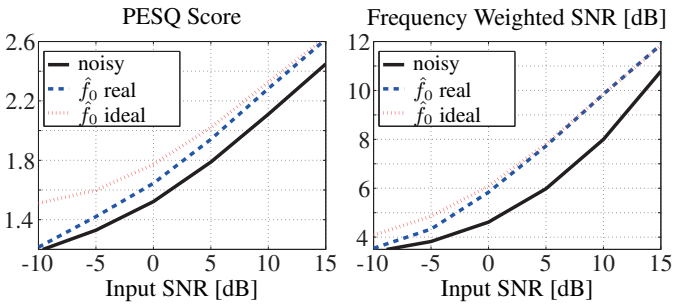

Fig. 2. PESQ Score and frequency weighted SNR, respectively, as a function of the input SNR for white Gaussian noise.

5. EVALUATION

For the evaluation of the proposed algorithm, a randomly chosen subset of 10 female and 10 male speakers taken from the TIMIT database is deteriorated by additive white Gaussian noise with global SNRs ranging from -10 dB to 15 dB in steps of 5 dB. The segment length is set to 32 ms and a segment shift of 4 ms is chosen to allow for a high temporal resolution. With a sampling frequency of 8 kHz, this corresponds to $N = 256$ samples and $L = 32$ samples. PESQ and frequency weighted SNR are employed as objective measures for speech quality and presented in Fig. 2. Implementations of these measures are taken from [7]. The fundamental frequency is estimated on the noisy signal using YIN implemented according to [6], but with the threshold for minimum selection set to 0.5 and a segment advance of 4 ms. For an analysis of the upper bound, we also present the results when the fundamental period is estimated on clean speech.

In Fig. 2 it can be seen that the proposed algorithm achieves improvements of the PESQ score as well as of the frequency weighted SNR for the whole range of evaluated input SNRs. Towards even higher SNRs, the possible enhancement reduces due to the fact that the noisy phase for higher SNRs is indeed a decent estimator for the clean phase. This is stated also in [2], where it is found that above a certain SNR, phase errors are not perceived by human listeners. First studies showed that a comparable performance can be achieved also for more realistic noise types like babble noise.

Instrumental measures are computed on the entire speech signals, while phase enhancement is only performed during voiced speech as detected by YIN. Hence, the results depend on the relation of voiced speech to the total signal length as well as on the percentage of voiced speech that is detected by YIN. For low SNR conditions, YIN detects only few voiced sounds due to the strong influence of the background noise on the estimation of f_0. Although the performance within the detected segments is still good, the overall performance gain reduces. Towards higher SNRs, the influence of the noise to the pitch estimation reduces and more voiced sounds are detected. This can be seen in Fig. 2, where for high input SNRs the difference between the ideal f_0-estimation, based on the clean speech, and the one based on the noisy signal becomes minor. For lower SNRs, however, the pitch estimation deteriorates. Hence, a more sophisticated pitch estimation algorithm might help to improve the performance, especially for low input SNRs.

On the right hand side of Fig. 3 the amplitude spectrogram and

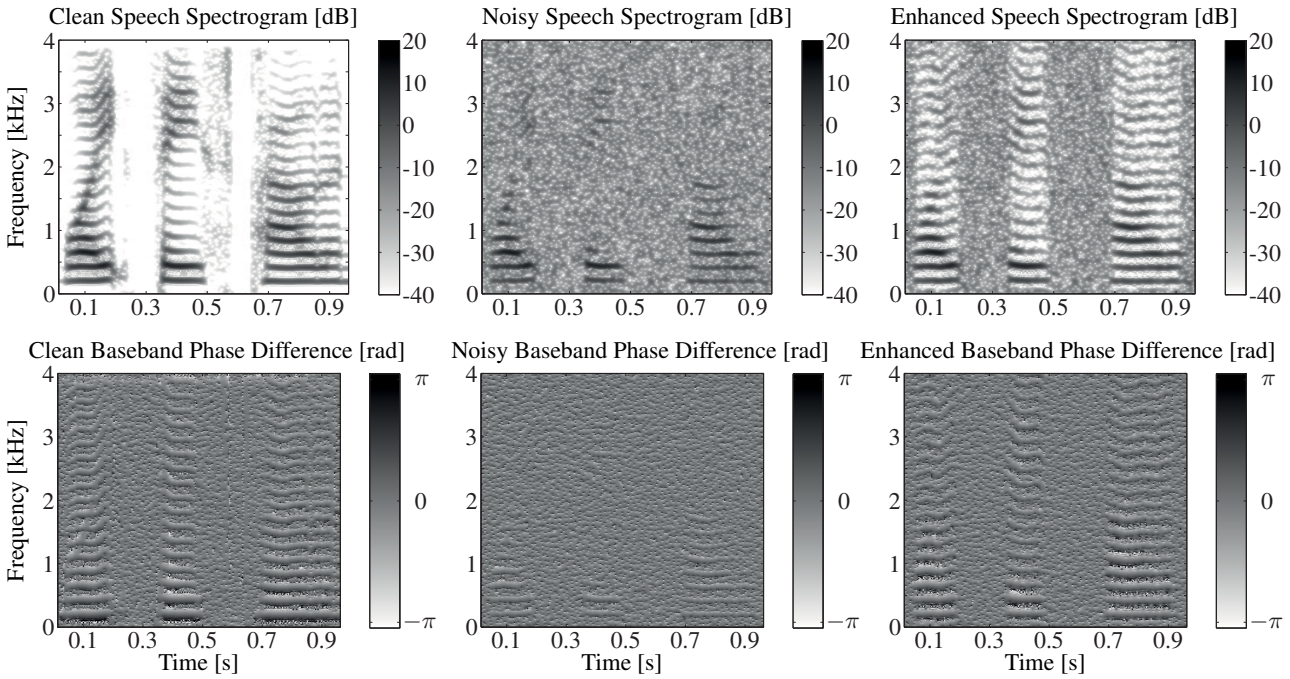

Fig. 3. Amplitude spectra of clean (left), noisy (middle), and enhanced (right) speech signals are presented in the upper row, together with the corresponding baseband phase difference from segment to segment, $\phi(k,l) - \phi(k,l-1)$, in the lower row. The speech signal is degraded by white noise with 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.

the baseband phase difference from one segment to the next is depicted for the signal enhanced by the proposed algorithm. It is compared to the noisy input signal in the middle, degraded by white noise at a global SNR of 0dB, together with the clean speech signal on the left. Pitch estimation as well as voiced/unvoiced classification is performed on the noisy input. It can be observed that, during voiced sounds, structures in the phase that were lost due to noise are successfully reconstructed. More importantly, after reconstructing the enhanced signal via overlap-add and recomputing the STFT-spectra for visualization, the amplitude appears enhanced as well. Not only is the noise reduced in between the harmonics, but even harmonic components in very low SNR regions are reconstructed. Note, again, that this noise reduction after reconstruction is only achieved based on phase improvement.

Informal listening tests further confirm the presented results, where speech enhancement is perceived in terms of noise reduction. Especially for vowels of longer duration the 'roughness' introduced by the noisy phase [2] is reduced. However, in the current version also some artifacts are introduced: firstly, at changes from voiced, processed to unprocessed regions the noise floor changes abruptly. Secondly, the employed harmonic model (3) does not hold perfectly for all voiced speech sounds, like sounds with mixed excitation. Future work will aim at reducing these artifacts and increasing the naturalness of the enhanced signal.

6. CONCLUSION

In this paper we propose an algorithm for the enhancement of noisy speech, based on STFT-phase reconstruction during voiced segments. It is shown that with phase enhancement alone, an improvement of signal quality can be achieved. Besides the stand-alone performance of this method, a combination with STFT-amplitude enhancement algorithms seems promising and will be subject of future work.

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

[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] 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.

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

[6] 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.

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