On the Importance of Harmonic Phase Modification for Improved Speech Signal Reconstruction

ANNA MALY[1], PEJMAN MOWLAEE[2]

1) Intelligent Acoustic Solutions, JOANNEUM RESEARCH, Austria

2) Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria

Abstract

- Phase importance in single-channel speech enhancement
- 1) Amplitude Estimation
- **2) Signal Reconstruction** [1]
- The current study addresses two questions:
- 1) **STFT** or **harmonic** phase?
- 2) Harmonic Phase: **Unwrapped phase** versus **linear phase**?

1. Notations

- $y(n), x(n)$ . . . noisy and clean speech at time index n
- $Y(k,l), X(k,l)$ . . . noisy and clean speech spectra, at frequency index k and time frame index l
- $\phi_y(k,l), \phi_x(k,l)$ . . . noisy and clean spectral phase
- $\hat{\phi}_x(k,l)$ . . . estimated clean spectral phase
- $\hat{x}_p(n), \hat{X}_p(k,l)$ . . . phase-enhanced signal and spectrum
- $G(k,l)$ . . . amplitude enhancement gain function
- $|\hat{X}(k,l)|$ . . . amplitude-enhanced speech spectrum

2. Harmonic Model Plus Phase Decomposition

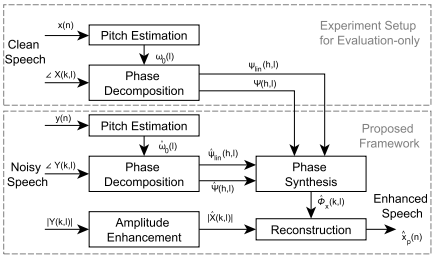
- Speech segments are modelled as sum of harmonics.
- Harmonic phase decomspotion results in *linear* and *unwrapped* phase parts:

$$\psi(h,l) = \underbrace{h \sum_{l'=0}^{l} \omega_0(l')\big(t(l') - t(l'-1)\big)}_{\text{Linear phase: } \psi_{\text{lin}}(h,l)} + \underbrace{\angle V(hf_0(l),l) + \psi_d(h,l)}_{\text{Unwrapped phase: } \Psi(h,l)},$$

where $\psi_{\text{lin}}(h,l)$ is the linear phase and $\Psi(h,l)$ is the unwrapped phase at harmonic h and time frame l.

- Unwrapped phase consists of minimum phase (vocal tract) $\psi_{\text{min}}(h,l) = \angle V(hf_0(l),l)$ and phase dispersion $\psi_d(h,l)$.
- Unwrapped phase estimate is given by subtracting linear phase from $\psi(h,l)$.

3. Proposed Framework



4. Experiment Setup

- 50 GRID speech files corrupted with babble noise at -5 to 15 dB SNR.
- **Fundamental frequency** estimation using PEFAC.
- **Harmonic model phase decomposition** from COVAREP.
- Combination with **conventional amplitude enhancement** MMSE-LSA

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

where $G(k,l)$ is the gain function tabularized with *prior* and *posterior* SNRs.

- Using overlap-add on the phase enhanced spectrum gives

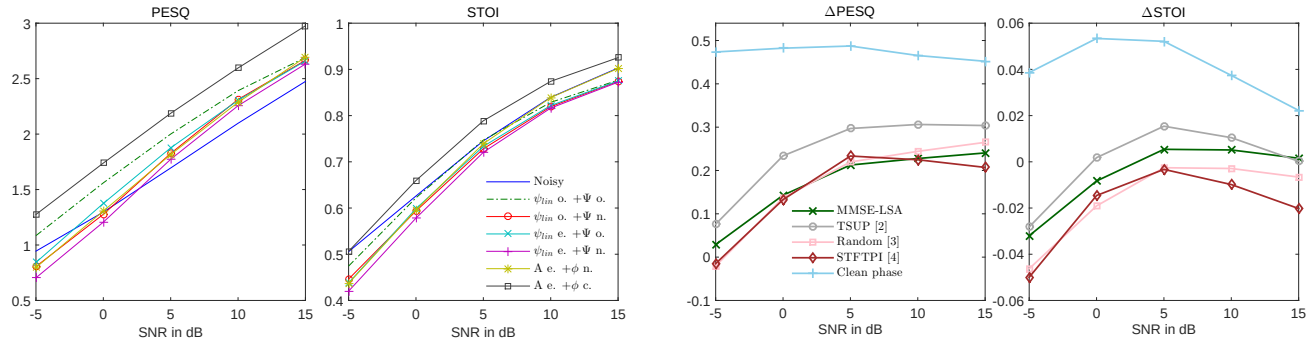
$$\hat{X}_p(k,l) = |\hat{X}(k,l)|e^{j\hat{\phi}_x(k,l)}.$$

The phase-enhanced signal $\hat{x}_p(n)$ is given by applying overlap-add routine.

5. Comparison Category Rating (CCR)-Test

Comparison: *oracle linear phase with oracle unwrapped phase* versus *phase-enhanced signal using clean STFT phase*. The results are rated within $[-3, 3]$.
Result: on average the oracle unwrapped phase achieves -0.8 versus clean STFT phase with a confidence interval of ± 0.1.

6. Importance of Harmonic Phase Parts (left) and Comparisons with Other Phase Estimators (right)



7. Observations

- CCR Test revealed that a proper modification of harmonic phase and smoothed unwrapped phase provides **similar perceptual quality versus clean phase**.
- Proof-of-concept experiment showed that a **proper linear phase together with a successful modification of the unwrapped phase contribute the most** to an improved signal reconstruction. Audio examples available at *http://www2.spsc.tugraz.at/people/pmowlaee/ICASSP2016.html*
- The **importance of joint modification of linear and unwrapped phase** was validated.

8. References

[1] P. Mowlaee, J. Kulmer, "Phase estimation in single-channel speech enhancement: Limits-potential", IEEE Trans. Audio, Speech , and Language Proc., vol. 23, no. 8, pp. 1283-1294, Aug 2015
[2] J. Kulmer, P. Mowlaee,"Phase Estimation in Single Channel Speech Enhancement Using Phase Decomposition", IEEE Signal Processing Letters, vol. 22, no. 5, pp. 598-602, May 2015.
[3] A. Sugiyama, 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, 2013
[4] M. Krawczyk, T. Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement", IEEE Trans. Audio, Speech, Language Process., vol. 22, no. 12, pp. 1931-1940, Dec 2014

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).