# On The Use of Discrete Cosine Transform Polarity Spectrum in Speech Enhancement


Sisi Shi
*Signal Processing Laboratory*
*Griffith University*
Brisbane, Australia
sisi.shi@griffithuni.edu.au

Andrew Busch
*School of Engineering*
*Griffith University*
Brisbane, Australia
a.busch@griffith.edu.au

Kuldip Paliwal
*Signal Processing Laboratory*
*Griffith University*
Brisbane, Australia
k.paliwal@griffith.edu.au

Thomas Fickenscher
*High-Frequency Engineering*
*Helmut Schmidt University*
Hamburg, Germany
thomas.fickenscher@hsu-hh.de



***Abstract*—This paper investigates the use of short-time Discrete Cosine Transform (DCT) for speech enhancement. We denote the absolute values and signs of the DCT spectral coefficients as the Absolute Spectrum (AS) and Polarity Spectrum (PoS), respectively. We theoretically show that the noisy PoS is the best estimate of the original, under the constrained MMSE criterion. To verify this experimentally, the effect of using the noisy PoS for signal resynthesis is analysed through objective and subjective measures. The results show that when the Instantaneous SNR (ISNR) is above 0 dB, deemed as perfect, recovery of the original speech signal can be obtained only by modifying the DCT absolute spectrum. However, an accurate DFT Phase Spectrum (PhS) estimation might be required to achieve the same improvement in perceived speech quality. When the perceived quality is measured against the Segmental SNR (SSNR), it shows the PoS is more capable to conserve the speech quality than the PhS for the same level of global distortion. The results show that the noisy PoS can be used as an estimate of the clean PoS without perceivable degradation in speech quality, only if the ISNR of the noisy speech signal is above 0 dB or the SSNR is above 10.5 dB.**



***Index Terms*—Speech enhancement, Discrete cosine transform (DCT), Just noticeable difference (JND)**


## I. Introduction

Despite the theoretical advantages of the DCT [1], most speech enhancement techniques still prefer the DFT spectrum, which has readily available short-time Spectral Amplitude (STSA) estimators [2]–[6]. The short-time[1] DCT analysis of a sequence $\{y(n), 0 \leqslant n \leqslant N-1\}$ is given by

$$Y_{\mathcal{C}}(i,k) = m_k \sum_{n=0}^{N_w-1} y(n+iN_s)w(n)\cos\left[\frac{(2n+1)k\pi}{2L}\right] \quad (1)$$

where $0 \leqslant k \leqslant L-1$ and:

$$m_k = \begin{cases} \sqrt{\frac{1}{L}} & \text{for } k=0 \\ \sqrt{\frac{2}{L}} & \text{for } k \neq 0 \end{cases} \quad (2)$$

$n$, $k$ and $i$ are the discrete time, frequency and frame index, respectively. $w(n)$ is the analysis window function of duration $N_w$. $N_s$ and $L$ are the length of the frame shift and frequency analysis, respectively. In speech processing, a window duration between 20 to 40 ms is typically used, so that the properties of the signal do not change appreciably; and the windows must overlap by at least 75% to avoid aliasing [7]. For this study, we denote the modulus and signs of the DCT spectral coefficients $Y_{\mathcal{C}}(i,k)$ as the Absolute Spectrum (AS) and Polarity Spectrum (PoS), respectively. The use of the DCT spectra for speech enhancement has not been extensively researched to date.

On the other hand, the role of DFT spectral components has been extensively discussed. The DFT analysis is given by [8]:

$$Y_{\mathcal{F}}(i,k) = \sum_{n=0}^{N_w-1} y(n+iN_s)w(n)e^{-j\frac{2\pi}{L}nk},\ 0 \leqslant k \leqslant L-1, \quad (3)$$

The absolute values and phases of the DFT coefficients are known as the Magnitude Spectrum (MS) and Phase Spectrum (PhS), respectively. Many DFT-based algorithms enhance only the MS, while the noisy PhS is used as such, i.e., [3]–[6]. This is justified by the assumption that phase is perceptually unimportant [9] and the MMSE estimator of the original phase is in fact the noisy phase [3]. However, Paliwal et al. have suggested that phase has useful information towards speech quality [10]–[12], especially for low SNRs. The effect of using noisy phase for speech synthesis was discussed in [13], [14]. In these studies, an Analysis-Modification-Synthesis (AMS) framework was used to create phase-only (PhO) stimuli. The PhO stimuli was generated by adding a controlled level of distortion in the PhS, while keeping the MS fixed from the clean input. Thus the effects on the perceived speech quality are a result from the changes in PhS only. It was found that the re-synthesised speech sounds like the original as long as the level of distortion is below a certain threshold. Above that threshold, some roughness is perceivable by the listener. Through informal listening tests, the threshold or the Just Noticeable Difference (JND), of perception of phase deviation was found to be roughly 6 dB in Instantaneous spectral Signal-to-Noise Ratio (ISNR) [13]. This demonstrates speech enhancement can be achieved by modifying only the MS and leaving the noisy PhS uncorrected, provided that the local SNR at a frequency bin is at least 6 dB [13]. Chappel et al. qualified the JND with respect to a global Segmental SNR (SSNR) using a form of "A or B" listening tests [14]. These binary tests were adequate for evaluating the quality of generated stimuli but rather unfavorable for threshold tracking, because

[1]In this paper the short-time modifier is implied when referring to the DFT, DCT and their corresponding spectra unless otherwise stated.

they were inevitably biased by the false positive error. That is, listeners might prefer the modified stimuli over the clean due to an internal criterion that under their conscious control [15]. Also, only two utterances were used to construct the stimuli, which may introduce more bias to results reported. To date, there is no equivalent work showing what JND is required to achieve an optimal result for the DCT-based methods.

The aim of this study is to explore the relevance of DCT Polarity Spectrum in the context of STSA estimation-based speech enhancement. To achieve this, we derive the optimal MMSE polarity estimator base on modelling speech and noise DCT spectral coefficients as zero mean statistically independent Gaussian random variables (Section II). We show that the noisy PoS is the constrained MMSE estimator of the clean PoS. Hence, the use of noisy PoS to construct the enhanced signal is justified on theoretical grounds. To verify these findings experimentally, we use an approach reported in [14] and describe it in Section III. For this, we create the Polarity-Only (PO) stimuli to investigate the effect of modified PoS on the perceived speech quality, and compare these results with Phase-Only (PhO) stimuli with both objective (PESQ) and subjective (JND) tests. Both of these metrics were measured with respect to ISNR and Segmental SNR for listeners with normal hearing.

## II. MMSE Polarity Estimator

Let the clean speech signal, noisy speech signal, and noise signal be denoted by x(n), y(n) and d(n), respectively. The additive noise model can be expressed as:

$$y(n) = x(n) + d(n), \quad 0 \leqslant n \leqslant N-1 \tag{4}$$

Let $Y_{\mathcal{C}}(i,k) \triangleq \phi_Y(i,k)|Y_{\mathcal{C}}(i,k)|$, $X_{\mathcal{C}}(i,k)$, $D_{\mathcal{C}}(i,k)$ denote the DCT spectral coefficients of the noisy $y(n)$, the clean $x(n)$, and the noise signal $d(n)$, respectively. $|Y_{\mathcal{C}}(i,k)|$ is the AS and $\phi_Y(i,k) = \mathrm{sgn}(Y_{\mathcal{C}}(i,k))$ is the PoS. The frame index $i$ and the frequency index $k$ are subsequently omitted for the sake of brevity. Equation (4) can be represented in the DCT domain as:

$$\phi_Y|Y_{\mathcal{C}}| = \phi_X|X_{\mathcal{C}}| + \phi_D|D_{\mathcal{C}}| \tag{5}$$

with $Y_{\mathcal{C}}$ (similar with $X_{\mathcal{C}}$ and $D_{\mathcal{C}}$) is given by (1).

With the assumption that the DCT spectral coefficients are statistically independent, the MMSE polarity estimator, $\widetilde{\phi}_X$, can be obtained from $Y_{\mathcal{C}}$ as follows:

$$\begin{aligned}\widetilde{\phi}_{\mathrm{X}} &= \mathrm{E}\left\{\phi_{\mathrm{X}}|Y_{\mathcal{C}}\right\} \\ &= \int_{-\infty}^{\infty} \phi_X f(X_{\mathcal{C}}|Y_{\mathcal{C}})\,\mathrm{d}x_{\mathcal{C}}\end{aligned} \tag{6}$$

where $\mathrm{E}\{\cdot\}$ denotes the expectation operator, and $f(\cdot)$ denotes the Probability Density Function (PDF).

Under the Gaussian distribution assumptions, $f(X_{\mathcal{C}}|Y_{\mathcal{C}})$ is given by

$$f(X_{\mathcal{C}}|Y_{\mathcal{C}}) = \sqrt{\frac{\lambda_x+\lambda_d}{2\pi\lambda_x\lambda_d}}\exp\left\{-\frac{[(\lambda_x+\lambda_d)X_{\mathcal{C}} - \lambda_x Y_{\mathcal{C}}]^2}{2\lambda_x\lambda_d(\lambda_x+\lambda_d)}\right\} \tag{7}$$

where $\lambda_x(k) = \mathrm{E}\left\{|X_{\mathcal{C}}|^2\right\}$ and $\lambda_d(k) = \mathrm{E}\left\{|D_{\mathcal{C}}|^2\right\}$, are the variances of the speech and the noise coefficients, respectively. Substituting (7) into (6) gives

$$\widetilde{\phi}_{\mathrm{X}} = \mathrm{erf}\left(\sqrt{\frac{v_k}{2}}\right)\phi_Y \tag{8}$$

with $v_k \triangleq \frac{\xi}{1+\xi}\gamma$, the error function $\mathrm{erf}\,(\cdot)$ [16, eq.8.250.1], the *a priori* SNR $\xi \triangleq \lambda_x/\lambda_d$ and the *a posteriori* SNR $\gamma \triangleq Y_{\mathcal{C}}^2/\lambda_d$. However, as the modulus of $\widetilde{\phi}_{\mathrm{X}}$ is not equal to unity, the MMSE estimator given by (8) degrades the amplitude estimation when combining with an independently derived amplitude estimator. To solve this, a second estimator was derived with the constraint that the modulus of the resulting estimator is one:

$$\begin{aligned}&\min_{\widehat{\phi}_X} = \mathrm{E}\left\{|\phi_X - \widehat{\phi}_X|^2\right\} \\ &\text{subject to } |\widehat{\phi}_X| = 1\end{aligned} \tag{9}$$

Using the method of Lagrange multipliers [17], we have

$$\frac{\mathrm{d}\,\mathrm{E}\left\{|\phi_X - \widehat{\phi}_X|^2\right\}}{\mathrm{d}\,\widehat{\phi}_X} = \lambda \cdot \frac{\mathrm{d}\Big(|\widehat{\phi}_X| - 1\Big)}{\mathrm{d}\,\widehat{\phi}_X} \tag{10}$$

then,

$$\widehat{\phi}_X = \mathrm{E}\left\{\phi_{\mathrm{X}}\right\}\frac{2}{2-\lambda} \tag{11}$$

Substituting (8) into the above equation yields

$$\widehat{\phi}_X = \mathrm{erf}\left(\sqrt{\frac{v_k}{2}}\right)\phi_Y\frac{2}{2-\lambda} \tag{12}$$

By using the constraint $|\widehat{\phi}_X| = 1$, we get from (12)

$$\lambda = 2\left[1 \pm \mathrm{erf}\left(\sqrt{\frac{v_k}{2}}\right)\right] \tag{13}$$

Finally, substituting (13) back to (12) gives

$$\widehat{\phi}_X = \pm\phi_Y \tag{14}$$

and (9) attains a minimum on

$$\widehat{\phi}_X = \phi_Y \tag{15}$$

Hence, the polarity of the noisy signal, $\phi_Y$, is the MMSE estimator of the clean polarity, independent of the amplitude estimation.

## III. Objective experiments

### A. Speech corpus

Speech samples are from the TSP speech database [18]. TSP corpus contains over 1400 utterances, belonging to 24 speakers (12 male and 12 female). These recordings were filtered with a linear phase, low-pass FIR filter and down-sampled to 16 kHz. Corresponding noisy stimuli were generated by degrading the clean stimuli with Additive White Gaussian Noise (AWGN) at various SNRs.

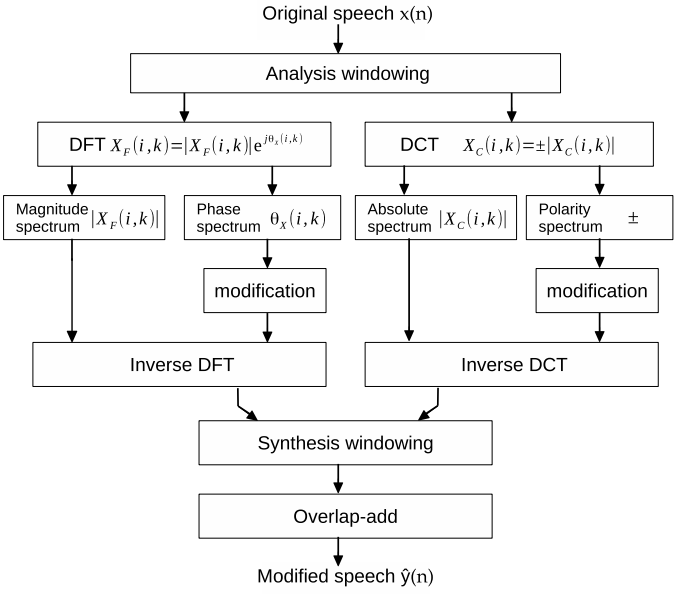



Fig. 1: Block diagram of the AMS procedure

### B. Stimuli construction

The clean corpus are processed through the AMS framework (Fig.1) to obtain the Polarity-Only (PO) and Phase-Only (PhO) stimuli. At the analysis stage, the original signal is segmented into 32ms frames, each with an overlap of 75%. For consistency with the earlier work of [13], [14], in the first scenario we modify the clean spectrum with respect to the ISNR. In practice, the ISNRs will not be fixed across all time instances and frequencies. Therefore, in the second scenario, the modification is done in terms of Segmental SNR (SSNR), which can be calculated across an entire speech utterance [14].

*a) Modification with respect to Instantaneous SNR:* To construct the PO stimuli, (1) is used to obtain the DCT spectral components. A noisy signal $Y_{\mathrm{C}}$ is generated such that, the ISNR for each frequency bin is constant, and equal to:

$$\mathrm{ISNR} = 10\log_{10}\left(\frac{|X_{\mathcal{C}}|^2}{|D_{\mathcal{C}}|^2}\right) \tag{16}$$

Rearranging (16), we obtain the required amplitude of the noise signal:

$$|D_{\mathcal{C}}| = \frac{|X_{\mathcal{C}}|}{\sqrt{10^{\frac{\mathrm{ISNR}}{10}}}} \tag{17}$$

The noise PoS, $\phi_D$, is a random vector with length L and each element is chosen uniformly from the set $\{-1, 1\}$. The noisy signal is generated by using (5). The PoS of the noisy signal is then combined with AS of the clean signal, that is,

$$\widehat{Y}_{\mathrm{C}} = \phi_Y |X_{\mathrm{C}}| \tag{18}$$

Similar procedure is used to produce the phase-only (PhO) stimuli. The DFT representation of (4) is:

$$|Y_{\mathcal{F}}|e^{j\theta_Y} = |X_{\mathcal{F}}|e^{j\theta_X} + |D_{\mathcal{F}}|e^{j\theta_D} \tag{19}$$

The MS of the noise signal, $|D_{\mathcal{F}}|$, was generated by using (17). The random noise phase $\theta_D \triangleq \angle D_{\mathcal{F}}$ is uniformly distributed in the interval $[-\pi, \pi]$. The modified DFT domain signal is given by:

$$\widehat{Y}_{\mathcal{F}} = |X_{\mathcal{F}}|e^{j\theta_Y}. \tag{20}$$

*b) Modification with respect to Segmental SNR:* The SSNR is defined as

$$\mathrm{SSNR} = \frac{1}{M}\sum_{i=0}^{M-1} 10\log_{10}\frac{\sum_{n=i\cdot N_w}^{i\cdot N_w+N_w-1} x^2(n)}{\sum_{n=i\cdot N_w}^{i\cdot N_w+N_w-1}(x(n)-y(n))^2} \tag{21}$$

where $M$ is the number of signal segments that contain speech. The speech frames with energy greater than $-45$dB with respect to the maximum frame energy are concluded for summation in (21). To generate the noisy utterance, the level of AWGN required to achieve the desired SSNR is first calculated as

$$\varphi = \sqrt{\frac{10^{P_{Cln}/10}}{10^{SSNR/10}}} \tag{22}$$

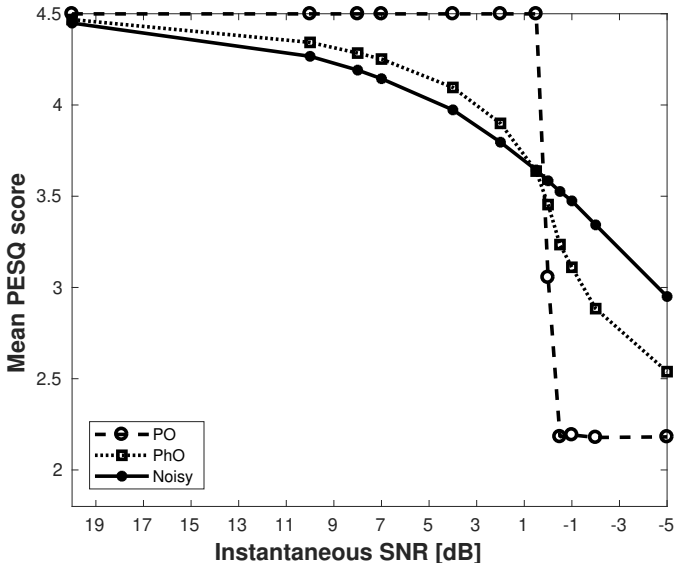
where $P_{Cln}$ is the power of the clean speech. The noise signal $d(n)$ was found as the product of $\varphi$ and a Gaussian distributed vector with length N, initialized with zero mean and unit variance. The generated noise then was added to the clean signal $x(n)$ to produce the noisy signal $y(n)$ as in (4). Upon framing and taking the DCT or the DFT, we obtain the spectral components of the noisy signal. Using (18) for the DCT domain signals and (20) for the DFT domain signals, the corresponding modified DCT signal $\widehat{Y}_{\mathcal{C}}$ and modified DFT signal $\widehat{Y}_{\mathcal{F}}$ are generated.

Finally, the least-squares overlap-add [19] method is applied to the modified speech frames, $\widehat{Y}_{\mathcal{C}}$ or $\widehat{Y}_{\mathcal{F}}$, to construct the synthesized signal, $\widehat{y}(n)$. The Hamming window used for analysis and the modified Hann window [20], which reduces the computation requirements for partial window overlap is employed for synthesis.

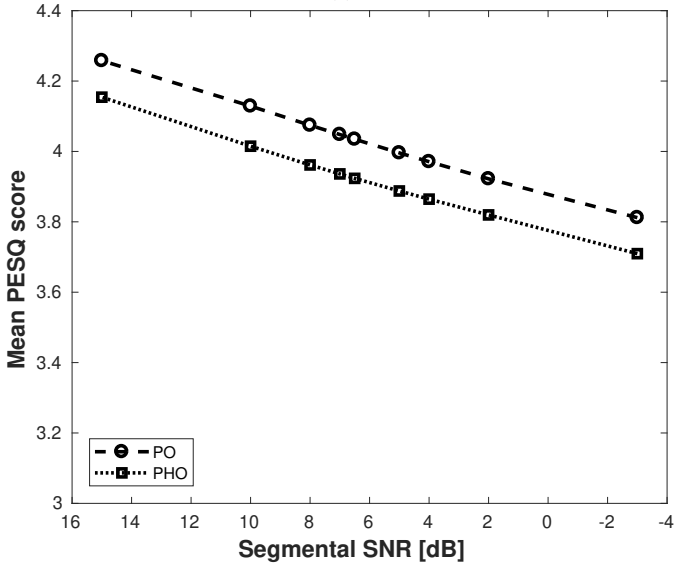
### C. Experiment Results

The Perceptual Estimation of the Speech Quality (PESQ) [21] metric was employed as an objective speech quality measure. The PESQ provides a score between 0.5 and 4.5 which predicts the quality of a degraded speech signal. A score of 1.0 corresponds to low quality and 4.5 corresponds to distortionless. In our objective experiments, mean PESQ scores were computed over the 30 sentences of the corpus for each treatment being investigated.

Fig.2a shows when ISNR $>$ 0 dB, the quality measure of the PO stimuli stays at optimum. This signifies that modifying the PoS alone had no effect on the perceived quality when the local SNR is above 0 dB, and the effects solely came from modifying the AS. This result is expected by the theoretical analysis, as the noise in each frequency bin is guaranteed to be smaller in amplitude than the clean speech signal and thus it is impossible for the polarity of that bin to change. On the other hand, the quality measure of the PhO stimuli declines non-linearly with the ISNR, which implies the PhS simultaneously




(a)




(b)

Fig. 2: Perceived quality estimation for Polarity-Only (PO, dashed line), Phase-Only (PhO, dotted line) and noisy (solid) stimuli as a function of (a) Instantaneous SNR and (b) Segmental SNR.

contributed to the perceived quality of the reconstructed speech signal; thus, retrieving the DFT MS alone might not provide the optimal improvement in speech quality. Moreover, it shows the DCT polarities and DFT phases play more significant roles than their associated spectral amplitudes when ISNR $\leqslant$ 0 dB, which implies the original signal's energy is less than that of the noise. In this case, if the resultant DCT polarity opposes to the original's, it causes an equivalent $\pi$ radians phase deviation in the polarity spectrum. As a result, the quality measure of the PO stimuli dropped substantially at this threshold. However, in practice it only occurs rarely during a speech active region as the overall speech energy is generally higher than the noise.

The results of testing using the noisy signals with specified SSNR, which is a more realistic distortion estimator, are shown in Fig.2b. These results show that for both PO and PhO stimuli the quality measure declines linearly as the distortion increases; however, it graded higher for the PO stimuli than the PhO stimuli at all given SSNR values. It suggests that the DCT polarity spectrum is more capable of conserving the speech quality than the DFT phase spectrum for the same level of global distortion. Since unless the noise energy is greater than the speech energy at a particular frequency bin, the PoS will not be corrupted. This allows PoS to have a higher degree of distortion tolerance than the PhS.

## IV. Subjective experiments

Accurate measurements of the JND are fundamental indicators of optimal speech enhancement schemes due to the psycho-acoustically motivated criterion. We estimate the JND of perception of DCT polarity aberration via the adaptive psycho-acoustic procedure [22]. A three-interval, three-alternative forced-choice (3I-3AFC) task was used to track the JND. Unlike a "yes/no" task used in [14], a forced-choice task is not biased by the false positive errors, because there is only one correct response. The adaptive procedure is described as follows.

### A. Procedures

Twelve participants (6 females and 6 males) aged between 24 and 41 years were recruited from Griffith University. At the beginning of each trial, a clean corpus was randomly chosen from the TSP database as the reference signal. Then the variable stimulus was generated by adding noise to one of its spectrum adaptively. Two types of stimuli were performed to measure the JND and the impairment of the stimulus was evaluated with respect to the ISNR or the SSNR (as described in Sec.III-B).

In each trial, a set of three stimuli was presented in temporal succession and separated by silent gaps. One variable stimulus contained the distortion, whereas the others were kept clean as references and the order of these stimuli was randomized. The participant was instructed to report which of the three intervals contained the degraded stimuli, and depending on the response, the stimulus level were varied across trials. The 2-up, 1-down procedure was implemented for threshold tracking. In such a case, the level of the variable stimulus incremented toward the threshold after two consecutive correct responses and stepped down from the threshold after one incorrect response. A relative large step size was used to approach the threshold quickly, and a smaller one for staying close to the threshold in successive runs. The final JND was evaluated by averaging the various threshold estimates collected at the reversal points.

### B. Experiment Results

The results, summarised in Table.I, suggest that no degradation in the DCT polarity spectrum is discriminated by the

listeners when the ISNR is above 0 dB or the SSNR is above 10.5 dB. Contrarily, DFT phase deviation in the speech become perceptible if the ISNR falls below 5 dB or the SSNR falls below 15.5 dB. Therefore, the noisy PoS can be used as an estimate of the clean PoS if the ISNR is above 0 dB or the SSNR is above 10.5 dB. However, the noisy PhS can only be used as an approximation of the original if the ISNR is at least about 5 dB or the SSNR is at least 15.5 dB. Consequently, DFT-based speech enhancement methods may require 5 dB higher improvement in terms of ISNR and SSNR than the DCT-based methods.

TABLE I: Comparison of JNDs in terms of ISNR and SSNR

| Type | Metric [dB] | Mean | Std[c] | Geo. Mean | Median |
|---|---|---|---|---|---|
| PO[a] | ISNR | 0.21 | 0.09 | 0.00 | 0.25 |
| PhO[b] | ISNR | 5.10 | 1.04 | 4.99 | 5.50 |
| PO | SSNR | 10.43 | 1.8 | 10.10 | 10.50 |
| PhO | SSNR | 15.45 | 1.7 | 15.36 | 16.03 |

[a]DCT Polarity-Only, [b]DFT Phase-Only, [c]Standard deviation.

## V. Conclusion

In this paper, the use of short-time DCT Polarity Spectrum (PoS) in speech enhancement is investigated. A theoretical analysis showed that the optimal estimate of the clean PoS is the noisy PoS under the constrained MMSE criterion, justifying the use of the noisy PoS for signal reconstruction. To verify this result experimentally, we evaluated the relevance of the PoS towards perceived speech quality using both objective (PESQ) and subjective (JND) testing for both ISNR and SSNR generated stimuli. The objective results show that when the ISNR is above 0 dB, modifying the PoS has no effect on perceived speech quality; however, an accurate DFT phase estimation might be required to achieve the same improvement in perceived speech quality. Additionally, when the perceived speech quality was measured against Segmental SNR (SSNR), it shows that the DCT polarity spectrum is better able to conserve the speech quality than the DFT phase spectrum for the same level of global distortion. Subjective testing was also conducted by determining the JND via the adaptive psychophysical procedure. The results of this testing showed that the JND resulting from polarity distortion is 0 dB in ISNR and 10.5 dB in SSNR, which are about 5 dB lower than the corresponding JNDs in DFT phase spectrum. Overall, both the theoretical and experimental results demonstrate that speech enhancement can be achieved by combining the noisy PoS with accurately estimated DCT Absolute Spectrum (AS), with no degradation in speech quality perceived due to the use of the noisy PoS, provided that the ISNR is above 0 dB or the SSNR is greater than 10.5 dB. Future work will aim to develop accurate estimators for the DCT AS in order to further test the effectiveness of the DCT representation for speech enhancement.

## References


[1] Y. Soon, S. N. Koh, and C. K. Yeo, "Noisy speech enhancement using discrete cosine transform," *Speech communication*, vol. 24, no. 3, pp. 249–257, 1998.

[2] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 2, pp. 137–145, 1980.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[4] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE transactions on acoustics, speech, and signal processing*, vol. 33, no. 2, pp. 443–445, 1985.

[5] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1741–1752, 2007.

[6] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2014.

[7] L. R. Rabiner and R. W. Schafer, *Theory and applications of digital speech processing*. Pearson Upper Saddle River, NJ, 2011, vol. 64, ch. 7.

[8] P. Vary and R. Martin, *Digital speech transmission: Enhancement, coding and error concealment*. John Wiley & Sons, 2006.

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

[10] K. K. Paliwal and L. Alsteris, "Usefulness of phase spectrum in human speech perception," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[11] L. Alsteris and K. Paliwal, "Further intelligibility results from human listening tests using the short-time phase spectrum," *Speech Commun.*, vol. 48, no. 6, pp. 727–736, Jun 2006.

[12] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *speech communication*, vol. 53, no. 4, pp. 465–494, 2011.

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

[14] S. B. Chappel, R. and K. Paliwal, "Phase distortion resulting in a just noticeable difference in the perceived quality of speech," *Speech Commun.*, vol. 23, no. 8, pp. 138–147, 2016.

[15] W. H. Ehrenstein and A. Ehrenstein, "Psychophysical methods," in *Modern techniques in neuroscience research*. Springer, 1999, pp. 1211–1241.

[16] J. Laneau, J. Wouters, and M. Moonen, "Relative contributions of temporal and place pitch cues to fundamental frequency discrimination in cochlear implantees," *The Journal of the Acoustical Society of America*, vol. 116, no. 6, pp. 3606–3619, 2004.

[17] W. Fulks, *Advanced calculus; an introduction to analysis*, 2nd ed. John Wiley & Sons, 1961, ch. 10, pp. 297–300.

[18] P. Kabal, "Tsp speech database," *McGill University, Database Version*, vol. 1, no. 0, pp. 09–02, 2002.

[19] T. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. Upper Saddle River, NJ: Prentice Hall, 2002.

[20] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[21] 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," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[22] M. R. Leek, "Adaptive procedures in psychophysical research," *Perception & psychophysics*, vol. 63, no. 8, pp. 1279–1292, 2001.