Least Squares Estimate of the Initial Phases in STFT based Speech Enhancement

Sidsel Marie Nørholm[1], *Martin Krawczyk-Becker*[2], *Timo Gerkmann*[2],
Steven van de Par[3], *Jesper Rindom Jensen*[1] *and Mads Græsbøll Christensen*[1]

[1]Audio Analysis Lab, AD:MT, Aalborg University, Denmark
[2]Speech Signal Processing Group and [3]Acoustics Group,
Cluster of Excellence ”Hearing4all”,
Dept. Medical Physics and Acoustics, University of Oldenburg, Germany
{smn, jrj, mgc}@create.aau.dk
{martin.krawczyk-becker, timo.gerkmann, steven.van.de.par}@uni-oldenburg.de

Abstract

In this paper, we consider single-channel speech enhancement in the short time Fourier transform (STFT) domain. We suggest to improve an STFT phase estimate by estimating the initial phases. The method is based on the harmonic model and a model for the phase evolution over time. The initial phases are estimated by setting up a least squares problem between the noisy phase and the model for phase evolution. Simulations on synthetic and speech signals show a decreased error on the phase when an estimate of the initial phase is included compared to using the noisy phase as an initialisation. The error on the phase is decreased at input SNRs from -10 to 10 dB. Reconstructing the signal using the clean amplitude, the mean squared error is decreased and the PESQ score is increased.

Index Terms: speech enhancement, single-channel, STFT domain, phase estimation, signal reconstruction

1. Introduction

Single-channel speech enhancement is important in many systems such as mobile phones and hearing aids where it is desirable to estimate a speech signal from a mixture of the signal buried in noise. Some enhancement methods work directly in the time domain [1, 2] whereas other methods work by transforming the signal into another domain. This could for example be the subspace methods where, e.g., the eigenvalue decomposition of a signal matrix is computed [3]. Another domain, that we will focus on in this paper because it is computational effective [4], is the short time Fourier transform (STFT) domain. Here, some well-known methods are spectral subtraction [5] and the Short-Time Spectral Amplitude Estimator [6]. Common for these methods, and most other methods in this domain, is that they enhance the STFT amplitude, whereas the phase is left unaltered. This is motivated by [7, 8] who conclude that modifying the noisy STFT phase only gives a minor gain compared to modifying the noisy STFT amplitude. However, later work by [9] shows that the importance of the phase depends on the settings and that it can be beneficial to estimate the STFT phase. Recently, in [10, 11], improved STFT amplitude estimates are obtained by using STFT phase estimates in the process.

Different approaches have been taken to modify the noisy STFT phase. In [12, 13], the change of STFT phase is based on the fact that not all STFT representations are consistent. Given a spectrum of a speech signal, an inverse STFT followed by an STFT leads back to the same spectrum, but if changes are made to the amplitude or phase of the spectrum, this is not necessarily the case for the altered spectrum, and it is, therefore, not consistent [14]. The quality of the resulting signal can be improved by minimising this inconsistency. In [12, 13] this is done by modifying the STFT phase to make a better match to the STFT amplitude estimate. The error on the phase is, therefore, not guaranteed to decrease because the phase is only modified to match the enhanced STFT amplitude. In [15], the STFT phase change in voiced speech periods is estimated based on the harmonic model and knowledge about the fundamental frequency. The phase in unvoiced periods is left unaltered, but since the major constituent of speech is voiced, changing the phase in these periods can still make a difference in terms of speech enhancement. Since only the phase change is estimated in [15], an initial phase estimate is needed as an anchor. In [15], the noisy phase is used as the initial STFT phase at the harmonic frequencies which gives a constant offset at each harmonic between the clean speech phase and the estimated phase and changes the relation between harmonics. This results in a significant error on the enhanced STFT phase, and the waveform of the resulting signal will be changed. In terms of perception, this is not a major problem if only a single harmonic is present, but in the case of more harmonics, as is the case in speech signals, it can have an influence on how the sound is perceived [16, 17].

To minimise the error on the phase, we propose a method to estimate the initial STFT phases in voiced speech periods. The method is based on the harmonic model and the model for phase evolution over time presented in [15]. The initial phases are estimated by setting up a least squares (LS) problem between the noisy phase and the signal model.

The paper is organised as follows: in Section 2 the harmonic signal model and the STFT are shortly introduced, in Section 3 the method from [15] is introduced, in Section 4 the proposed method is explained, results are presented in Section 5, and Section 6 concludes the work.

This work was funded by the Villum Foundation and the Danish Council for Independent Research, grant ID: DFF 1337-00084.

The work was carried out while Sidsel Marie Nørholm visited the Speech Signal Processing Group at the University of Oldenburg

2. Signal Model

We here use the harmonic signal model which is a good approximation to voiced speech. With this model the signal is composed of a set of harmonics with sinusoids having frequencies given by multiples of a fundamental frequency. For discrete time indices, $m = 0, ..., M - 1$, the signal can be represented as:

$$s(m) = \sum_{h=1}^{H} 2A_h \cos(\omega_0 hm + \varphi_h), \tag{1}$$

where H is the number of harmonics, A_h the amplitude of the h'th harmonic, $\omega_0 = 2\pi f_0/f_s$ the normalised fundamental angular frequency, with f_0 being the fundamental frequency and f_s the sampling frequency, and φ_h is the initial phase of the h'th harmonic. The desired signal is estimated from a mixture, $x(m)$, of the desired signal, $s(m)$, and additive noise, $v(m)$,

$$x(m) = s(m) + v(m). \tag{2}$$

The processing is done in the short-time Fourier transform (STFT) domain. The transformation to this domain is done by splitting the noisy signal into segments of length N, overlapping by $N - L$ samples, applying a window function $w(n)$ and computing the Discrete Fourier Transform (DFT), i.e.,

$$\begin{aligned} X(k,l) &= \sum_{n=0}^{N-1} x(lL+n)w(n)e^{-j\omega_k n} && (3)\\ &= |X(k,l)|e^{j\phi_X(k,l)}, && (4)\\ &= S(k,l) + V(k,l), && (5)\\ &= |S(k,l)|e^{j\phi_S(k,l)} + |V(k,l)|e^{j\phi_V(k,l)}, && (6) \end{aligned}$$

with k being the frequency index, l the segment index and $\omega_k = 2\pi k/N$ the normalised angular frequency of frequency band k. It can be seen in (4) that the signal in the STFT domain can be split into an amplitude part $|X(k,l)|$ and a phase part $e^{j\phi_X(k,l)}$. In many existing approaches only the amplitude is modified whereas the phase is not estimated, and the noisy phase is used directly, i.e., $\widehat{S(k,l)} = |\widehat{S(k,l)}|e^{j\phi_X(k,l)}$, where $\widehat{\{\cdot\}}$ denotes an estimated quantity. In this paper we will focus on estimating the clean phase $\phi_S(k,l)$ from the noisy phase $\phi_X(k,l)$.

3. Phase Reconstruction

In [15], the change in instantaneous phase in frequency bins containing the harmonic frequencies is estimated as a piecewise linear function when the harmonic frequency $\omega_h^{k,l} = h\omega_0^{k,l}$ is known, i.e.,

$$\begin{aligned} \Delta\phi_S(k,l) &= \phi_S(k,l) - \phi_S(k,l-1)\\ &= \omega_h^{k,l} L. && (7) \end{aligned}$$

The last equality holds under the assumption that the fundamental frequency in segments $l - 1$ and l are the same. Reformulation of (7) gives the instantaneous phase in segment l from the phase in segment $l - 1$

$$\widehat{\phi}_S(k,l) = \widehat{\phi}_S(k,l-1) + \omega_h^{k,l} L. \tag{8}$$

To get the instantaneous phase in segment l, it is therefore necessary to have information about the instantaneous phase in segment $l - 1$. In the very beginning of a piece of voiced speech, the algorithm has to be initialised with a phase for the first segment, i.e., information about the initial phases, φ_h, is needed. In [15], the noisy phase is used as an initialisation. This is illustrated in Fig. 1 where the baseband transformed phase (see [15]) in a frequency band containing a single harmonic of a frequency modulated signal is shown. It is seen that even though the phase evolution over time is correctly estimated with the method in [15] (KG[15]), using the noisy phase as an initialisation will give a constant offset between the clean phase and the estimated phase due to a wrong initial phase, ϕ_h. If only a single sinusoid is present, the initial phase is not that important in terms of perception, but if several harmonics are present, the relationship between the initial phases of the different harmonics has an influence on the shape of the waveform of the resulting signal and can also have an influence on how the sound is perceived [16, 17]. Therefore, we estimate the initial phases in the next section.

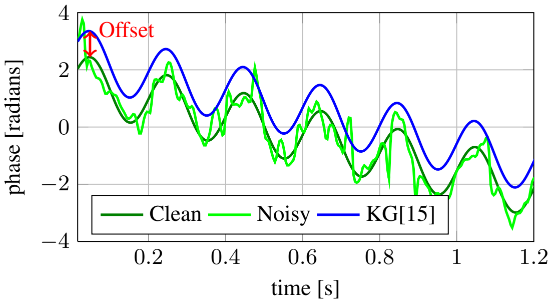


Figure 1: Reconstruction of the STFT phase based on KG[15] where the noisy phase is used as initialisation leading to a constant offset between the clean phase and the reconstructed phase.

4. Estimation of Initial Phases

The estimation of the initial phases is set up as a least squares (LS) problem between the instantaneous phases estimated using (8) with an initialisation of $\varphi_h = 0$ and the noisy phase for each harmonic separately

$$\widehat{\varphi}_h = \arg\min_{\varphi_h} \sum_{l=l_0}^{l_0+P-1} (\phi_X(k,l) - \widehat{\phi}_S(k,l) - \varphi_h)^2, \tag{9}$$

where P is the number of segments used for the estimation. The solution is found by differentiating the expression and equating with zero, i.e.,

$$\widehat{\varphi}_h = \frac{1}{P} \sum_{l=l_0}^{l_0+P-1} \phi_X(k,l) - \widehat{\phi}_S(k,l). \tag{10}$$

Due to the properties of the phase seen in (4), every $b2\pi, b \in \mathbb{Z}$, multiple of the phase gives rise to the same phase contribution to the resulting signal. This has to be taken into account in the estimation of the initial phase and, therefore, every phase difference in (10) is mapped to the interval $[-\pi, \pi]$, and the final estimate of the initial phase of harmonic h is given by:

$$\widehat{\varphi}_h = \frac{1}{P} \sum_{l=l_0}^{l_0+P-1} \angle(e^{j\phi_X(k,l) - j\widehat{\phi}_S(k,l)}), \tag{11}$$

where $\angle(\cdot)$ denotes the angle of the argument. To keep the right relation between frequency bins, all bins dominated by the given harmonic (see [15]) are also shifted according to the given estimate.

The method is implemented in two different ways. One where an entire piece of voiced speech is used for the estimation of the initial phase (denoted LS1 in the results section) and one where the initial phase of a given harmonic is reestimated each time the harmonic jumps to a new frequency bin (denoted LS2 in the results section). The first method has the advantage of more data used in the estimation and, therefore, if the model is perfectly correct, it should give a better estimate. However, it is vulnerable to errors in the model, e.g., a slightly wrong fundamental frequency estimation would lead to a model that over time deviates more and more from the clean signal and, thereby, gives larger errors in the estimation of the initial phase with more time segments used. The second method should do a better job in the case of a erroneous fundamental frequency estimate. However, in the transformation to the STFT domain the signal is overlapped which means that the noise in neighbouring time frames is not uncorrelated and, therefore, an estimation based on only a few frames would give an unreliable estimate.

The estimate of the initial phases introduces a latency in the system according to P. LS1 introduces a delay of one voiced speech period. The latency introduced by LS2 will depend on when the harmonics jump from one frequency bin to another and will, therefore, be smaller or equal to the latency introduced by LS1.

5. Results

The least squares estimates of the initial phases are first tested by means of a synthetic signal. After testing the concept on synthetic data, we turn to real speech signals. The synthetic signal used is a frequency modulated harmonic signal, i.e.,

$$s(m) = \sum_{h=1}^{H} A_h \cos(\omega_0 hm + \frac{\omega_\Delta}{\omega_m} h \cos(\omega_m m) + \varphi_h).$$

Here, $\omega_\Delta = 2\pi f_\Delta / f_s$ is the maximum deviation of the first harmonic away from ω_0 in one direction and $\omega_m = 2\pi f_m / f_s$ is the normalised angular modulation frequency. The signal is chosen because of its harmonic structure which is the basis of the proposed method and, further, it is a more interesting case than a pure harmonic signal since the fundamental frequency is modulated and, therefore, the harmonics will jump between different frequency bins when it is transformed to the STFT domain. Due to the multiplication by h in the modulation, the maximum deviation away from the harmonic frequency is increasing for higher harmonics, and they will, therefore, also have a higher tendency to jump between frequency bins. This will also be the case for speech signals. In the simulations $H = 10$, $f_s = 8000$, $M = 20000$, and f_0, f_Δ, f_m and φ_h are chosen randomly in intervals as $f_0 \in [100, 200]$ Hz, $f_\Delta \in [0, 10]$ Hz, $f_m \in [0, 10]$ Hz and $\varphi_h \in [-\pi, \pi]$. The frequency modulated signal is degraded by white Gaussian noise at signal-to-noise ratios (SNRs) from -10 dB to 10 dB in steps of 2.5 dB. The signal is transformed to the STFT domain in segments of 256 samples (corresponding to 32 ms) with an overlap of 87.5% and the window applied is a square root Hann window. In the evolution of the phase in the frequency domain, the true fundamental frequency is assumed to be known. The results are averaged over 1000 Monte Carlo simulations (MCS) [18]. The methods are evaluated both in the frequency and in the time domain. In the frequency domain, the phase ambiguities are again taken into account by using the circular phase error [6]:

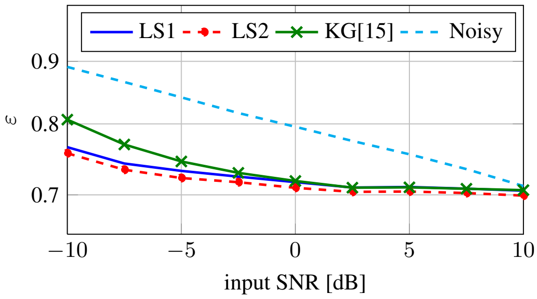


Figure 2: Phase error, ε, as a function of the input SNR averaged over all frequency bins and time for a synthetic signal. Averaged over 1000 MCS.

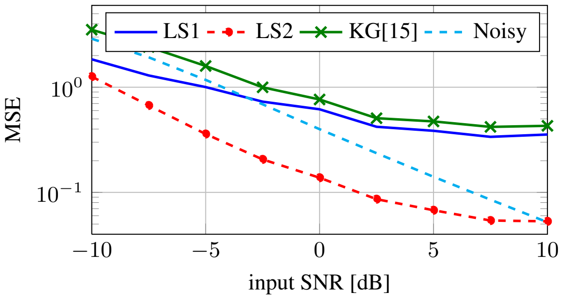


Figure 3: Mean squared error of reconstructed signal as a function of the input SNR for a synthetic signal. Combination of phase and clean amplitude. Average over 1000 MCS.

$$\varepsilon(k, l) = 1 - \cos(\phi^S_{k,l} - \widehat{\phi}^S_{k,l}), \tag{12}$$

which is in the range [0,2]. In the time domain they are evaluated by means of the mean squared error (MSE) between the clean signal, $s(m)$, and the reconstructed signal, $\widehat{s}(m)$, MSE $= (s(m) - \widehat{s}(m))^2$. The two methods are compared to the method in [15] where the noisy phase is used as an initialisation, here denoted by KG[15], and the noisy phase denoted by Noisy. In Fig. 2, the phase error averaged over all frequency bins and time is shown. It is seen that at all input SNRs considered here there is an advantage in estimating the instantaneous phase compared to using the noisy phase. Also, a smaller error can be obtained by estimating the initial phase. Both LS estimates give smaller errors than KG[15] up to approximately 0 dB input SNR, above 0 dB, LS2 gives a smaller error than KG[15] whereas LS1 gives the same error as KG[15]. The signal is thereafter reconstructed using an inverse STFT. Before doing that, the STFT phase term has to be multiplied with the STFT amplitude. For calculation of the mean squared error, we have used the clean speech amplitude, and the result is shown in Fig. 3. Now, KG[15] gives the highest error at all input SNRs, LS2 gives the lowest error whereas LS1 and the noisy phase give errors in between. The lower error of LS2 compared to LS1 shows that it is reasonable to take the jumps between frequency bins into account in the estimation process.

The methods are also evaluated using five male and five female speech signals from the TIMIT database degraded by

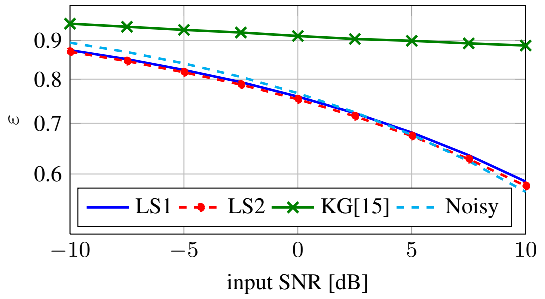


Figure 4: Phase error, ε, as a function of the input SNR averaged over all frequency bins and voiced speech periods for 5 male and 5 female speakers from the TIMIT database. Average over 50 MCS for each speaker.

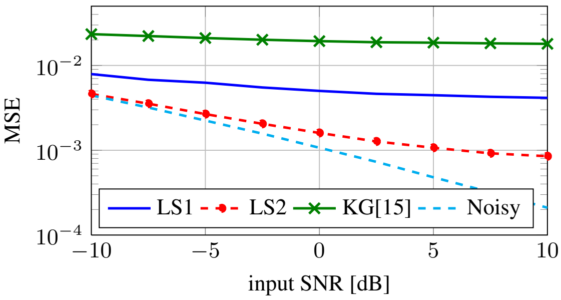
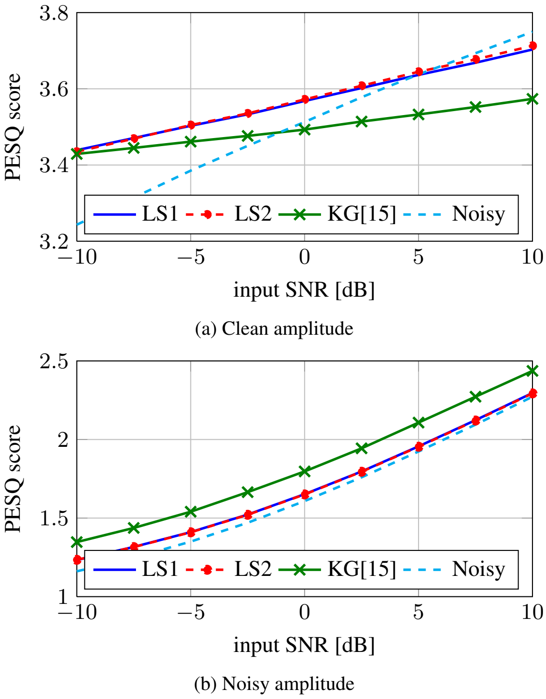


Figure 5: Mean squared error of reconstructed voiced speech parts as a function of the input SNR for 5 male and 5 female speakers from the TIMIT database. Combination of phase and clean amplitude. Average over 50 MCS for each speaker.

white Gaussian noise. The signals are downsampled to 8 kHz and the fundamental frequency is estimated from the clean speech signal using a nonlinear least squares estimator [19]. In the estimation, a search interval around (±10 Hz) the pitch obtained from the corresponding laryngograph signal [17] is used. The voiced periods are also chosen using the laryngograph track as the periods where the fundamental frequency is larger than zero. It is found that best results are obtained if only the lowest harmonics are modified so here the initial phases for the three first harmonics are estimated and changed. As in [15], the noisy phase is used directly in periods of unvoiced speech. The phase error is shown in Fig. 4, this time averaged over 50 MCS for each speaker, voiced speech periods and all frequency bins. The error on the phase using LS1 or LS2 is considerably decreased compared to KG[15], and LS2 again performs slightly better than LS1. Here, however, the error on the noisy phase is very similar to the error of LS1 and LS2, being slightly higher below 5 dB input SNR and slightly lower above 5 dB input SNR. Looking at the mean squared error of the reconstructed voiced speech parts in Fig. 5, it is seen that the error is again decreased when estimating the initial phase with LS1 or LS2 compared to using the noisy initial phase in KG[15], and again it is also more beneficial to use LS2 than LS1. Here, on the other hand, using the noisy phase at all times gives a slightly lower error on the reconstructed signal than LS2. The Perceptual Evaluation of Speech Quality (PESQ) score [20] of the reconstructed speech signals is also found. We have used two different choices of amplitudes in the reconstruction. These are the clean amplitude and the noisy amplitude. Using the clean amplitude, LS1 and LS2 performs best over the most of the range of input SNRs as seen in Fig. 6a whereas Fig. 6b shows that using the noisy amplitude, KG[15] gives the best PESQ score over the entire range. It would be more intuitive if a smaller error on the phase always would lead to a better reconstructed signal. The reason for this might be due to the inconsistency discussed in [14] and suggest that more work should be put into making consistent STFT representations based on both an amplitude and a phase estimate. However, better phase estimates on its own can still be used in, e.g., [10, 11] to give better reconstructed signals.



(a) Clean amplitude

(b) Noisy amplitude

Figure 6: PESQ score of reconstructed signal as a function of the input SNR for 5 male and 5 female speakers from the TIMIT database. Combination of phase and (a) clean amplitude and (b) noisy amplitude. Average over 50 MCS for each speaker.

6. Conclusion

In this paper, we considered speech enhancement in the STFT domain. Most prior work has been done on enhancing the noisy STFT amplitude, but the focus of this paper was the STFT phase. We suggest a least squares method to estimate the initial STFT phases in voiced speech periods. The initial phases are found by minimising the squared error between the noisy phase and the model-based phase estimates suggested in [15]. Simulations show that the error on the phase can be decreased considerably when estimating the initial phase as compared to using the noisy phase as the initial phase as proposed in [15]. The error on the phase is also reduced compared to the noisy phase in the ideal case with a synthetic signal and also slightly up to an input SNR of 5 dB when speech signals are considered. Reconstruction in combination with the clean amplitude gives an increase in PESQ score relative to KG[15] and an increase relative to the noisy phase up to an input SNR of 5 dB.

7. References

[1] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 4, pp. 1218–1234, 2006.

[2] J. R. Jensen, J. Benesty, M. G. Christensen, and S. H. Jensen, "Enhancement of single-channel periodic signals in the time-domain," *IEEE Trans. on Audio, Speech and Language Process.*, vol. 20, no. 7, pp. 1948–1963, Sep. 2012.

[3] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, Jul. 1995.

[4] J. Benesty, J. Chen, and E. A. P. Habets, *Speech enhancement in the STFT domain*. Springer, 2012.

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

[6] 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.

[7] D. Wang and J. S. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 30, no. 4, pp. 679–681, 1982.

[8] P. Vary, "Noise suppression by spectral magnitude estimation - mechanism and theoretical limits," *Signal Processing*, vol. 8, no. 4, pp. 387 – 400, 1985.

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

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

[11] 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.

[12] J. Le Roux, E. Vincent, Y. Mizuno, H. Kameoka, N. Ono, and S. Sagayama, "Consistent Wiener filtering: Generalized time-frequency masking respecting spectrogram consistency," in *Latent Variable Analysis and Signal Separation*. Springer Berlin Heidelberg, 2010, pp. 89–96.

[13] 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.

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

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

[16] B. C. J. Moore, *An introduction to the psychology of hearing*. Brill, 2012.

[17] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and Models*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[18] N. Metropolis, "The beginning of the monte carlo method," *Los Alamos Science*, no. 15, pp. 125–130, 1987.

[19] M. G. Christensen and A. Jakobsson, "Multi-pitch estimation," *Synthesis Lectures on Speech and Audio Processing*, vol. 5, no. 1, pp. 1–160, 2009.

[20] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 16, no. 1, pp. 229–238, 2008.