PHASE ESTIMATION IN SINGLE-CHANNEL SPEECH ENHANCEMENT USING PHASE INVARIANCE CONSTRAINTS

Michael Pirolt, Johannes Stahl*, Pejman Mowlaee*,*
*Vasili I. Vorobiov**, Siarhei Y. Barysenka**, Andrew G. Davydov***

* Signal Processing and Speech Communication Lab, Graz University of Technology, Graz, Austria
** Belarusian State University of Informatics and Radioelectronics, Minsk, Belarus
michael.pirolt@student.tugraz.at, johannes.stahl@tugraz.at,
pejman.mowlaee@tugraz.at, viv314@gmail.com,
siarhei.barysenka@gmail.com, agdavydov81@gmail.com

ABSTRACT

Phase-aware signal processing has received increasing interest in many speech applications. The success of phase-aware processing depends strongly on the robustness of the clean spectral phase estimates to be obtained from a noisy observation. In this paper, we propose a novel harmonic phase estimator relying on the phase invariance property exploiting relations between harmonics using the phase structure. We present speech quality results achieved in speech enhancement to justify the effectiveness of the proposed phase estimator compared to noisy phase and other phase estimation benchmarks.

Index Terms— Phase estimation, phase invariance, speech enhancement, speech quality.

1. INTRODUCTION

Speech signal processing methods often ignore the processing of spectral phase information. Performance gain can be achieved when an enhanced spectral phase or some additional information about phase is incorporated. For a general review on recent advances in phase-aware signal processing, its application in speech communication and why it has been neglected we refer to [1–3].

In particular, in the field of noise reduction the importance of phase receives increasing attention by researchers. Some examples for recent works are, model-based short-time Fourier transform (STFT) phase improvement [4], maximum a posteriori harmonic (MAP) phase estimation [5], temporal smoothing of the unwrapped harmonic phase (TSUP) [6], and finally the phase estimation impact on enhancement have been investigated in [7]. Apart from improved signal reconstruction, spectral phase information can be also used to derive improved spectral amplitude estimators, see e.g. [8–10].

The work of Pejman Mowlaee, Michael Pirolt and Johannes Stahl was supported by the Austrian Science Fund (project number P28070-N33).

The benefits from phase-aware processing are limited by the accuracy of the estimated phase. Therefore, a challenging research topic is to find novel approaches that help to achieve more robust and accurate estimators of the clean spectral phase from the noisy speech observation.

In this paper, we propose exploiting the relation between the phase of harmonics of a speech signal. The so-derived harmonic phase estimator results in improved perceived quality and speech intelligibility, and a low phase estimation error.

The rest of the paper is organized as follows. Section 2 presents the background on phase invariance and phase quasi-invariance properties. Section 3 presents the proposed phase enhancement scheme. Section 4 presents a proof-of-concept experiment and speech enhancement results and Section 5 concludes the work.

2. BACKGROUND ON THE PHASE INVARIANCE PROPERTY

2.1. Phase Invariant

The phase invariant constraint (PI) was first introduced by Zverev in ultrasonic dispersion measurements [11], where it was reported that harmonic oscillation contains a phase structure which is invariant to the time reference. In harmonic signals, the PI can be determined for any triplet of harmonic components if their frequencies satisfy the set of equations:

$$\begin{cases} f_1 = K_1 F_0, & \text{where } K_1 = 1, 2, ... \\ f_2 = K_2 F_0, & \text{where } K_2 = K_1 + 1, K_1 + 2, ... \\ f_3 = K_3 F_0, & \text{where } K_3 = 2K_2 - K_1. \end{cases} \quad (1)$$

In these equations, F_0 denotes the fundamental frequency. We consider a polyharmonic signal $s(t)$ with time index t consisting of H_t harmonics. Given the fundamental frequency F_0, each of the harmonics is characterized by the harmonic index $h \in [1, H_t]$ and the corresponding amplitude $A(h,t)$ and phase $\Phi(h,t)$, both assumed to be slowly varying in time:

$$s(t) = \sum_{h=1}^{H_t} s(h,t) = \sum_{h=1}^{H_t} A(h,t)\cos\underbrace{(2\pi h F_0(t)t + \Phi(h,t))}_{\Psi(h,t)}. \tag{2}$$

The PI denoted by $\Delta\Psi(t)$, for $H_t = 3$ is given by:

$$\begin{aligned}\Delta\Psi(t) &= \frac{\Psi(1,t)+\Psi(3,t)}{2} - \Psi(2,t) \\ &= \frac{\Phi(1,t)+\Phi(3,t)}{2} - \Phi(2,t),\end{aligned} \tag{3}$$

where $\Psi(h,t)$ denotes the instantaneous phase. It is important to mention that cancellation of the linear items $2\pi h F_0(t)t$ can be achieved only when the instantaneous phase $\Psi(h,t)$ is continuous and has no wraps. This can be ensured by using the phase unwrapping procedure.

2.2. Phase Quasi-Invariant

The phase quasi-invariant constraint (PQI) was introduced by Vorobiov within the analysis of phase relations in speech [12]. The application of this constraint together with the PI was outlined for speech analysis [13, 14].

Again considering equation (2) the following relation $\Delta\Psi_{\bar{h}}(h,t)$ between components with frequencies $\bar{h}F_0(t)$ and $hF_0(t)$, where $\bar{h} < h$, is free of linear components similar to PI in Eq. (3):

$$\begin{aligned}\Delta\Psi_{\bar{h}}(h,t) &= \Psi(\bar{h},t) - \frac{\Psi(h,t)\cdot\bar{h}}{h} \\ &= \left(\Phi(\bar{h},t) - \frac{\Phi(h,t)\cdot\bar{h}}{h}\right)\Bigg|_{\frac{2\pi\bar{h}}{h}}.\end{aligned} \tag{4}$$

The equation above is called *phase quasi-invariant* (PQI). It is also required to unwrap the instantaneous phase functions $\Psi(\bar{h},t)$ and $\Psi(h,t)$ before calculating the PQI. The unambiguous definition range of the PQI is $[0, \frac{2\pi\bar{h}}{h})$ if the harmonic phase $\Phi(\bar{h},t)$ at time instant $t = 0$ is within the interval $[0, 2\pi)$.

Another signal representation based on the phase difference measure is the Relative Phase Shift (RPS) [15]. The relation between RPS and PQI can be depicted as:

$$RPS(h,t) = \Phi(h,t) - h\Phi(1,t) = -h\Delta\Psi_1(h,t). \tag{5}$$

Eq. (5) shows that the RPS can be represented by the negative PQI with $\bar{h} = 1$, multiplied with the harmonic index. While the RPS depicts the phase difference only between the fundamental frequency and its higher harmonics, the PQI is not limited to the fundamental frequency phase, as the reference harmonic $\bar{h}$ is free to choose.

2.3. Suitability for Phase-Aware Speech Processing

To demonstrate the smoothness of the PQI and the PI along time in voiced speech, records of sustained vowels A-E-I-O-U were analyzed in PI and PQI domain, respectively. Figure 1 illustrates the analysis for male and female speakers[1].

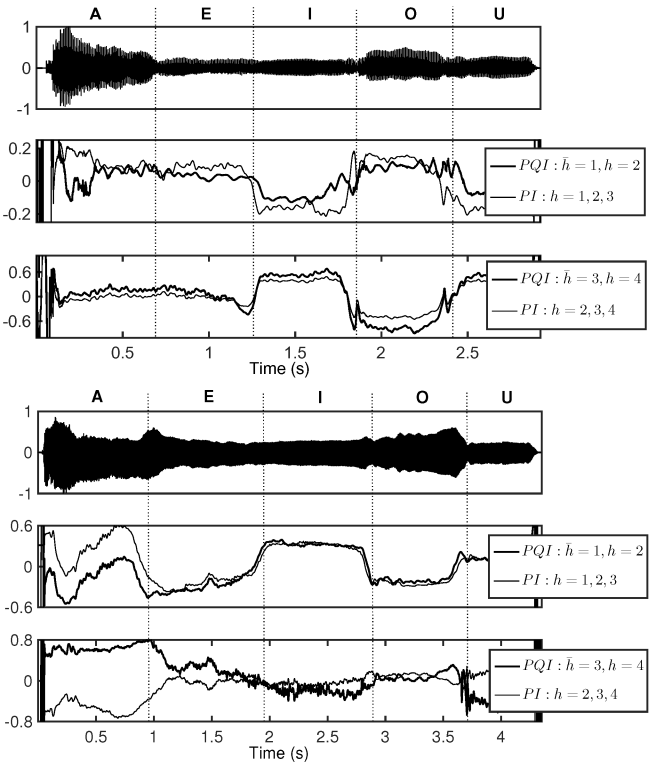

Fig. 1. The results of proof-of-concept experiments of PI and PQI for sustained A-E-I-O-U sequence. (Top) Male speaker. $F_0 = 118 \pm 2$ Hz along the whole record, (Bottom) Female speaker. $F_0 = 220 \pm 5$ Hz along the whole record.

First, the pitch estimate was obtained using the PEFAC algorithm [17]. The instantaneous phase functions $\Psi(h,t)$ were calculated using the Hilbert transform for filtered $hF_0(t)$ where $h \in [1, 4]$. After phase unwrapping the phase characteristics PI and PQI were calculated. In order to unify the scale of these representations for illustration purposes, the PQI was normalized to half of its unambiguous definition range, whereas the PI was normalized to π.

The PI and the PQI show similar trends, e.g., at phoneme transitions that entail abrupt changes in both curves. These results support that the PI and the PQI both carry information about the structure of voiced speech, so they are favorable candidates for phase-aware speech processing.

3. PROPOSED PHASE ESTIMATOR

The idea of this work is to apply temporal smoothing on the PQI extracted from the noisy speech signal in order to reduce its variance. This is motivated by the successful results reported in TSUP [6, 7] and will be justified within the proof-of-concept experiment presented in this Section. An overview of the proposed method is depicted in Figure 2.

[1]The implementation for this experiment can be found at [16].

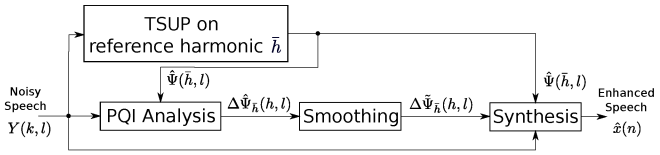

Fig. 2. Illustration of the proposed phase estimator where the phase-enhanced speech $\hat{x}(n)$ is estimated given the noisy speech.

3.1. PQI Framework

In this Section, the proposed phase enhancement framework is presented. As explained in Section 2, the PQI is based on the phase difference measures between two harmonics. Therefore we model the noisy signal as the sum of harmonics corresponding to the clean signal $x(n)$ with some noise added. The noisy signal is represented by an assembly of signal frames $y(n,l)$ where $n \in [0, N-1]$ denotes the discrete time index, N the frame length and l denotes the frame index and we have:

$$y(n,l) = \underbrace{\sum_{h=1}^{H_l} A(h,l)\cos\left(h \cdot 2\pi \frac{F_0(l)}{f_s} n + \Phi(h,l)\right)}_{x(n)\ldots\text{clean signal}} + \nu(n,l), \tag{6}$$

where $\nu(n,l)$ denotes the noise and h denotes the harmonic index with $h \in [1, H_l]$ and H_l denotes the number of harmonics at frame l. The time instances at each frame t_l are calculated according to [18]:

$$t_l = t_{l-1} + \frac{1}{4 \cdot F_0(l-1)}. \tag{7}$$

3.2. Calculation of PQI

The PQI values are calculated based on Eq. (4). Since the output of Eq. (4) gives us a cyclic random variable with the unambiguous definition range of $\left[\frac{-\pi\bar{h}}{h}, \frac{\pi\bar{h}}{h}\right)$, it is recommended to add a scaling factor after wrapping to ensure an unambiguous definition range of $[-\pi, \pi)$, please note that the PQI is independent of the fundamental frequency, therefore it yields:

$$\begin{aligned}\Delta\Psi_{\bar{h}}(h,l) &= \frac{h}{\bar{h}}\left(\Phi(\bar{h},l) - \frac{\Phi(h,l)\cdot\bar{h}}{h}\right)\Bigg|_{\frac{2\pi\cdot\bar{h}}{h}} \\ &= \frac{h}{\bar{h}}\left(\Psi(\bar{h},l) - \frac{\Psi(h,l)\cdot\bar{h}}{h}\right)\Bigg|_{\frac{2\pi\cdot\bar{h}}{h}}.\end{aligned} \tag{8}$$

The PQI can be evaluated for every arbitrary pair $\{h, \bar{h}\} \in [1, H_l]$. For all further observations, the harmonic index $\bar{h}$ is referred to as PQI reference harmonic, while h denotes the harmonic index. Furthermore, the reference harmonic $\bar{h}$ is set to 2 and therefore does not change during the process.

3.3. Temporal Smoothing of PQI

From Eq. (8), the harmonic phase of an arbitrary harmonic $h \in [1, H_l]$ can be reformulated using PQI and the corresponding reference harmonic phase $\bar{h}$:

$$\Psi(h,l) = \frac{h \cdot \Psi(\bar{h},l)}{\bar{h}} - \Delta\Psi_{\bar{h}}(h,l). \tag{9}$$

The PQI reference phase $\Psi(\bar{h},l)$ is of high importance, as corruption with noise leads to erroneous results for the corresponding harmonic phases throughout the harmonics. Therefore it is recommended to pre-enhance the reference phase. For the following observations, we used TSUP [6] solely on the reference phase.

The PQI values are then calculated based on the pre-enhanced reference phases $\hat{\Psi}(\bar{h},l)$:

$$\Delta\hat{\Psi}_{\bar{h}}(h,l) = \frac{h}{\bar{h}}\left(\hat{\Psi}(\bar{h},l) - \frac{\Psi(h,l)\cdot\bar{h}}{h}\right)\Bigg|_{\frac{2\pi\cdot\bar{h}}{h}}. \tag{10}$$

The differential phases obtained from Eq. (10) are then smoothed across time, by mean averaging:

$$\Delta\tilde{\Psi}_{\bar{h}}(h,l) = \angle\frac{1}{|\mathcal{W}|}\sum_{\tilde{l}\in\mathcal{W}} e^{j\Delta\hat{\Psi}_{\bar{h}}(h,\tilde{l})}, \tag{11}$$

where $\mathcal{W}$ denotes the set of frames that lie within a range of 100 milliseconds around frame l. This filter length was chosen empirically, after analysing the PQI behaviour over time.

3.4. Synthesizing Phase-Enhanced Speech

The signal synthesis is based on [5], as the enhanced harmonic phase is transformed to the STFT domain by modifying the frequency bins within the main lobe width of the analysis window. We define $Y(k,l)$ as the DFT of the noisy signal with k as the corresponding frequency bin and K as the DFT length with $k \in [0, K-1]$. Further, $|Y(k,l)|$ denotes the noisy spectral amplitude and $\vartheta(k,l) = \angle Y(k,l)$ the noisy STFT phase. The enhanced STFT phase is then given by:

$$\begin{aligned}\hat{\vartheta}(\lfloor h\omega_0(l)K\rfloor + i, l) &= \left(\frac{h\cdot\hat{\Psi}(\bar{h},l)}{\bar{h}} - \Delta\tilde{\Psi}_{\bar{h}}(h,l)\right), \\ &\forall i \in [-N_p(l)/2, N_p(l)/2].\end{aligned} \tag{12}$$

where $N_p(l)$ denotes the minimum value of either the main lobe width of the analysis window N_w or frequencies close to neighboring harmonic $N_p(l) = \min(N_w, \omega_0(l)K/(2\pi))$. We obtain the phase enhanced signal in STFT domain by:

$$\hat{X}(k,l) = |Y(k,l)|e^{j\hat{\vartheta}(k,l)}. \tag{13}$$

The corresponding time domain signal $\hat{x}(n)$ is obtained by the inverse DFT of $\hat{X}(k,l)$ followed by the overlap-add procedure. Alternatively, the amplitude $|Y(k,l)|$ in Eq. (13) can be replaced by any enhanced amplitude $|\hat{X}(k,l)|$ available.

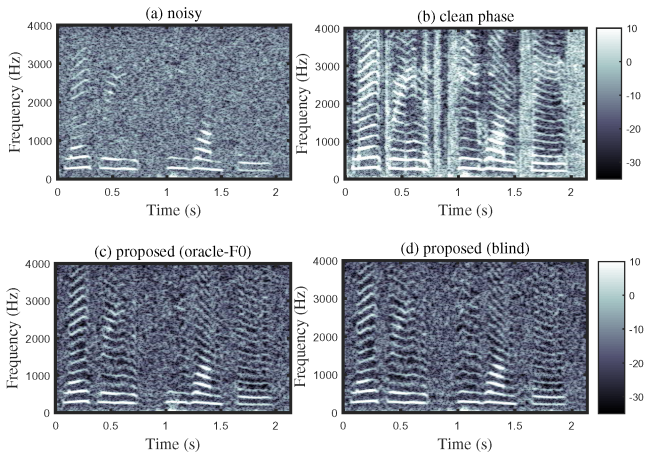

Fig. 3. Spectrogram of a female utterance with white noise at SNR $= 5$ dB. a: Noisy phase. b: Clean phase. c: Proposed method and oracle F_0. d: Proposed method blind.

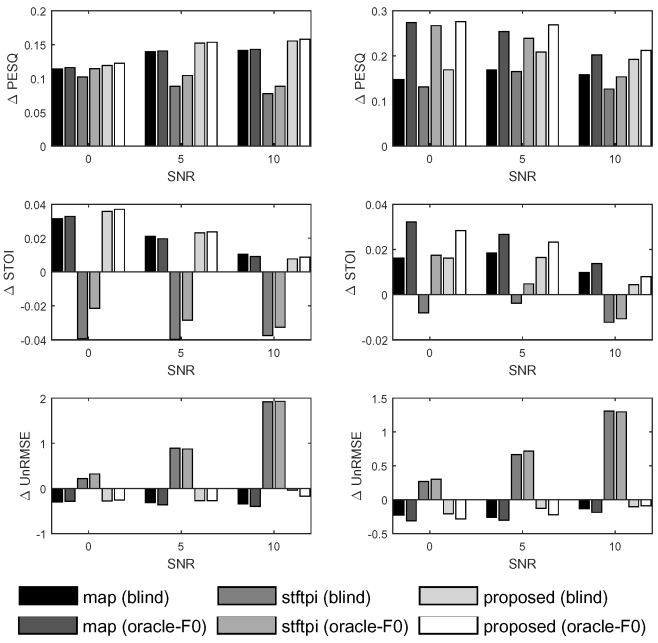

Fig. 4. PESQ improvement, STOI improvement and UnRMSE improvement in dB for (left) white and (right) babble noise.

4. RESULTS

4.1. Experiment Setup

We randomly chose 50 utterances spoken by 20 speakers (10 female and 10 male) from GRID [19] and mixed them with white and babble noise from NOISEX-92 [20] at SNRs between 0 to 10 dB in 5 dB steps. As evaluation criteria we chose perceptual evaluation of speech quality (PESQ) [21], short-term objective intelligibility measure (STOI) [22] and unwrapped root mean square estimation error (UnRMSE) [23] in decibels.

4.2. Speech Enhancement Results

Figure 3 shows the proof of concept experiment carried out on a female speech sample mixed with white noise at a global SNR $= 5$ dB. The phase-enhanced results using the proposed method illustrate an improved harmonic structure, closer to that observed in the clean phase. This harmonic structure is lost if only the noisy spectral phase is available.

The quantitative results are reported in Figure 4, namely the delta improvement compared to the noisy signal by means of (top) perceived quality, (middle) speech intelligibility, and (bottom) UnRMSE [23]. The scores are averaged over all utterances for both, white and babble noise. The reported results for noisy phase represent the lower-bound. As benchmarks, we include the performance of STFTPI [4] and MAP [5], both in combination with PEFAC [17] as noise-robust F_0-estimator.

In white noise, the proposed phase estimation method is not that sensitive to F_0 estimation accuracy. However, in the babble noise scenario the achievable performance by the phase enhancement methods is dependent on F_0 estimation accuracy. Overall, the proposed method improves the perceived quality, speech intelligibility and phase estimation error for all SNRs and noise types. This is an important finding since speech enhancement methods are often reported to degrade speech intelligibility or not being capable of improving perceived quality and intelligibility jointly.

In terms of speech quality, the proposed method outperforms all benchmark methods at all SNRs. In terms of the speech intelligibility, the proposed method impacts less at high SNRs. In terms of UnRMSE, the MAP estimate [5] is superior which is attributed to the fact that it relies on prior information about SNR, which is not taken into account in the proposed estimator. The results for PESQ enhancement are statistically significant for a significance level of 5%. For listening examples we refer to the accompanying webpage [24].

5. CONCLUSION

The paper proposed a new harmonic phase estimator for speech enhancement relying on relations between harmonics using the phase structure across harmonics. Temporal smoothing of the phase invariance representation allows for selective smoothing at harmonic level and contributes to improved speech quality. In this work the phase estimate is only used for signal reconstruction. Since an enhanced spectral phase was also reported to be useful in speech recognition [25] and separation [26] these applications are to be considered as future works.

6. REFERENCES

[1] P. Mowlaee, R. Saeidi, and Y. Stylianou, "Advances in phase-aware signal processing in speech communication," *Speech communication*, vol. 81, pp. 1–29, 2016.

[2] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase Processing for Single-Channel Speech Enhancement: History and recent advances," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 55–66, March 2015.

[3] P. Mowlaee, J. Kulmer, J. Stahl, and F. Mayer, "Phase-Aware Signal Processing in Speech Communication: History, Theory and Practice," *John Wiley & Sons*, 2016.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] T. Gerkmann, "Bayesian Estimation of Clean Speech Spectral Coefficients Given a Priori Knowledge of the Phase," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 62, no. 16, pp. 4199–4208, Aug. 2014.

[10] P. Mowlaee, J. Stahl, and J. Kulmer, "Iterative joint MAP single-channel speech enhancement given non-uniform phase prior," *Speech Communication*, vol. 86, pp. 85–96, 2017.

[11] V. A. Zverev, "Modulation method of ultrasonic dispersion measurements (in Russian)," *The Papers of the USSR Academy of Sciences*, vol. 91/4, pp. 791–794, 1953.

[12] V. I. Vorobiov, "Inter-component phase processing of speech signals for their recognition and identification of announcers," in *Proceedings of the 18th session of the Russian Acoustical Society*. Russian Acoustical Society, 2006, vol. 3, pp. 48–51.

[13] V. I. Vorobiov, G. V. Davydov, and Y. V. Shamgin, "Phase relation between fundamental tones and vowel sounds obertones (in Russian)," in *The reports of BSUIR*. Belarusian State University of Informatics and Radioelectronics, 2006, vol. 2/14, pp. 64–68.

[14] V. I. Vorobiov, "Inter-component phase processing of speech signals in time and frequency domains," in *Proceedings of the 19th session of the Russian Acoustical Society*. Russian Acoustical Society, 2007, vol. 3, pp. 46–49.

[15] I. Saratxaga, I. Hernaez, D. Erro, E. Navas, and J. Sanchez, "Simple representation of signal phase for harmonic speech models," *Electronics Letters*, vol. 45, no. 7, pp. 381–383, 2009.

[16] V. I. Vorobiov, A. G. Davydov, and S. Barysenka, "The interactive bispectrum calculation tool for MATLAB," *Software, available [Sep. 2016] from https://github.com/agdavydov81/bispectrum/*, 2016.

[17] S. Gonzalez and M. Brookes, "PEFAC - A pitch estimation algorithm robust to high levels of noise," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 22, no. 2, pp. 518–530, 2014.

[18] G. Degottex and D. Erro, "A measure of phase randomness for the harmonic model in speech synthesis.," in *Proc. Interspeech*, 2014, pp. 1638–1642.

[19] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, November 2006.

[20] 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," Tech. Rep., DRA Speech Research Unit, 1992.

[21] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs.," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2001, pp. 749–752.

[22] 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 & Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[23] A. Gaich and P. Mowlaee, "On Speech Intelligibility Estimation of Phase-Aware Signal Processing for Automatic Speech Recognition," *Proc. Interspeech*, pp. 2553–2557, 2016.

[24] M. Pirolt, J. Stahl, P. Mowlaee, V. I. Vorobiov, S. Y. Barysenka, and A. G. Davydov, "Phase Estimation in Single-channel Speech Enhancement Using Phase Invariance Constraints: supporting webpage with some audio examples. [Online]. Available: `http://www2.spsc.tugraz.at/people/pmowlaee/PQI`," Sept. 2016.

[25] J. Fahringer, T. Schrank, J. Stahl, P. Mowlaee, and F. Pernkopf, "Phase-Aware Signal Processing for Automatic Speech Recognition," in *Proc. Interspeech*, 2016, pp. 3374–3378.

[26] F. Mayer and P. Mowlaee, "Improved phase reconstruction in single-channel speech separation," in *Proc. Interspeech*, 2015, pp. 1795–1799.