TIME-FREQUENCY CONSTRAINTS FOR PHASE ESTIMATION IN SINGLE-CHANNEL SPEECH ENHANCEMENT

Pejman Mowlaee[†], *Rahim Saeidi*[‡]

[†]Signal Processing and Speech Communication Lab, Graz University of Technology, Austria
[‡]Speech and Image Processing Unit, School of Computing, University of Eastern Finland, Finland

ABSTRACT

Previous single-channel speech enhancement algorithms often employ noisy phase while reconstructing the enhanced signal. In this paper, we propose novel phase estimation methods by employing several temporal and spectral constraints imposed on the phase spectrum of speech signal. We pose the phase estimation problem as estimating the unknown clean speech phase at sinusoids observed in additive noise. To resolve the ambiguity in phase estimation problem, we introduce individual time-frequency constraints: group delay deviation, instantaneous frequency deviation, and relative phase shift. Through extensive simulations, the effectiveness of the proposed phase estimation methods in single-channel speech enhancement is demonstrated. Employing the estimated phase for signal reconstruction in medium-to-high SNRs leads to consistent improvement in perceived quality compared to when noisy phase is used.

***Index Terms*—** Phase estimation, single-channel speech enhancement, time-frequency constraints, perceived speech quality.

1. INTRODUCTION

Enhancement of speech signals observed in background noise is of great importance for the sake of robustness of different speech applications including: automatic speech recognition, mobile telephony and hearing aids. Much effort has been dedicated to derive optimal estimators for frequency and amplitude spectrum of desired signal [1]. The use of phase information in speech signal processing has been a controversial topic. In previous studies [2], the phase information has been considered of little importance in terms of its impact on the perceived signal quality within amplitude estimation and signal reconstruction shown in Figure 1. On the other hand, recent studies presented the importance of phase information in human speech perception [3], speech enhancement and separation [4–11]. The issue of estimating clean phase spectrum and its impact on the ultimate achievable performance is not adequately addressed yet.

While the choice of noisy phase at high enough SNR signal components is not critical and was shown to provide the MMSE estimation of clean phase [12], the choice of noisy phase spectrum for all signal components in signal reconstruction has been well observed to introduce certain distortions like musical noise as reported in [7, 13]. The MMSE estimation of phase spectrum is based on the independence assumption for all time-frequency discrete Fourier transform (DFT) coefficients which is not the case for speech signals. Therefore, a proper phase estimation method, mainly replacing noisy phase at signal components of low or moderate SNR level has the potential to improve the perceived speech quality.

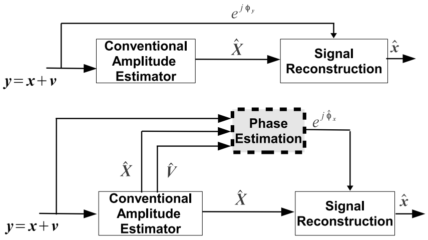

Fig. 1. (Top) Block diagram for typical single-channel speech enhancement composed of two stages: (1) amplitude spectrum estimation, and (2) signal reconstruction, (bottom) proposed phase estimation algorithm.

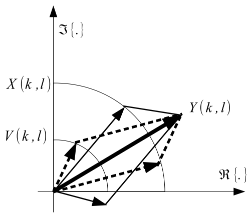

Fig. 2. The ambiguity in phase values for underlying sources results in two different ways on building the noisy observation $Y(k,l)$.

Considering a vector sum of speech and noise shown in Figure 2, at every time-frequency cell, there are two set of phase values which satisfy the problem geometry. To deal with ambiguity, in [11] we proposed a group-delay based phase estimation in the context of source separation setup where enhanced amplitude of speech and estimated noise were used. We showed in [11] that even by employing oracle amplitudes for underlying sources, the phase ambiguity causes a big drop in perceived speech quality.

In this paper, we introduce new constraints by employing instantaneous frequency deviation [14] and relative phase shift (RPS) [15] concepts from speech coding and speech synthesis fields and assemble them as metrics to handle the ambiguity in geometry-based phase estimation. The estimated phase is evaluated in signal reconstruction stage and in phase-aware amplitude estimator [7, 8]. The rest of the paper is organized as follow; Section 2 presents the problem formulation and conventional speech enhancement. Section 3 presents the proposed phase estimation methods, section 4 presents the results and section 5 concludes the work.

2. SPEECH ENHANCEMENT PROBLEM FORMULATION AND CONVENTIONAL SPEECH ENHANCEMENT

Let $x(n)$ and $v(n)$ be speech and noise signals, respectively, and let $y(n) = x(n) + v(n)$ as their noisy observation in discrete time

This project has received funding from the European Unions Seventh Framework Programme for research, technological development and demonstration under grant agreement no FP7-ICT-2011-7-288121.

978-1-4799-6808-4/14/$31.00 ©2014 IEEE

domain, with n as time index. Taking Fourier transformation, we further define $Y^c(k,l) = Y(k,l)e^{j\phi_y(k,l)}$ as the complex Fourier representation of the noisy signal defined for the kth frequency bin and the lth frame with $Y(k,l)$ and $\phi_y(k,l)$ as the noisy spectral amplitude and phase spectrum, respectively. Similarly, we define $X^c(k,l) = X(k,l)e^{j\phi_x(k,l)}$ and $V^c(k,l) = V(k,l)e^{j\phi_v(k,l)}$ as the complex spectrum for speech and noise, with $X(k,l)$ and $V(k,l)$ as the spectral amplitudes for speech and noise, respectively. For the observed noisy signal we have:

$$Y(k,l)e^{j\phi_y(k,l)} = X(k,l)e^{j\phi_x(k,l)} + V(k,l)e^{j\phi_v(k,l)}. \quad (1)$$

The spectral amplitude of the noisy signal is the absolute value of the vector sum of the underlying components and we have

$$Y(k,l) = \sqrt{X^2(k,l) + V^2(k,l) + 2X(k,l)V(k,l)\cos\Delta\phi_{k,l}}, \quad (2)$$

where we define $\Delta\phi_{k,l} = \phi_x(k,l) - \phi_v(k,l)$. It is obvious that $\pm\Delta\phi_{k,l}$ are both valid solutions for (2). This ambiguity in the sign is because of the lack of knowledge about the sign of $\sin\Delta\phi_{k,l}$. The observed noisy phase is given by:

$$\phi_y(k,l) = \pm m\pi + \tan^{-1}\frac{X(k,l)\sin\phi_x(k,l) + V(k,l)\sin\phi_v(k,l)}{X(k,l)\cos\phi_x(k,l) + V(k,l)\cos\phi_v(k,l)}, \quad (3)$$

where m is an integer number. Clearly, even given the oracle spectral amplitude of speech and noise, equation (3) is one equation with two unknowns, i.e., $\phi_x(k,l)$ and $\phi_v(k,l)$ as speech and noise phase.

Given the noisy signal, the conventional methods are focused on obtaining the MMSE estimation for the spectral amplitude. This is found as a parametric estimator in [16] and expressed in the form of a softmask function $G(k,l)$ multiplied to the observed signal as $\hat{X}(k,l) = G(\xi(k,l),\zeta(k,l))Y(k,l)$ where $\xi(k,l)$ and $\zeta(k,l) = Y^2(k,l)/P_v(k,l)$ are defined as the *a priori* and the *a posteriori* signal-to-noise ratios (SNRs), respectively, with $P_v = E\{V^2(k,l)\}$ as the noise power. In this work, as the baseline enhancement method, we choose the MMSE-LSA enhanced amplitude spectrum given by [17]: $\hat{X}(k,l) = G^{LSA}(\xi(k,l),\zeta(k,l))Y(k,l)$, where

$$G^{LSA}(\xi(k,l),\zeta(k,l)) = \frac{\xi(k,l)}{1+\xi(k,l)}\exp\left(\frac{1}{2}\int_{\nu(k,l)}^{\infty}\frac{e^{-t}}{t}dt\right), \quad (4)$$

and $\nu(k,l) = \frac{\zeta(k,l)\xi(k,l)}{1+\xi(k,l)}$. The noisy phase $\phi_y(k,l)$ is used to reconstruct the enhanced time-domain signal at frame l calculated as

$$\hat{x}_l(n) = \mathscr{F}^{-1}\{\hat{X}(k,l)e^{j\phi_y(k,l)}\}, \quad (5)$$

where $\mathscr{F}^{-1}(\cdot)$ is the inverse short-time Fourier transformation. Finally, overlap-and-add method [18] is applied to $\hat{x}_l(n)$ at all frames to reconstruct the enhanced speech signal $\hat{x}(n)$.

3. PROPOSED PHASE ESTIMATION METHODS

3.1. Geometry-based Phase Estimation Approach

We define $\hat{\phi}_x^a(k,l)$ and $\hat{\phi}_v^a(k,l)$ as the ambiguous phase set estimates for speech and noise sources defined for the kth frequency bin at the lth time-frame [11]. The ambiguity in the trigonometric functions results in four candidates for $\{\cos\phi_v(k,l), \sin\phi_x(k,l)\}$, and two candidates for $\{\cos\phi_x(k,l), \sin\phi_v(k,l)\}$ and two candidates for $\pm\Delta\phi(k,l)$ [11]. From Figure 2, it is obvious that at each time-frequency cell (k,l), there are two phase sets of the sources $\hat{\phi}_x^{(a)}(k,l) = \{\hat{\phi}_x^{(1)}(k,l), \hat{\phi}_x^{(2)}(k,l)\}$ and $\hat{\phi}_v^{(a)}(k,l) = \{\hat{\phi}_v^{(1)}(k,l), \hat{\phi}_v^{(2)}(k,l)\}$, for speech and noise signals, respectively, which both satisfy all observations regarding the noisy complex spectrum and the spectral amplitude of the underlying signals. The two sets of phase candidate only differ in their resulting sign in $\Delta\phi$. We impose the minimum reconstruction error criterion, in order to find the best pair of ambiguous phase values at the current time-frequency cell, defined as below:

$$\begin{aligned} e(k,l) &= |Y^c(k,l) - \hat{Y}(k,l)e^{j\hat{\phi}_y(k,l)}|, & (6)\\ \hat{Y}(k,l)e^{j\hat{\phi}_y(k,l)} &= X(k,l)e^{j\hat{\phi}_x(k,l)} + V(k,l)e^{j\hat{\phi}_v(k,l)}. & (7)\end{aligned}$$

3.2. Phase Estimation at Sinusoids

It is already well observed in [19] that for the spectral components of high SNR (SNR > 6(dB)), the choice of noisy phase is a reasonable estimation of clean speech phase. On the other hand, for spectral components with SNR lower than 6 decibel, the phase deviation gets larger than the threshold of speech perception [19]. However, in practice the estimation of local SNR for every time frequency bin is rather unreliable due to errors in noise estimator. Furthermore, the redundant STFT representation introduce many signal components with low amplitude level which gets easily masked by noise. To mitigate these, here we are focused on enhancing the phase of those signal components deteriorated by noise but contributing the most in representing the underlying speech signal. Hence, in this work we choose only the frequency components that show high amplitude spectrum (spectral peaks) as a representative for high energy components and perform phase estimation on them. The spectral peaks are supposedly arising from medium-to-high strong signal components.

To detect the spectral peaks we can either apply peakpicking or fit a sinusoidal model to the enhanced speech amplitude spectrum with a relatively low model order. For the sake of the simplicity and to avoid the erroneous model order selection in sinusoidal model, in the following, we apply the proposed phase estimation methods only to the spectral peaks found by a simple peak picking [20]. The frequency of the p-th sinusoidal peak is denoted by $\{k_p\}_{p=1}^{P_l}$ with P_l as the number of peaks detected at frame l whose value varies across frames, and we further define $\hat{X}(k_p,l)$ as the amplitude of sinusoids for the pth peak selected by peakpicking. Figure 3 graphically represents each of the proposed individual constraints across time and frequency for a real speech signal.

3.3. Instantaneous Frequency Deviation Constraint

Instantaneous frequency (IF) is defined as the first time-derivative of the phase spectrum [21]. For the p-th harmonic component at frame l and assuming a hop size of H samples between consecutive frames, the instantaneous frequency estimate $\hat{\omega}_x^{\Delta}(k_p,l)$ is given as [**?**, 22]:

$$\hat{\omega}_x^{\Delta}(k_p,l) = \frac{\hat{\phi}_x(k_p,l) - \hat{\phi}_x(k_p,l-1)}{2\pi H}. \quad (8)$$

We approximate the IF value by $\hat{\omega}_x^{\Delta}(k_p,l) \approx k_p/N_{\text{DFT}}$ with N_{DFT} defined as number of DFT points and we obtain an IF-based phase estimate given by

$$\hat{\phi}_x^{\text{IFD}}(k_p,l) = \frac{2\pi H k_p}{N_{\text{DFT}}} + \hat{\phi}_x^{\text{IFD}}(k_p,l-1), \quad (9)$$

An estimation for the current frame phase value $\hat{\phi}_x^{\text{IFD}}(k_p,l)$ is obtained based upon the phase value of the previous frame $\hat{\phi}_x(k_p,l-1)$

and under the assumption of having a stationary enough instantaneous frequency (e.g. at smooth trajectories with no abrupt changes) within the time interval of the harmonic trajectory under consideration. In order to remove the ambiguity in the two phase candidates, we rely on the fact that the IF-based phase estimate of the noisy signal denoted by $\hat{\phi}_y^{\text{IFD}}(k_p, l)$ still exhibits similarity with that of the clean signal, so can be used as a reference point to define *distortion metric* as the time-derivative constraint

$$d_{\text{IFD}} = 1 - \cos(\hat{\text{IFD}}_y^{\text{IFD}}(k_p, l) - \hat{\phi}_x^{\text{IFD}}(k_p, l)). \quad (10)$$

The rationale behind employing the cosine operator in the definition of the metric is to make it invariant to modulo of 2π and eventually to avoid the wrong error calculation due to the periodicity of phase components. Similar treatment was employed for phase-based estimators studied in [22]. The phase distortion metric of type $d_\phi(\hat{\phi}(k_p, l), \phi(k_p, l)) = 1 - \cos(\hat{\phi}(k_p, l) - \phi(k_p, l))$ was also used in [23] for small estimation errors it is well resembling the squared-error distortion measure. Finally, the optimal phase values at each frame at pth spectral peak denoted by k_p is given by drawing all combinations from $\hat{\phi}_x^a(k_p, l)$ and is given by

$$\hat{\phi}_x^*(k_p, l) = \underset{\hat{\phi}_x^a(k_p, l)}{\arg\min}\, d_{\text{IFD}}. \quad (11)$$

3.4. Relative Phase Shift Constraint

We employ the relative phase shift (RPS) representation of phase recently introduced in [15] where the authors justified the perceptual importance of the phase related information in speech signals that allowed the direct analysis of phase structure in analysis, modification and synthesis. The RPS relates the instantaneous phase of the fundamental frequency component and the instantaneous phase value at the pth harmonic [15] as: $\hat{\phi}_x^{\text{RPS}}(k_p, l) = p\phi_x(k_1, l)$ where $\phi_x(k_1, l)$ refers to the instantaneous phase of the fundamental frequency component. Here, we approximate the fundamental frequency as the first peak frequency denoted by k_1 estimated via fitting the sinusoidal model to signal, and k_p referring to the frequency of the pth sinusoid. For the initialization of RPS constraint, we set $\phi_x(k_1, l)$ equal to the phase of the sinusoidal peak estimated from noisy observation, as it is a dominant peak and less deteriorated by noise contribution.

In order to attain minimum relative phase shift distortion, we define the following distortion metric

$$d_{\text{RPS}} = 1 - \cos(\hat{\phi}_x(k_p, l) - \hat{\phi}_x^{\text{RPS}}(k_p, l)). \quad (12)$$

Then the optimal phase value at the k_pth frequency bin is:

$$\hat{\phi}_x^*(k_p, l) = \underset{\hat{\phi}_x^a(k_p, l)}{\arg\min}\, d_{\text{RPS}}. \quad (13)$$

3.5. Group Delay Deviation

Group delay is defined as the first frequency-derivative of the phase spectrum [24]:

$$\tau_x(k, l) = -\Delta_k\{\phi_x(k, l)\}, \quad (14)$$

where Δ_k is frequency derivative operator in discrete domain. Assuming a short-time Fourier analysis, for a rectangular window type of finite support of length N as $w(l) = 1$ for $l \in [0, N-1]$, its Fourier transform $W(e^{j\omega}) = \frac{1}{N}\frac{\sin(\frac{N\omega}{2})}{\sin(\frac{\omega}{2})}e^{-j\omega(\frac{N-1}{2})}$ comprises of only a linear phase term. The group delay for the linear phase $\phi(\omega) = -\omega(\frac{N-1}{2})$, will be a constant value of $\tau_w = \frac{N-1}{2}$. In [25], *group delay deviation* was defined as the deviation in group delay of $\tau_x(k, l)$ with respect to τ_w, given as below:

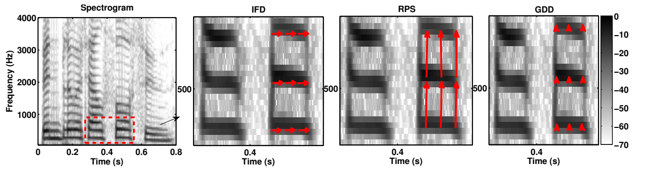

Fig. 3. From left to right, showing how different constraints function on the phase spectrum across time and frequency. The red arrows show the coordination to which the proposed constraints are applied on the phase spectrum.

$$\Delta\tau_x(k, l) = \tau_w - \tau_x(k, l). \quad (15)$$

The group delay deviation is well observed to exhibit minima at spectral peaks [25]. Using this constraint along with the geometry, in [11], we presented phase estimation solutions for single-channel source separation problem in [11]. The minimum group delay deviation constraint around harmonic peaks helped to select the correct phase candidate which was unknown due to the ambiguity in the sign difference between the two spectra. We define the group delay deviation-based distance metric as:

$$d_{\text{GDD}} = 1 - \cos(\tau_w - (\hat{\phi}_x(k_p, l) - \hat{\phi}_x(k_p + 1, l))). \quad (16)$$

We employ d_{GDD} to remove the ambiguity in phase candidates, and the optimal phase value for frequency k_p is given by

$$\hat{\phi}_x^*(k_p, l) = \underset{\hat{\phi}_x^a(k_p, l)}{\arg\min}\, d_{\text{GDD}}. \quad (17)$$

3.6. Utilization of the Proposed Metrics

We confine the proposed time-spectral metrics to be applied only at spectral peaks with normalized magnitude of -30 (dB) and above as the spectral peaks with the magnitude lower than -30 (dB) do not contribute to the perceived signal quality and are most likely originated by a noise-like component. The proposed constraints used in the phase estimation methods require the phase at some reference point to rely on in order to calculate the phase of the next time or frequency cell. The phase estimation procedure is as follow; The IFD constraint functions on the same frequency bin across two consecutive frames, RPS is applied across phase of harmonic multiples with respect to the fundamental frequency phase calculated within the same frame, and GDD is applied on the phase values at frequencies in the vicinity of the peak i.e. k_p and $k_p + 1$ at the same time frame. For all metrics, combinations of the phase candidates in the Ambiguous phase candidate set are examined and one with the minimum distortion is chosen.

4. RESULTS

We extract fifty sentences from GRID corpus [26] including 18 male and 16 female speakers. The noisy speech signals are produced by mixing speech with white and babble noise selected from NOISEX-92 database [27]. As our performance evaluation criterion, we employ PESQ measure. The results reported here are averaged over fifty utterances and are swept over a range of SNR from -5 to 20 *(dB)*. The audio material is sampled at 8 kHz. We use a hamming window length of $N = 32$ ms, with $H = 4$ ms of frame shift in processing the speech signal. To initialize the noise tracker we use the first ten frames as noise-only frames.

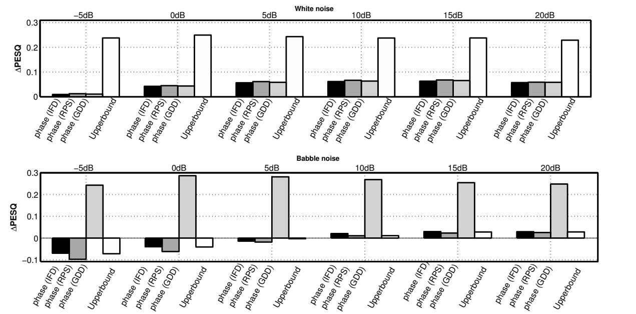

Fig. 4. $\Delta PESQ$ results averaged over 50 utterances obtained by the proposed phase estimation methods compared to others for blind scenario for (top) white, and (bottom) babble noise scenarios.

4.1. Phase Estimation for Signal Reconstruction

We evaluate the effectiveness of the proposed phase estimation methods. Figure 4 shows the PESQ results obtained by the proposed phase estimation methods for blind scenario where both speech and noise spectra as well as phase are estimated. For a clear comparison, we further report the $\Delta PESQ$ obtained by the phase estimation methods compared to the conventional speech enhancement using noisy phase using

$$\Delta PESQ = PESQ_{\text{MMSE-LSA + proposed phase}} - PESQ_{\text{MMSE-LSA + noisy phase}}.$$

The proposed methods lead to a consistent improvement in PESQ, in particular for mid to high SNR for both white and babble noise. The level of improvement in PESQ is slightly larger in white noise compared to babble noise. For white noise scenario, the proposed phase estimation methods bring an average PESQ improvement of 0.05 compared to noisy phase for medium to high SNRs (SNR ≥ 5 (dB)). For babble noise, the improvement for SNR > 5 (dB) in PESQ are rather negligible.

All the phase estimation methods proposed here rely on the correctness of the problem geometry shown in Figure 2. As soon as the geometry is distorted due to erroneous speech and noise estimates, the estimated phase candidates will be less accurate. Furthermore, due to over/under-estimation of signal-to-noise ratio, the selected peaks might not be correctly chosen, and misclassified by selecting the noise signal component rather than speech spectral peak. This will lead to degradation in performance with wrong phase assignment in all methods. From noise known scenario (not shown here) it was observed that the success of the phase estimator is highly dependent on the performance of the noise estimation.

4.2. Phase Estimation for Amplitude Estimation

In the conventional MMSE amplitude estimation [12, 17], the speech phase information is neglected originally due to the fact that the circularly symmetric speech prior distribution is exploited in the derivation of Wiener filter. If an estimation of the clean phase spectrum is available, a phase-aware MMSE amplitude estimator, as recently was shown in [7, 9] can be used. In this section, we justify the effectiveness of the proposed phase estimation methods in terms of improving the estimates for amplitude spectrum and eventually enhancing the estimated complex spectrum. For this purpose, we employ the phase estimation using the GDD metric in the structure of the iterative phase and amplitude estimator [8]. The estimated amplitude is then used together with the estimated phase to reconstruct the enhanced signal. As lower bound and upper bound we include the results of unprocessed and phase-aware amplitude given oracle phase, respectively[1] . Figure 5 shows the $\Delta PESQ$ results categorized to white (top panel) and babble (bottom panel) noise scenarios. In the evaluation of this section we calculate $\Delta PESQ$ as the following:

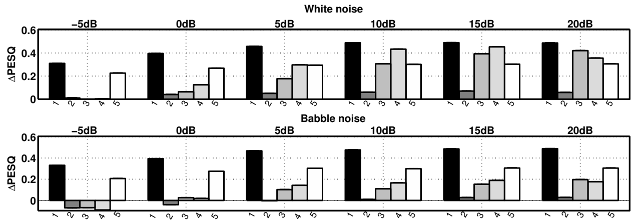

Fig. 5. Results in $\Delta PESQ$ shown for blind scenario for white (top) and babble (bottom) obtained by the following methods: 1) amplitude (phase-aware) + phase (oracle), 2) amplitude (MMSE-LSA) + phase (GDD), 3) amplitude (phase-aware) + phase (GDD), 4) iterative closed-loop phase-aware [8], 5) amplitude (MMSE-LSA) + phase (oracle).

$$\Delta PESQ = PESQ_{\text{Enhanced Complex Spectrum}} - PESQ_{\text{MMSE-LSA + noisy phase}}.$$

Incorporating the estimated phase in the phase-aware amplitude estimator results in over 0.1 improvement in PESQ compared to conventional phase-unaware amplitude estimator (MMSE-LSA) for SNR ≥ 5 (dB). For white noise case, the combination surpasses the perceived quality obtained by the upper-bound for the conventional speech enhancement using oracle phase in signal reconstruction highlighting the effectiveness of the proposed phase estimation method in the framework of the phase-aware amplitude estimation. In babble noise, the improvement because of phase estimation in phase-aware amplitude estimator increases as the input SNR is rising. However, the gap between phase-aware amplitude estimator using oracle phase and estimated phase performance is quite visible for babble noise. Inferior performance for the babble noise could be explained by deficiency of non-stationary noise estimation. The degradation in phase enhancement-only at low SNRs is due to harmonics in babble noise leading to difficulty in local estimation of phase (computation of k_p). The performance provided by the phase-aware amplitude estimator using the estimated phase asymptotes that obtained by oracle phase as SNR increases.

5. CONCLUSION

We presented new spectro-temporal constraints on phase spectrum to solve the phase estimation problem in single-channel speech enhancement. The proposed constraints were employed to resolve the ambiguity in phase estimation in single-channel speech enhancement problem considering only the geometry of speech and noise. The current study indicates the effectiveness of the proposed phase estimation approach to push the limits of conventional single-channel speech enhancement in which the noisy phase is used for signal reconstruction. Our experiments showed that for SNR ≥ 5 (dB), the proposed phase estimation methods consistently improve the perceived speech quality compared to the case where noisy phase is used. The estimated phase could further improve the spectral amplitude estimation resulting in substantial improvement in perceived speech quality.

[1] Sample wave files are available online at the following link: http://www.spsc.tugraz.at/iwaenc2014

6. REFERENCES

[1] P. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, Boca Raton, 2007.

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

[3] K. K. Paliwal and L. D. Alsteris, “On the usefulness of STFT phase spectrum in human listening tests,” *Speech Communication*, vol. 45, no. 2, pp. 153 – 170, 2005.

[4] 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.

[5] P. Mowlaee and R. Martin, “On phase importance in parameter estimation for single-channel source separation,” in *The International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2012, pp. 1–4.

[6] P. Mowlaee and M. Watanabe, “Partial phase reconstruction using sinusoidal model in single-channel speech separation,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 1–5.

[7] P. Mowlaee and R. Saeidi, “On phase importance in parameter estimation in single-channel speech enhancement,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7462–7466.

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

[9] T. Gerkmann and M. Krawczyk, “MMSE-optimal spectral amplitude estimation given the STFT-phase,” *SPL, IEEE*, vol. 20, no. 2, pp. 129 –132, Feb. 2013.

[10] M. Krawczyk and T. Gerkmann, “STFT phase improvement for single channel speech enhancement,” in *International Workshop on Acoustic Signal Enhancement; Proceedings of IWAENC*, 2012, pp. 1–4.

[11] P. Mowlaee, R. Saiedi, and R. Martin, “Phase estimation for signal reconstruction in single-channel speech separation,” in *Proceedings of the International Conference on Spoken Language Processing*, 2012, pp. 1–4.

[12] 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.

[13] J. Le Roux and E. Vincent, “Consistent Wiener filtering for audio source separation,” *IEEE SPL*, vol. 20, no. 3, pp. 217 – 220, 2013.

[14] A. P. Stark and K. K. Paliwal, “Speech analysis using instantaneous frequency deviation,” in *9th Annual Conference of the International Speech Communication Association*, 2008, pp. 22–26.

[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] C. Breithaupt, M. Krawczyk, and R. Martin, “Parameterized MMSE spectral magnitude estimation for the enhancement of noisy speech,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, March 2008, pp. 4037–4040.

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

[18] L. Rabiner and J.B. Allen, “On the implementation of a short-time spectral analysis method for system identification,” *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 1, pp. 69–78, Feb 1980.

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

[20] R. McAulay and T. Quatieri, “Speech analysis/synthesis based on a sinusoidal representation,” *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 34, no. 4, pp. 744 – 754, aug 1986.

[21] J. R. Carson and T. C. Fry, “Variable Frequency Electric Circuit Theory with Application to the Theory of Frequency Modulation,” *Bell System Technical Journal*, vol. 16, pp. 513–540, 1937.

[22] M. Lagrange and S. Marchand, “Estimating the instantaneous frequency of sinusoidal components using phase-based methods,” *J. Audio Eng. Soc*, vol. 55, no. 5, pp. 385–399, 2007.

[23] I. Cohen, “Relaxed statistical model for speech enhancement and a priori SNR estimation,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 870–881, 2005.

[24] B. Yegnanarayana and H.A. Murthy, “Significance of group delay functions in spectrum estimation,” *IEEE Transactions on Signal Processing*, vol. 40, no. 9, pp. 2281–2289, Sep 1992.

[25] A. P. Stark and K. K. Paliwal, “Group-delay-deviation based spectral analysis of speech,” in *INTERSPEECH*, 2009, pp. 1083–1086.

[26] M. Cooke, J. R. Hershey, and S. J. Rennie, “Monaural speech separation and recognition challenge,” *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 1–15, 2010.

[27] 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.