Phase Estimation in Single-Channel Speech Enhancement Using Phase Invariance Constraint

MICHAEL PIROLT[†], JOHANNES STAHL[†], PEJMAN MOWLAEE[†], VASILI I. VOROBIOV[‡], SIARHEI Y. BARYSENKA[‡], ANDREW G. DAVYDOV[‡]

(PRESENTED BY GERNOT KUBIN)

[†]Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria

[‡] Belarusian State University of Informatics and Radioelectronics, Minsk, Belarus

Abstract

- Phase-aware speech processing: **fundamentals-applications** [1,8].
- Is phase processing **possible**? [2,4]
- Existing methods: Temporal Smoothing of Unwrapped Phase (TSUP) [2] and Maximum a Posteriori (MAP) [6], STFT phase improvement [4]
- Phase-aware processing in **speech enhancement** (see [1, Ch. 3])
- **Here: phase invariance property to estimate clean spectral phase**

1. Notations and Signal Model

For each frame index l, the noisy speech is given by:

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

$\nu(n,l)$: noise at frame l and time n
H_l : number of harmonics at frame l
$F_0(l)$: fundamental frequency,
$Y(k,l)$: noisy DFT spectrum
$\vartheta(k,l) = \angle Y(k,l)$: noisy DFT phase
K : DFT length with $k \in [0, K-1]$
N_w : analysis window main lobe width

h : harmonic index with $h \in [1, H_l]$
$\Phi(h,l)$: unwrapped phase
f_s : sampling frequency
$\hat{X}(k,l)$: phase-enhanced spectrum
$\hat{\vartheta}(k,l)$: enhanced DFT phase
l : frame index
$\hat{x}(n)$: phase-enhanced signal

2. Phase Invariant (PI) and Phase Quasi Invariant (PQI) Properties

Phase Invariant (PI): for a harmonic signal, Zverev [9] proposed *phase invariant* (PI) property, determined for each triplet of the harmonic components if their frequencies satisfy the following set of equations:

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

Then the PI denoted by $\Delta\Psi(l)$, is given by:

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

Phase Quasi-Invariant (PQI): was introduced by Vorobiov [5], between components with frequencies $\bar{h}F_0(t)$ and $hF_0(t)$. For an arbitrary pair $\{h,\bar{h}\} \in [1, H_l]$ we obtain:

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

3. Temporal Smoothing of PQI

Given pre-enhanced reference phases $\hat{\Psi}(\bar{h},l)$, PQI values are given by:

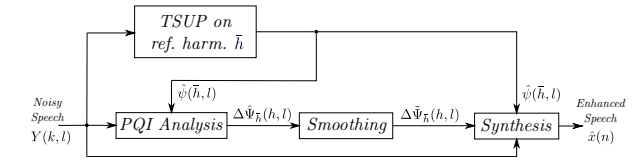
$$\Delta\hat{\Psi}_{\bar{h}}(h,l) = \frac{h}{\bar{h}}\left(\hat{\Psi}(\bar{h},l) - \frac{\Psi(h,l)\cdot\bar{h}}{h}\right)\Bigg|_{\frac{2\pi\cdot\bar{h}}{h}}.$$

We temporally smooth the PQI as follows:

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

$\mathcal{W}$: the set of frames that lie within a range of 100 milliseconds around frame l.

4. Proposed Phase Estimation Method



5. Signal Synthesis

The enhanced harmonic phase is transformed to the STFT domain by modifying the frequency bins within the main lobe width of the analysis window:

$$\hat{\vartheta}(\lfloor h\omega_0(l)K\rfloor + i, l) = \underbrace{\left(\frac{h\cdot\hat{\Psi}(\bar{h},l)}{\bar{h}} - \Delta\tilde{\Psi}_{\bar{h}}(h,l)\right)}_{\hat{\psi}(\bar{h},l)}$$

where $\forall i \in [-N_p(l)/2, N_p(l)/2]$, $N_p(l)$ minimum value of either 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:

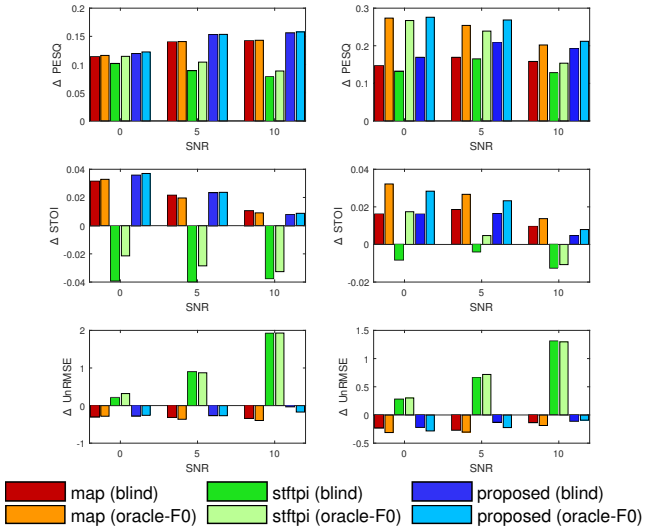
$$\hat{X}(k,l) = |Y(k,l)|e^{j\hat{\vartheta}(k,l)},$$

$\hat{x}(n)$ is given by applying the inverse DFT of $\hat{X}(k,l)$ followed by overlap-add.

6. Experiment Setup

- Speech: GRID corpus, 50 utterances (specakers: 10 female, 10 male)
- Noise: NOISEX-92: white and babble, mixed at SNR $\in \{0, 5, 10\}$ (dB)
- Methods: MAP [6], STFT phase improvement (STFTPI) [4], proposed.
- Robustness to F_0 estimation errors: oracle-F_0 vs. blind scenario
- Evaluation criteria: ΔPESQ, ΔSTOI, ΔUnRMSE in (dB) [7].
- In PESQ, the proposed method outperforms others.
- In terms of phase estimation error (UnRMSE), MAP [6] is the best.
- Listening examples at: `www2.spsc.tugraz.at/people/pmowlaee/PQI`

7. Experimental Results



8. References

[1] P. Mowlaee, J. Kulmer, J. Stahl, and F. Mayer, "Phase-Aware Signal Processing in Speech Communication: History, Theory and Practice," John Wiley & Sons, 2016.
[2] P. Mowlaee and J. Kulmer, "Phase Estimation in Single-Channel Speech Enhancement: Limits-Potential," TASL, 23(8), pp. 1283-1294, 2015.
[3] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase Processing for Single-Channel Speech Enhancement: History and recent advances," IEEE Signal Processing Magazine, 32(2), pp. 55-66, 2015.
[4] M. Krawczyk and T. Gerkmann, "STFT Phase Reconstruction in Voiced Speech for an Improved Single-Channel Speech Enhancement," TASL, 22(12), pp. 1931-1940, 2014.
[5] V. I. Vorobiov, "Inter-component phase processing of speech signals for their recognition and identification of announcers," in Proc. Russian Acoustical Society. pp. 48-51, 2006.
[6] J. Kulmer and P. Mowlaee, "Harmonic phase estimation in single-channel speech enhancement using von mises distribution and prior snr," ICASSP, pp. 5063-5067, 2015.
[7] A. Gaich and P. Mowlaee, "On Speech Intelligibility Estimation of Phase-Aware Single-Channel Speech Enhancement", INTERSPEECH, pp. 2553-2557, 2015.
[8] www.spsc.tugraz.at/PhaseLab
[9] V. A. Zverev, "Modulation method of ultrasonic dispersion measurements (in Russian)," USSR Academy of Sciences, 91(4), pp. 791-794, 1953.

TU Graz Graz University of Technology

Signal Processing & Speech Communication Laboratory
Graz University of Technology
Inffeldgasse 16c, A-8010 Graz, AUT
http://www.spsc.tugraz.at

This work was supported by the Austrian Science Fund (project number P28070-N33).
FWF Der Wissenschaftsfonds.