ON THE IMPORTANCE OF HARMONIC PHASE MODIFICATION FOR IMPROVED SPEECH SIGNAL RECONSTRUCTION

Anna Maly

Intelligent Acoustic Solutions
JOANNEUM RESEARCH, Austria
anna.maly@joanneum.at

Pejman Mowlaee

Signal Processing and Speech Communication Lab.
Graz University of Technology, Austria
pejman.mowlaee@tugraz.at

ABSTRACT

Conventional single-channel speech enhancement is mainly focused on modifying the noisy short-time Fourier transform amplitude spectrum while for signal reconstruction the noisy phase is used. Recent advances demonstrate the positive improvements in speech enhancement when the noisy phase is replaced with an estimated clean phase for signal reconstruction. In this paper, we study the impact of the linear phase and unwrapped phase components provided by harmonic phase decomposition on the speech quality at signal reconstruction. We present objective and subjective results comparing the proposed harmonic phase modification with other phase estimation methods. Our results show that enhancement of decomposed phase parts suffices for improved reconstruction in speech enhancement.

Index Terms: Phase spectrum, speech enhancement, phase decomposition, phase estimation.

1. INTRODUCTION

Conventional single-channel speech enhancement mostly relies on filtering the spectral amplitude of the noisy speech. For signal reconstruction the noisy spectral phase is often employed. The spectral phase information has been believed to be unimportant [1] or partly important for low signal-to-n ise ratio (SNRs) [2]. In contrast, recent studies [3–7] demonstrate the usefulness of the spectral phase in single-channel speech enhancement in terms of an improved perceived quality and speech intelligibility.

The studies on the phase importance in speech enhancement have been conducted in the short-time Fourier transform (STFT) domain where amplitude and phase spectra are extracted by applying Fourier transform on the windowed speech segments. The Fourier transformation is chosen because of its computational efficiency (for a review see e.g. [8]). While the spectral amplitude contains perceptually relevant information and harmonic structure, the STFT phase follows a uniform distribution and is often reported perceptually unimportant [1, 2]. More recently, harmonic model plus phase decomposition (HMPD) was proposed in [9], providing a compact representation for speech synthesis relying on phase variance and spectral envelope features. Further, the phase variance was shown to be a reliable metric to assess the voice quality of a synthesized speech signal [10]. Finally, positive improvements in speech enhancement have been reported due to phase-aware processing for signal reconstruction [4, 5, 11–15], joint amplitude and phase estimation [7, 16], and speech quality estimation [17, 18].

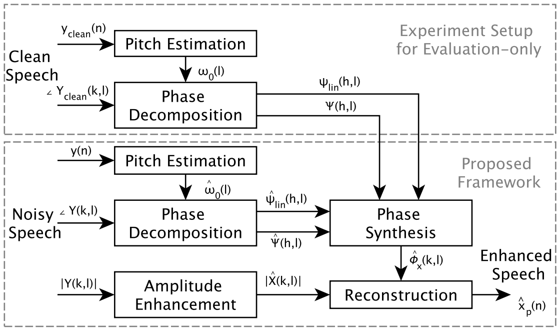

Fig. 1. (Lower part): block diagram for the proposed framework: harmonic phase modification combined with amplitude enhancement, (upper part): oracle phase parts for evaluation-only.

In this paper, we address the question of the importance of harmonic phase versus the STFT phase in single-channel speech enhancement. The block diagram for the proposed framework is shown in Figure 1. We argue the sufficiency of the harmonic phase components for signal reconstruction rather than seeking an improved STFT phase. To this end, we focus on the unwrapped harmonic phase after removing the linear phase using harmonic phase decomposition [9]. We investigate the importance of the *linear* and *unwrapped* phase separately when combined with a conventional amplitude-only speech enhancement. Objective and subjective evaluations are conducted to assess the perceived speech quality of the reconstructed signal.

The paper is organized as follows; In Section 2 we present the importance of phase for signal reconstruction and some previous studies. In Section 3, we present different combinations for harmonic phase components used for signal reconstruction. In section 4 we present results. Section 5 concludes on the work.

2. PHASE IMPORTANCE IN SIGNAL RECONSTRUCTION

Given an enhanced STFT spectral amplitude, a proper spectral phase information is required to reconstruct a speech signal. However, for the following reasons the noisy spectral phase has been often selected as a common choice in speech enhancement: i) the phase has been considered unimportant for human perception [1], ii) the wrapping issue in the phase spectrum makes it difficult to estimate a clean phase given the noisy spectrum, iii) the noisy phase spectrum has been shown to be sufficient when the local signal-to-noise ratio is above 6 decibels [2], iv) given the independence assumption

A part of this work was done when Anna Maly was affiliated with SPSC, Graz. This work was supported by the Austrian Science Fund (project number P28070-N33). We thank Gilles Degottex for his productive discussions.

between the DFT amplitude and phase spectra, it is equal to the minimum mean square error (MMSE) [19] and maximum a posteriori (MAP) [20] estimate for the clean spectral phase.

Recently, researchers demonstrated the usefulness of the spectral phase to improve the naturalness of the synthesized speech signal by incorporating a proper phase at signal reconstruction stage [21]. We consider different combinations of harmonic phase components at signal reconstruction stage in single-channel speech enhancement.

3. HARMONIC PHASE DECOMPOSITION

3.1. Notations

Let $x(n)$ and $d(n)$ be clean speech and noise signals, respectively, with $y(n) = x(n) + d(n)$ as the noisy signal with n as time sample index. Let k and l be the frequency and time frame indices. We define $X(k,l) = |X(k,l)|e^{j\phi_x(k,l)}$ and $Y(k,l) = |Y(k,l)|e^{j\phi_y(k,l)}$ as the DFT coefficients for clean and noisy signals, respectively, and $|X(k,l)|$ and $|Y(k,l)|$ as the spectral amplitude for clean and noisy speech and $\phi_y(k,l)$ and $\phi_x(k,l)$ as spectral phase.

3.2. Harmonic Model Plus Phase Decomposition

Harmonic model with phase decomposition was proposed in [21]. As signal segmentation, a pitch-synchronous analysis is required to perform the harmonic model phase decomposition using $t(l) = t(l-1) + \frac{1}{4f_0(l-1)}$ where $t(l)$ and $t(l-1)$ are the time instants for the lth and $(l-1)$th frames, respectively, with $f_0(l)$ denoting the fundamental frequency at the lth frame. In voiced regions each frame can be modeled as sum of harmonics, consisting of amplitude $|X(h,l)|$ and instantaneous phase $\phi(h,l)$ with h as the harmonic index. The instantaneous STFT phase sampled at harmonic h and frame l is decomposed into the *linear* and *unwrapped* parts:

$$\psi(h,l)=\underbrace{h\sum_{l'=0}^{l}\omega_0(l')\big(t(l')-t(l'-1)\big)}_{\text{Linear phase: }\psi_{\text{lin}}(h,l)}+\underbrace{\angle V(h,l)+\psi_d(h,l)}_{\text{Unwrapped phase: }\Psi(h,l)}, \tag{1}$$

where $\psi_{\text{lin}}(h,l)$ is the linear phase and $\Psi(h,l)$ is the unwrapped phase. We define $V(h,l)$ as the vocal tract filter. Then, the minimum phase part is the phase response of the vocal tract filter $V(h,l)$ and we have $\psi_{\text{min}}(h,l) = \angle V(hf_0(l),l)$. The linear phase component imposes wrapping in the instantaneous STFT phase. Discontinuity in the linear phase results in certain degradation in the perceived speech quality of the reconstructed speech signal [22]. Given a fundamental frequency estimate at frame l denoted by $\omega_0(l) = 2\pi f_0(l)/f_s$, where $f_0(l)$ and f_s denote the fundamental frequency and the sampling frequency, respectively. Then the linear phase is approximated as:

$$\psi_{\text{lin}}(h,l) = h\sum_{l'=0}^{l}\omega_0(l')(t(l')-t(l'-1)). \tag{2}$$

The linear phase wraps the instantaneous phase across time according to h and the time gap $t(l)-t(l-1)$. The unwrapped phase defined in (1) denoted by $\Psi(h,l)$ is itself composed of minimum phase $\psi_{\text{min}}(h,l)$ and phase dispersion $\psi_d(h,l)$:

$$\Psi(h,l) = \psi_{\text{min}}(h,l) + \psi_d(h,l). \tag{3}$$

The unwrapped phase is calculated via subtracting the linear phase part from the instantaneous phase. Finally, the last term in (3), called phase dispersion or source shape, captures the pulse shape in the underlying speech segment. It captures the stochastic characteristics of the harmonic phase. Unlike the STFT phase it represents a non-uniform distribution, as von Mises distribution characterized by mean and variance parameters, as used for phase estimation in single-channel speech enhancement [14, 15].

3.3. Phase Reconstruction

In speech coding and speech synthesis literature, several different combinations of the harmonic phase components have been recommended; the combination of a linear phase and the minimum phase was used in the sinusoidal analysis/synthesis model proposed by McAualay and Quatieri [23]. Vary in [2] studied the impact of phase modification in DFT-based speech enhancement concluding that zero phase and random phase contribute to harmonic monotonous or some perceptually audible rough quality, respectively. Model-based STFT phase reconstruction (STFTPI) was proposed in [11] relying on modifying phase at harmonics using phase prediction across frames and applying window phase compensation across frequency.

More recently, the source shape phase component was taken into account modeled in terms of mean and variance capturing the noisiness and breathness of the synthesized speech using HMPD model [9], reporting a high perceived reconstructed speech quality. Further, phase variance was reported as reliable metric for quality assessment of synthesized speech [10]. At voiced frames, the variance of phase distortion should be small as the glottal pulse shape is preserved in time [10]. This property motivated for the Temporal Smoothing of Unwrapped Phase (TSUP) for speech enhancement [15]. In TSUP method first the linear phase is subtracted from the instantaneous phase using a fundamental frequency estimate as described in (3). The unwrapped phase is then smoothed along time. Finally, the linear phase part is added back to the temporally smoothed unwrapped phase to obtain an enhanced STFT phase eventually used for signal reconstruction.

In [12] Sugiyama proposed phase randomization showing positive impact on the achievable noise reduction performance. The randomization of the noisy phase was shown effective at certain regions given a reliable estimation for the signal-to-noise ratio. The requirement of an accurate SNR information restricts the effectiveness of the phase randomization scheme for noise reduction.

While both linear and unwrapped phase estimation are carried out at harmonics, in order to combine it with an amplitude enhancement scheme, an enhanced phase is required in the STFT domain (see Figure 1). Therefore, similar to [15], we transform harmonic phase estimate $\hat{\psi}(h,l)$ back to STFT by modifying the frequency bins within the width of the window main-lobe denoted by N_p:

$$\hat{\phi}_x(\lfloor h\omega_0 K\rfloor + i, l) = \hat{\psi}(h,l), \quad \forall i \in [-N_p/2, N_p/2]\,, \tag{4}$$

with K as DFT size.

4. RESULTS

4.1. Experiment Setup

We selected 50 utterances from the GRID corpus [24] composed of male and female speakers. Each speech utterance was corrupted with babble noise taken from NOISEX-92 [25], mixed at signal-to-noise ratios from -5 to 15 decibels. As our evaluation criteria, we chose the perceptual evaluation of speech quality (PESQ) [26] and the short-time objective intelligibility measure (STOI) [27] as predictor of perceived quality and speech intelligibility of the enhanced speech. The fundamental frequency for phase decomposition is given by pitch estimation filter with amplitude compression (PEFAC) [28]. The phase information is extracted by applying harmonic model phase decomposition implemented in COVAREP [29]. The length of the analysis window is set equal to 24 ms at a sampling frequency of 16 kHz. As frame setup, we chose Blackman window with a frame shift of 2 ms.

As shown in Figure 1, in order to study the potential of an oracle phase scenario, the results obtained when clean signal is provided for pitch estimation or clean phase are also included. This comparative study reveals the potential by the oracle scenario for linear phase and unwrapped phase components. In all experiments, the phase estimation is combined with the conventional amplitude enhancement (denoted as $|\hat{X}(k,l)|$) given by:

$$|\hat{X}(k,l)| = G(k,l)|Y(k,l)|, \tag{5}$$

where $G(k,l)$ is the noise suppression rule given by a look up table determined by *prior* and *posterior* SNRs. As noise estimator minimum statistic noise estimator [30] was used and as speech PSD estimator we used minimum mean square error log-spectral amplitude (MMSE-LSA) estimator [31] with the decision-directed approach [19]. Finally, the phase enhanced spectrum is given by

$$\hat{X}_p(k,l) = |\hat{X}(k,l)|e^{j\hat{\phi}_x(k,l)}. \tag{6}$$

Using overlap-add, the phase-enhanced signal $\hat{x}_p(n)$ is produced.

4.2. Proof-of-Concept Experiment

Figure 2 shows the proof-of-concept visualizing the impact of selecting different phase parts at signal reconstruction. The results are shown as spectrogram (top), group delay (middle) and phase variance (bottom). The female utterance saying *"been blue at L four soon"* is corrupted in white noise at SNR $=0$ decibels. The utterance consists of plosive [b], vowel [U:], fricative [s],[f], known for their high speech intelligibility contribution [32].

The output signals are reconstructed using different combinations of unwrapped (Ψ) and linear (ψ_{lin}) phase, as noisy or estimated versions. The first column shows the clean signal as the upper bound while the other columns show the enhanced amplitude ($A.e.$) with noisy (n) phase (lower bound). With linear phase-only, there will be over-harmonization, obviously not sufficient to provide an improved signal reconstruct on. Similar buzzyness effect was reported in the reconstructed speech signals reported in [11, 17, 33], where a lower speech intelligibility [18] than the noisy phase was experienced (see ΔSTOI results in Figure 5). Further inspection of the phase dispersion component elaborates the role of an unwrapped phase in capturing the noisiness or breathness (the stochastic phase part), not modeled by a linear phase-only, still playing an important role for a high quality synthesized speech [9, 10]. The results using the oracle unwrapped phase are much closer to the results of the clean signal (clean amplitude + clean phase), confirming that the linear phase plus a proper unwrapped phase is the desired combination to employ for an improved signal reconstruction of speech enhancement.

4.3. Subjective Listening Results

In order to justify the results reported in the previous Section predicted by instrumental measures, here we conduct a subjective listening test following the comparison category rating (CCR) test [34]. A panel of thirteen listeners participated in the test, all expert listeners from TU Graz. The listening test was conducted in a quiet room using closed-back professional monitor headphones. The listeners were asked to rate the oracle linear phase together with oracle unwrapped phase from *much better* to *much worse* compared to the enhanced signals using clean STFT phase. The comparison was quantified into seven steps where 3 corresponds to *much better*, and -3 corresponds to *much worse*.

The result is shown in Figure 4. The speech outcome from oracle linear and unwrapped phase was rated between indistinguishable to slightly worse vrsus to the clean STFT phase. This result confirms that a proper modification of the noisy phase at harmonics, using an oracle fundamental frequency together with a proper smoothed unwrapped phase (hence a reduced phase variance) provides a similar perceptual quality compared to clean phase when combined with amplitude-only enhancement using MMSE-LSA [31].

4.4. Speech Quality and Speech Intelligibility Evaluation

4.4.1. Importance of the Harmonic Phase Parts

In this section, we investigate the importance of unwrapped phase and linear phase components. Via comparing the oracle and the estimated versions of these phase components and considering their possible combinations, we address the question which part of the harmonic phase components plays the most important role. This also answers the question, whether fundamental frequency estimation error (for linear phase), or a successful smoothing filter (for a continuous unwrapped phase), plays the main role in phase estimation for speech enhancement.

Figure 3 shows speech quality and speech intelligibility results instrumentally predicted by PESQ (left) and STOI (right), respectively. All the harmonic phase modification methods showed improved performance versus the noisy phase outcome. The curves with noisy unwrapped phase illustrate the lower bound of all possible phase combinations. The PESQ and STOI scores achieved by the oracle STFT phase are reported for comparison purposes. Phase decomposition with the oracle knowledge about the fundamental frequency results in a performance close to the clean STFT phase, outperforming the noisy phase in PESQ. With oracle linear and unwrapped phase the intelligibility performance is closest to the noisy phase scenario. This observation highlights the fact that a proper phase modification reduces the well-known degraded intelligibility effect due to amplitude-only enhancement schemes [35, 36]. These results further justify that a proper linear phase provided by an accurate f_0-estimator together with a successful modification of the unwrapped phase contribute the most to an improved quality at signal reconstruction in speech enhancement.

4.4.2. Comparisons with Other Phase Estimators

To address the question of sufficiency of harmonic phase modification versus the STFT phase modification, here, we consider several recent phase estimation methods. Our comparative study demonstrates the potential and limits of a proper harmonic phase modification. The benchmarks are: i) phase randomization [12], ii) STFTPI [11] and iii) Temporal Smoothing of Unwrapped Phase (TSUP) [15]. Further, for comparison purposes, we include the upper-bound performance provided by the STFT clean phase. In all methods, the output signal is produced via combining the enhanced amplitude associated with the modified phase following (6). Some audio examples are available at http://www2.spsc.tugraz.at/people/pmowlaee/ICASSP2016.html.

The results are shown in Figure 5. The improvement achieved in terms of perceived quality and speech intelligibility are reported

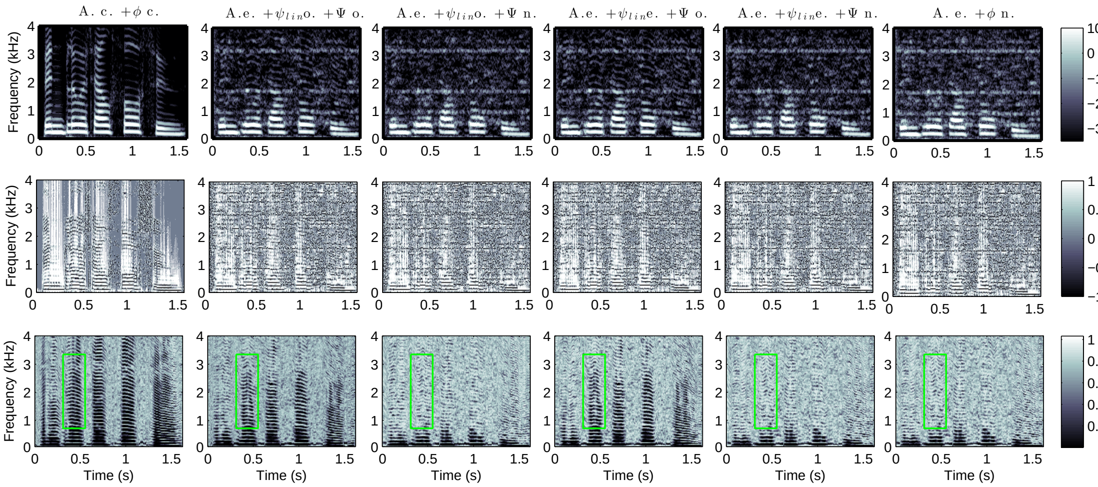

Fig. 2. Proof-of-concept results for enhanced speech in white noise; (top) spectrogram, (middle) group delay, (bottom) phase variance. The spectral amplitude is taken from MMSE-LSA, while the spectral phase is used using the following candidates (from left to right): clean signal, oracle linear phase + oracle unwrapped phase, oracle linear phase + noisy unwrapped phase, noisy linear phase + oracle unwrapped phase, noisy linear phase + noisy unwrapped phase, noisy phase. The following notations are used; o: oracle, n: noisy, e: estimated, c: clean.

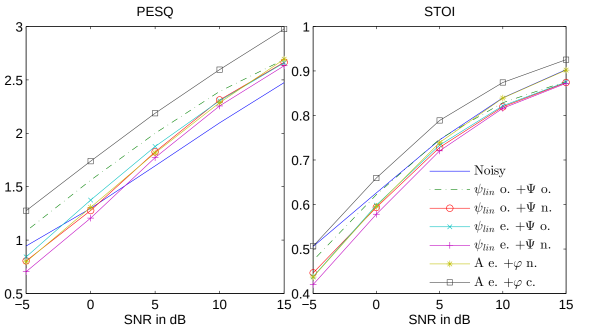

Fig. 3. PESQ and STOI results for different harmonic phase modifications in babble noise (o: oracle, n: noisy, e: estimated, c: clean).

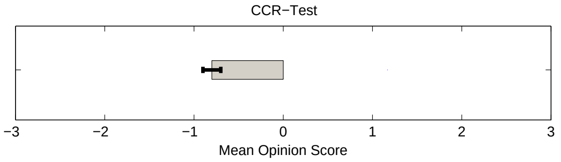

Fig. 4. Subjective listening results for CCR test, comparing the oracle harmonic phase versus clean STFT phase.

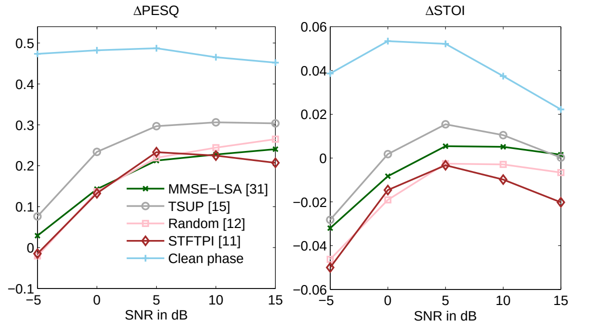

Fig. 5. ΔPESQ and ΔSTOI phase-only enhancement results for different phase estimation in babble noise.

in terms of ΔPESQ and ΔSTOI quantifying the improvement compared to that achieved by the conventional amplitude-only enhancement scheme (where noisy phase is employed at signal reconstruction). TSUP [15] results in joint improvement in both perceived quality and speech intelligibility. TSUP relies on modification of linear phase and unwrapped phase both accessed by the harmonic phase decomposition. The method outperforms other methods relying on modification of noisy STFT phase or liner phase. This observation validates the importance of joint modification of linear and unwrapped phase in order to achieve improved signal reconstruction (quality and intelligibility) in single-channel speech enhancement.

5. CONCLUSION

In this paper, we studied the importance of harmonic phase modification for improved single-channel speech enhancement performance. The oracle linear and unwrapped phase components were shown to suffice to reach to an indistinguishable performance compared to clean STFT phase upper-bound. Throughout comparison with existing phase enhancement methods relying on STFT phase, our objective and subjective results showed that a proper modification of harmonic phase contribute to improved phase-only speech enhancement in terms of speech quality and intelligibility performance. Future work will be dedicated to improving the phase estimation performance by applying more accurate phase unwrapping proposals in [37] together with accurate pitch estimator in noise. With a significant improvement in the estimation of the clean speech phase, further improved speech enhancement is expected when such accurate phase knowledge is used in a phase-aware amplitude estimator.

6. REFERENCES

[1] D. Wang and J. Lim, “The unimportance of phase in speech enhancement,” *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 30, no. 4, pp. 679–681, 1982.

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

[3] K. K. Paliwal, K. K. Wojcicki, and B. J. Shannon, “The importance of phase in speech enhancement,” *speech communication*, vol. 53, no. 4, pp. 465–494, 2011.

[4] P. Mowlaee and J. Kulmer, “Phase estimation in single-channel speech enhancement: Limits-potential,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 23, no. 8, pp. 1283–1294, Aug. 2015.

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

[6] P. Mowlaee, R. Saeidi, and Y. Stylianou, “Interspeech 2014 special session on phase importance in speech processing,” *Proc. Interspeech*, pp. 1623–1627, 2014.

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

[8] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement*, Synthesis Lectures on Speech and Audio Processing. Morgan & Claypool Publishers, 2013.

[9] G. Degottex and D. Erro, “A uniform phase representation for the harmonic model in speech synthesis applications,” *EURASIP J. on Audio, Speech, and Music Processing*, vol. 2014, no. 1, pp. 38, Oct. 2014.

[10] M. Koutsogiannaki, O. Simantiraki, G. Degottex, and Y. Stylianou, “The importance of phase on voice quality assessment,” in *Proc. Interspeech*, Sept. 2014.

[11] 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.

[12] A. Sugiyama and R. Miyahara, “Phase randomization - a new paradigm for single-channel signal enhancement,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2013, pp. 7487–7491.

[13] P. Mowlaee, R. Saiedi, and R. Martin, “Phase estimation for signal reconstruction in single-channel speech separation,” in *Proc. Interspeech*, 2012, pp. 1–4.

[14] J. Kulmer and P. Mowlaee, “Harmonic phase estimation in single-channel speech enhancement using von Mises distribution and prior SNR,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Apr. 2015, pp. 5063–5067.

[15] J. Kulmer and P. Mowlaee, “Phase estimation in single channel speech enhancement using phase decomposition,” *IEEE Signal Process. Lett.*, vol. 22, no. 5, pp. 598–602, May. 2015.

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

[17] A. Gaich and P. Mowlaee, “On speech quality estimation of phase-aware single-channel speech enhancement,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2015, pp. 216–220.

[18] A. Gaich and P. Mowlaee, “On speech intelligibility estimation of phase-aware single-channel speech enhancement,” in *Proc. Interspeech*, 2015, pp. 2553–2557.

[19] Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[20] T. Lotter and P. Vary, “Speech enhancement by MAP spectral amplitude estimation using a super-gaussian speech model,” *EURASIP J. Adv. Sig. Proc.*, vol. 2005, no. 7, pp. 1110–1126, 2005.

[21] G. Degottex and D. Erro, “A measure of randomness for harmonic model in speech synthesis,” *in Proceedings of the International Conference on Spoken Language Processing*, 2014.

[22] Y. Stylianou, “Removing linear phase mismatches in concatenative speech synthesis,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 9, no. 3, pp. 232–239, Mar 2001.

[23] R. McAulay and T. Quatieri, “Speech analysis/synthesis based on a sinusoidal representation,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 4, pp. 744–754, Aug 1986.

[24] M. Cooke, J. Barker, S. Cunningham, and X. Shao, “An audio-visual corpus for speech perception and automatic speech recognition,” *J. Acoust. Soc. Am.*, vol. 120, no. 5, pp. 2421–2424, 2006.

[25] A. Varga, H. J. M. Steeneken, M. Tomlinson, and D. Jones, “The NOISEX–92 Study on the Effect of Additive Noise on Automatic Speech Recognition,” *Technical Report, DRA Speech Research Unit*, 1992.

[26] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs,” *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, pp. 749–752, Aug 2001.

[27] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time-frequency weighted noisy speech,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[28] S. Gonzalez and M. Brookes, “PEFAC-a pitch estimation algorithm robust to high levels of noise,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 22, no. 2, pp. 518–530, 2014.

[29] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, “COVAREP - a collaborative voice analysis repository for speech technologies,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, May 2014, pp. 960–964.

[30] R. Martin, “Noise power spectral density estimation based on optimal smoothing and minimum statistics,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 9, no. 5, pp. 504–512, 2001.

[31] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean square error log-spectral amplitude estimator,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. ASSP-33, pp. 443–445, 1985.

[32] L. Li, H. Jialong, and P. Günther, “Effects of phase on the perception of intervocalic stop consonants,” *speech communication*, vol. 22, no. 4, pp. 403–417, Sept. 1997.

[33] S. P. Patil and J. N. Gowdy, “Exploiting the baseband phase structure of the voiced speech for speech enhancement,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Apr. 2014, pp. 6133–6137.

[34] “ITU-T P.800. Methods for subjective determination of transmission quality - Series P: telephone transmission quality; methods for objective and subjective assessment of quality,” Aug. 1996.

[35] P. C. Loizou and G. Kim, “Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 1, pp. 47–56, Jan 2011.

[36] J. Jensen and C.H. Taal, “Speech intelligibility prediction based on mutual information,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 22, no. 2, pp. 430–440, Feb 2014.

[37] T. Drugman and Y. Stylianou, “Fast and accurate phase unwrapping,” in *Proc. Interspeech*, 2015, pp. 1171–1175.