Received 13 March 2023, accepted 30 March 2023, date of publication 5 April 2023, date of current version 10 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3264436

RESEARCH ARTICLE

Log-Spectral Amplitude and Spectral Polarity Estimation in Short-Time Discrete Cosine Transform Domain

SISI SHI, KULDIP K. PALIWAL, AND ANDREW BUSCH
School of Engineering, Griffith University, Brisbane, QLD 4111, Australia

Corresponding author: Sisi Shi (sisi.shi@alumni.griffithuni.edu.au)

This work was supported by the Signal Processing Laboratory, Griffith University, Nathan Campus, Brisbane, QLD, Australia.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Griffith University under Application No. 2018/671, and performed in line with the Human Research Ethics.

ABSTRACT Single-channel speech enhancement based on short-time spectral amplitude (STSA) estimation often uses the unmodified phase spectrum for speech re-synthesis, thereby introducing undesired artifacts to the enhanced speech. Using discrete Cosine transform (DCT) instead of discrete Fourier transform (DFT) reduces the effects of such issues because the consequences of using noisy DCT polarities for speech re-synthesis are less severe than using the noisy DFT phases. Although DFT-based STSA estimators have been adequately studied in the past, such estimators have not sufficiently been developed for the DCT domain. This study aims to demonstrate the superiority of DCT representation in STSA estimation-based speech enhancement. To achieve this, we first derive the DCT-based STSA estimator which minimizes the mean squared error (MSE) of the log-spectral amplitudes (LSA). We then propose a novel DCT polarity estimator to be used in combination with the STSA estimator. The clean speech DCT coefficients are modeled by a Gaussian or a Laplace density and the noise DCT coefficients are modeled by a Gaussian density. To assess the enhanced speech, objective and subjective quality measures are employed. Results show that the new estimators performed better and are widely preferred by listeners over the corresponding DFT-based estimators. Moreover, the proposed STSA estimators can be expressed in the closed-form, whereas the DFT-based estimator with super-Gaussian speech prior has no closed-form solutions.

INDEX TERMS Discrete Cosine transform (DCT), minimum mean-square error (MMSE) estimator, speech enhancement, perceptual distortion measure, polarity estimation.

I. INTRODUCTION

Speech communication devices including hearing aids, cochlear implants, and mobile communication are required to function robustly in adverse noisy environments. Ambient addictive noise can corrupt speech signals, leading to degraded device performance and listener fatigue. Hence, it is crucial to perform speech enhancement to improve the quality, and preferably the intelligibility of the noisy speech [1]. When the noisy speech alone is accessible (i.e., single-channel), most speech enhancement systems use DFT, which has readily available STSA estimators [2]. Many of these estimators enhance only the short-time[1] spectral amplitudes (STSA), while the noisy spectral phase is left unmodified, i.e., [3], [4], [5], [6]. This is justified by the assumption that phase is perceptually unimportant [7] and the MMSE estimator of the original phase is the noisy phase [3].

Although DFT-based STSA estimators have been well developed in the past, they suffer from two major limitations. First, using the noisy phase for speech re-synthesis (reconstruction) introduces an upper bound on the maximum improvement in speech quality [8]. Because the spectral phase in fact contributes to the perceived speech quality

The associate editor coordinating the review of this manuscript and approving it for publication was Lin Wang.

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

especially for low SNRs [9], [10], [11]. It has been shown that using noisy phase for speech re-synthesis does not degrade the perceived speech quality as long as the level of phase distortion is below a certain threshold, i.e., Just-Noticeable-Difference (JND) [8]; above the JND, some roughness is perceivable by the listener [8], [12]. Consequently, having a high JND in noise distortion (or corresponding to a low SNR) is a very desirable feature for speech enhancement applications. The JND of perception in DFT phase distortion is roughly 5-6 dB in instantaneous spectral signal-to-noise ratio (ISNR) [8], [13]. Below 5-6 dB ISNR, DFT-based STSA estimators can no longer effectively improve speech quality due to perceivable phase distortion. Second, DFT-based STSA estimators derived under super-Gaussian speech prior are sub-optimal solutions due to inaccurate assumptions. Especially, the following assumptions are commonly invoked to significantly simplify the derivation of DFT-based MMSE STSA estimators (Table 1): (1) the real and imaginary parts of the speech DFT coefficients are independent; (2) the speech STSA and phases are independent and their join pdf satisfies $p(|X|, \theta_X) = p(|X|)p(\theta_X)$; and (3) the speech phase is uniformly distributed. These assumptions are true for complex Gaussian random variables (R.V.) [14], however, it is not the case with the complex super-Gaussian distributions such as Laplace distribution that are used for modeling the speech DFT coefficients. For example, the real and imaginary parts of complex Laplace R.V. are not independent according to measured histograms presented in [15] and [5]. Furthermore, the analyses of the joint pdf of the speech STSA and phases $p(|X|, \theta_X)$ as well as the pdf of $p(\theta_X)$ illustrate the Laplace amplitude depends on the phase, i.e., $p(|X|, \theta_X) \neq p(|X|)p(\theta_X)$, and the density of $p(\theta_X)$ is clearly not uniform but oscillates near the value of $1/2\pi$ [16]. Nonetheless, despite all those efforts to simplify the derivation, DFT-based estimators derived under super-Gaussian speech prior [5], [6], [16], [17] have no closed-form solutions and require numerical approximations due to the induced mathematical complexity.

We propose that the aforementioned limitations related to the conventional DFT-based STSA estimators can be alleviated by using DCT instead of DFT and thereby improving the perceived speech quality. Three main reasons are as follows. First, the approximation of the clean DCT polarity spectrum (PoS) by its noisy counterpart can be considered superior, with a significantly lower JND (in SNR) when compared to DFT-based methods. As demonstrated in [13], the JND in the DCT PoS is 0 dB in ISNR, which are about 5-6 dB lower than the threshold in the DFT phase spectrum (PhS). This means when the ISNR is above 0 dB, leaving the DCT polarity unmodified 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. This advantage was further proved through subjective listening tests and the average of the scores given by the listeners, termed as the mean subjective preference (%) score, was used as an indicator for the perceived speech quality [19]. Explicitly, for all tested noise conditions, using noisy PoS for speech reconstruction achieved a higher preference score than using the noisy PhS (Fig. 1). This suggests that the PoS is more capable of conserving speech quality than the PhS for the same level of distortion. Second, since the DCT coefficient is real, the polarity component only depends on the individual DCT coefficient, and thus the potential issue caused by assumption (1) does not apply (Table 1). More importantly, it is sufficient to specify only the density for the speech DCT coefficient, i.e., it is not necessary to find the distribution of the DCT STSA nor the DCT polarity. The advantages are that the DCT MMSE STSA estimators can be derived directly from the density of speech coefficient without making the independence assumption of the speech STSA and its polarity part. Consequently, the potential issues caused by assumptions (2) and (3) can be essentially avoided. Lastly, the real component of DCT is mathematically easy to calculate and thus, the DCT-based STSA estimators can be elegantly reduced to closed-form solutions.

Since the consequences of using noisy DCT polarities for speech re-synthesis are less severe than using the noisy DFT phases, adopting DCT representations in STSA estimation-based speech enhancement can potentially lead to higher speech quality. It is important to note that, DFT-based estimators cannot simply be carried over to the DCT domain as DCT is a real-valued transform and DFT is complex. In this regard, DCT-based MMSE spectral amplitude (STSA) estimators were developed in [19] by minimizing the MSE between the original STSA and its estimator. To be used in conjunction with the DCT-based STSA estimators, an optimal MMSE estimator of the DCT noise power was also derived. It has been shown when the enhanced DCT spectral amplitudes combined with the noisy polarity spectrum, the resulting speech achieved significantly better perceived quality than the ones obtained by the DFT-based approaches.

While the distortion measure previously used in [19] leads to good results, it is not the most perceptually meaningful one. Studies show that the human auditory performs a logarithmic compression of the STSA [20], and thus the distortion measure which is based on the MSE of log-spectral amplitude (LSA) is more perceptually relevant than that of the STSA. Initially, Ephraim and Malah proposed the LSA estimator [4] which minimizes the log-domain MSE. Motivated by the central limit theorem, the complex Gaussian model was used for both DFT clean speech and noise components. Later studies show that while noise components can be appropriately modeled by Gaussian distributions, clean speech components in the decorrelated domains are more accurately described by super-Gaussian distributions such as Laplacian (double-sided Exponential) [18], [21], [22]. Thus, employing a super-Gaussian speech prior instead of the Gaussian can improve the performance of MMSE STSA estimators [18], [22], [23], [24]. In particular, Hendriks et al. [6] derived an

TABLE 1. Common assumptions or approximations used for deriving DFT-based MMSE STSA estimators as compared to the proposed DCT-based approach.

DFT Domain	**DCT** Domain (proposed)
(1) Independence assumption of real and imaginary parts [18]	Real transform, does not apply
(2) Independence assumption of amplitude and phase [5], [6], [16], [17] i.e., $p(\lvert X\rvert, \theta_X) \approx p(\lvert X\rvert)p(\theta_X)$	No assumption made
(3) Uniform distribution assumption of phase [5], [6], [16], [17] i.e., $p(\theta_X) \approx 1/2\pi$	No assumption made
Marginal distribution of complex noisy coefficient Y(ω) $p(Y) = \int_0^\infty \int_0^{2\pi} p\left[Y \lvert\lvert x\rvert, \theta_x\right] p(\lvert x\rvert, \theta_x) d\theta_x d\lvert x\rvert$ $\approx \int_0^\infty \int_0^{2\pi} p\left[Y \lvert\lvert x\rvert, \theta_x\right] p(\lvert x\rvert)p(\theta_x) d\theta_x d\lvert x\rvert$ $\approx 1/2\pi \int_0^\infty p\left[Y \lvert\lvert x\rvert\right] p(\lvert x\rvert) d\lvert x\rvert$ [5], [6], [16], [17]	Marginal distribution of real noisy coefficient Y $p(Y) = \int_{-\infty}^{\infty} p(Y\lvert x)p(x)\,\mathrm{d}x$ No approximations made
Independence assumption of spectral coefficients	Independence assumption of spectral coefficients

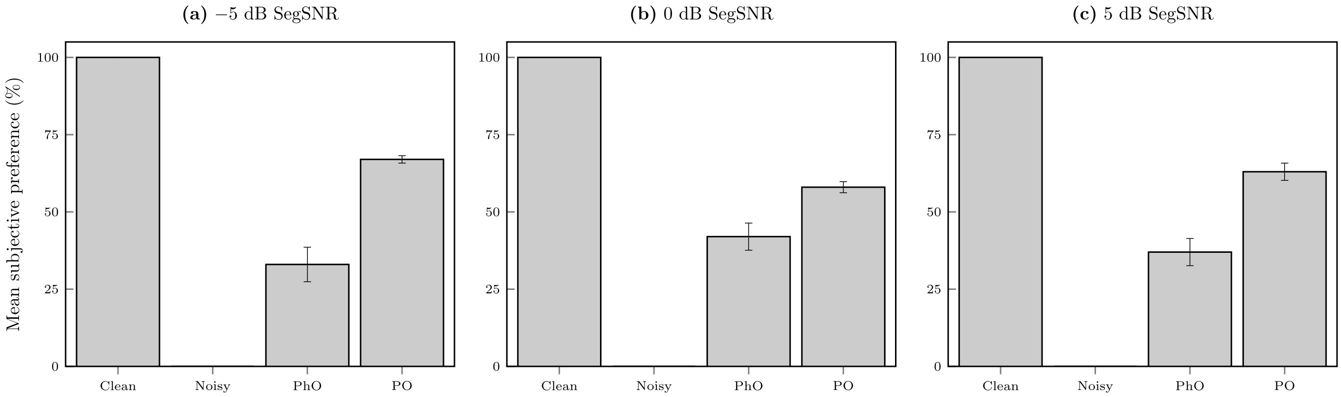

FIGURE 1. Mean preference scores (with standard error bars) for four stimuli types at (a) −5 dB, (b) 0 dB, and (c) 5 dB Segmental SNR (SegSNR) [19]. The polarity-only (PO) [or phase-only (PhO)] stimuli was generated by adding a controlled level of distortion into the DCT polarity spectrum (or the DFT phase spectrum) while keeping its spectral amplitudes fixed from the clean input. Thus the effects on the perceived speech quality result from the changes in the polarity spectrum or phase spectrum only. The distortion was added with respect to the SegSNR. We also include clean speech and noisy (unprocessed) speech as the upper bound and lower bound of the preference, respectively.

MMSE LSA estimator under the assumption that the clean speech DFT amplitudes follow a one-sided chi distribution. To exploit the phase information, [25] recently combined the phase compensation technique with a perceptually weighted β-order STSA estimator. For this method, the noisy speech signal is first framed and transformed into the DFT domain. The phase compensation function is then used to estimate the clean phase spectrum and the perceptually weighted β-order STSA estimator is used to enhance the magnitude spectrum. The modified speech is obtained by combining the estimated magnitude and phase spectra (refer to [25], Fig. 1 and Sec. IV). Although [25] improves speech quality in terms of objective quality metrics, it is unclear whether the improvement stems from a more reliable *a prior* SNR estimate or from the estimator itself because [25] used a different *a prior* SNR estimator for their approach when comparing to other methods. To date, perceptually-motivated STSA estimators have been concentrated in the DFT domain, to the authors' best knowledge, none has been developed in the DCT domain.

This paper aims to combine the advantages of using a real transform like DCT with perceptually-motivated distortion measure in enhancing noisy speech. To achieve this, the DCT-based MMSE LSA estimators are derived based on super-Gaussian speech prior and Gaussian noise prior. The proposed estimators can be expressed in closed form without making any approximations. Furthermore, we derive a novel polarity estimator (PoE) to be used in combination with the STSA estimator. Accordingly, the effect of using PoE on perceived speech quality is examined. The performance of our new estimators is compared to the DCT STSA estimator derived in [19], some of the well known DFT-based LSA estimators, e.g., [4] and [6], as well as the State-Of-The-Art (SOTA) DFT-based phase-aware system, e.g., [25].

This article is organized as follows. In Sections II and III we derive the DCT-based MMSE log-STSA estimators and polarity estimator, respectively. Section IV presents the objective and subjective experimental results, while a conclusion follows in Section V.

II. DERIVATION OF DCT MMSE LOG-SPECTRAL AMPLITUDE ESTIMATORS

A. SIGNAL MODEL AND NOTATION

Let the observed noisy speech of the i^{th} frame be

$$y_i(n) = x_i(n) + d_i(n), \quad 0 \leqslant n \leqslant N-1 \tag{1}$$

where $x_i(n)$, $y_i(n)$, $d_i(n)$, and N are the clean speech, noisy speech, additive noise, and the length of the observation interval in discrete-time, respectively. Let $Y(i,k) \triangleq \phi_Y(i,k)|Y(i,k)|$, $X(i,k)$, $D(i,k)$ denote the k^{th} DCT spectral coefficients of the noisy speech, the clean speech, and the noise, respectively. We assume that $X(i,k)$ and $D(i,k)$ are statistically independent with zero means. For better readability, the frame index i and the frequency index k are subsequently omitted, and a single-DCT coefficient at a given time-frequency instant is considered. We denote the modulus, $|Y|$, and signs of the DCT spectral coefficients, $\phi_Y = \text{sgn}(Y)$, as the Absolute Spectrum (AS) and Polarity Spectrum (PoS) of the DCT spectral coefficients Y, respectively (and similarly with X and D). Equation (1) can be represented in the DCT domain as:

$$\phi_Y|Y| = \phi_X|X| + \phi_D|D| \tag{2}$$

Our task is to obtain the estimator $|\widehat{X}|$, which minimizes the following distortion measure [4]:

$$E\left[\left(\log|X| - \log|\widehat{X}|\right)^2\right] \tag{3}$$

where $E[\cdot]$ denotes the expectation operator. We use capital letters and their corresponding lowercase letters to denote the random variable and its realization, respectively, and a hat symbol to denote its estimate, i.e.,$|\widehat{X}|$. From [4, eq. 3], the MMSE LSA estimator equals:

$$\begin{aligned}|\widehat{X}| &= \underset{|\widehat{X}|}{\arg\min}\, E\left[\left(\log|X| - \log|\widehat{X}|\right)^2\right] \\ &= \exp\{E\ [\ln|X| \mid Y]\}\end{aligned} \tag{4}$$

and it is independent of the basis chosen for the log in (3). The evaluation of $E\ [\ln|X||Y]$ can be obtained if we use the moment-generating function of $\ln|X|$ give Y as demonstrated in [4]

$$\begin{aligned}E\ [\ln|X||Y] &= \frac{\mathrm{d}}{\mathrm{d}t}E\left\{e^{t(\ln|X|)} \mid Y\right\}\Big|_{t=0} \\ &= \frac{\mathrm{d}}{\mathrm{d}t}E\ \left[|X|^t \mid Y\right]\Big|_{t=0}.\end{aligned} \tag{5}$$

Therefore, our task now is to compute $E\left[|X|^t \mid Y\right]$ and then attain $E\left[\ln|X||Y\right]$ by using (5). With the assumption that the DCT spectral coefficients are statistically independent, the conditional expectation $E\left[|X|^t \mid Y\right]$ is given by:

$$\begin{aligned}E\left[|X|^t \mid Y\right] &= \int_{-\infty}^{\infty} |x|^t p(x|Y)\,\mathrm{d}x \\ &= \frac{\int_{-\infty}^{\infty} |x|^t p(Y|x)p(x)\,\mathrm{d}x}{\int_{-\infty}^{\infty} p(Y|x)p(x)\,\mathrm{d}x}\end{aligned} \tag{6}$$

As discussed earlier, we assume a Gaussian distribution for the noise coefficients:

$$p(D) = \frac{1}{\sqrt{2\pi}\sigma_D}\exp\left(-\frac{D^2}{2\sigma_D^2}\right) \tag{7}$$

where $p(\cdot)$ denotes the probability density function (PDF) and σ_D^2 denotes the variance of the noise spectral coefficients. As given by (2), the noisy coefficient Y is the sum of two independent random variables, which implies that the conditional PDF of Y given X is

$$p(Y|X) = \frac{1}{\sqrt{2\pi}\sigma_D}\exp\left[\frac{-(Y-X)^2}{2\sigma_D^2}\right] \tag{8}$$

We employed either Gaussian or Laplacian speech prior to attaining the LSA estimator:

1. Motivated by the central limit theorem, the complex Gaussian PDF was used in the fundamental paper by Ephraim and Malah [4], and the real version is defined as in [26] and [19]

$$p(X) = \frac{1}{\sqrt{2\pi}\sigma_X}\exp\left(-\frac{X^2}{2\sigma_X^2}\right) \tag{9}$$

2. The suitability of the Laplacian PDF has been validated by fitting the PDFs to the empirical data. As demonstrated in [21] and [18], the Laplacian PDF provides a better fit to the empirical data than the Gaussian PDF, and it is defined as in [26] and [19]

$$p(X) = \frac{1}{\sqrt{2}\sigma_X}\exp\left(-\frac{\sqrt{2}\,|X|}{\sigma_X}\right) \tag{10}$$

where σ_X and σ_X^2 are the standard deviation and variance of the clean DCT coefficients, respectively.

Note that in statistical approaches to speech enhancement, the optimal solution can be described as a gain function multiplied by the noisy DCT STSA:

$$|\widehat{X}| \triangleq G(\cdot,\cdot)|Y| \tag{11}$$

Since the noisy PoS is the best estimate of the original PoS under the constrained MMSE criterion as proposed in [13], we can combine the enhanced AS, $|\widehat{X}|$, with the noisy PoS, ϕ_Y, to get the final estimate of the spectral component [Fig. 6 (a)]

$$\widehat{X} \cong \phi_Y|\widehat{X}| \tag{12}$$

Alternatively, we can use the proposed PoE, $\widehat{\phi}_X$, to replace the noisy PoS for reconstruction [Fig. 6 (b)]

$$\widehat{X} \cong \widehat{\phi}_X|\widehat{X}| \tag{13}$$

B. THE MMSE LSA ESTIMATOR FOR GAUSSIAN SPEECH PRIOR

For the Gaussian speech prior, the derivation of the MMSE LSA estimator of the DCT clean speech spectral amplitudes is comparable with [4]. Upon substituting (8), (9), into (6), and using relations [27, eq. 3.462.1, 9.240, 9.212.1, 8.335.1], we obtain:

$$E\left[|X|^{t} \mid Y\right] = \frac{1}{\sqrt{\pi}}(2\lambda)^{\frac{t}{2}}\Gamma\left(\frac{t}{2}+\frac{1}{2}\right)\ \Phi\left(-\frac{t}{2},\frac{1}{2};-\frac{v}{2}\right) \tag{14}$$

where $\Gamma(z)$ denotes the gamma function [27, Th.8.310.1] and $\Phi(\cdot)$ denotes the confluent hypergeometric function of the first kind [i.e., ${}_1F_1(\alpha;\beta;z)$], which is defined as in [27, eq. 9.210.1]

$$\Phi(\alpha,\beta;z) = {}_1F_1(\alpha;\beta;z) = \sum_{n=0}^{\infty}\frac{(\alpha)_n}{(\beta)_n}\cdot\frac{z^n}{n!} \tag{15}$$

where $(\alpha)_n$ is the Pochhammer symbol [28, eq. 6.1.22]

$$(\alpha)_n \triangleq \frac{\Gamma(\alpha+n)}{\Gamma(\alpha)} = \alpha(\alpha+1)\cdots(\alpha+n-1), \tag{16}$$

and $(\alpha)_0 \triangleq 1$. λ satisfies the following relation

$$\frac{1}{\lambda} = \frac{1}{\sigma_X^2}+\frac{1}{\sigma_D^2}. \tag{17}$$

Using the product rule of differentiation, the derivative of (14) with respect to t [which is needed for (5)] leads to

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}E\left[|X|^t \mid Y\right]\Big|_{t=0} &= \left[\frac{\mathrm{d}(2\lambda)^{\frac{t}{2}}}{\mathrm{d}t}\Big|_{t=0}\right] + \left[\frac{1}{\sqrt{\pi}}\frac{\mathrm{d}\Gamma\left(\frac{t}{2}+\frac{1}{2}\right)}{\mathrm{d}t}\Big|_{t=0}\right]\\ &\quad + \left[\frac{\mathrm{d}\Phi\left(-\frac{t}{2},\frac{1}{2};-\frac{v}{2}\right)}{\mathrm{d}t}\Big|_{t=0}\right]\end{aligned} \tag{18}$$

The derivative of $(2\lambda)^{\frac{t}{2}}$ at $t=0$ is given by

$$\frac{\mathrm{d}}{\mathrm{d}t}(2\lambda)^{\frac{t}{2}}\Big|_{t=0} = \frac{1}{2}\ln(2\lambda) \tag{19}$$

The derivative of $\Gamma\left(\frac{t}{2}+\frac{1}{2}\right)$ can be obtained by utilizing the Psi (Digamma) function, which is defined by [28, eq. 6.3.1]

$$\Psi(z) = \frac{\mathrm{d}}{\mathrm{d}z}\left[\ln\Gamma(z)\right] = \frac{\Gamma'(z)}{\Gamma(z)} \tag{20}$$

Rearrange (20) and use the expression given in [28, eq. 6.3.3] we attain

$$\frac{\mathrm{d}}{\mathrm{d}t}\Gamma\left(\frac{t}{2}+\frac{1}{2}\right)\Big|_{t=0} = \frac{1}{2}(-C-2\ln 2)\sqrt{\pi} \tag{21}$$

where C is the Euler's constant. Finally, to find the derivative of $\Phi\left(-\frac{t}{2},\frac{1}{2},-\frac{v}{2}\right)$ at $t=0$, we use [29, eq.38a] and find

$$\frac{\mathrm{d}\Phi\left(-\frac{t}{2},\frac{1}{2};-\frac{v}{2}\right)}{\mathrm{d}t}\Big|_{t=0} = \left(\frac{v}{2}\right){}_2F_2\left[1,1;\frac{3}{2},2;\left(-\frac{v}{2}\right)\right] \tag{22}$$

where ${}_2F_2\left[1,1;\frac{3}{2},2;\left(-\frac{v}{2}\right)\right]$ is the generalized hypergeometric function which is defined by means of a hypergeometric series [27, eq. 9.14.1]

$$\begin{aligned}&{}_pF_q(\alpha_1,\alpha_2,\ldots,\alpha_p;\beta_1,\beta_2,\ldots,\beta_q;z)\\ &\quad = \sum_{n=0}^{\infty}\frac{(\alpha_1)_n(\alpha_2)_n\cdots(\alpha_p)_n}{(\beta_1)_n(\beta_2)_n\cdots(\beta_q)_n}\cdot\frac{z^n}{n!}\end{aligned} \tag{23}$$

Nevertheless, the computation of the hypergeometric function for a wide dynamic range is not trivial, and numerical problems may result when the arguments are large. To improve numerical stability, we use the expressions [27, eq. 7.512.12, 9.212.1, 9.212.2, 9.215.1] and rewrite (22) in terms of the definite integral of combinations of exponential and algebraic functions

$$\frac{\mathrm{d}\Phi\left(-\frac{t}{2},\frac{1}{2};-\frac{v}{2}\right)}{\mathrm{d}t}\Big|_{t=0} = \frac{v}{4}\int_0^1(1-x)^{-\frac{1}{2}}\left[\frac{e^{\left(-\frac{v}{2}\right)x}-1}{\left(-\frac{v}{2}\right)x}\right]\mathrm{d}x \tag{24}$$

Now, by using (19), (21), and (24) we obtain from (18)

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}E\left[|X|^t \mid Y\right]\Big|_{t=0} &= \frac{1}{2}(\ln\lambda-\ln 2-C)\\ &\quad + \frac{v}{4}\int_0^1(1-x)^{-\frac{1}{2}}\left[\frac{e^{\left(-\frac{v}{2}\right)x}-1}{\left(-\frac{v}{2}\right)x}\right]\mathrm{d}x\end{aligned} \tag{25}$$

On substituting (25) into (5) and using (4), (17), we get the desired amplitude estimator

$$\begin{aligned}|\widehat{X}| &= \sqrt{\left(\frac{\xi}{1+\xi}\right)\frac{1}{2\exp(C)\,\gamma}}\,\exp\left[\frac{v}{4}\,\mathcal{I}(-\frac{v}{2})\right]|Y|\\ &\triangleq G_{N-LSA}(\xi,\gamma)\,|Y|\end{aligned} \tag{26}$$

where $C = 0.57721566490\ldots$ is the Euler's constant [27, eq. 8.367.1], $\mathcal{I}(\cdot)$ designates the definite integral

$$\mathcal{I}(z) \triangleq \int_0^1(1-x)^{-\frac{1}{2}}\left(\frac{e^{z\cdot x}-1}{z\cdot x}\right)\mathrm{d}x \tag{27}$$

and v is defined by

$$v \triangleq \frac{\xi}{1+\xi}\gamma\,;\qquad \xi \triangleq \frac{\sigma_X^2}{\sigma_D^2}\,;\qquad \gamma \triangleq \frac{|Y|^2}{\sigma_D^2} \tag{28}$$

The terms ξ and γ are referred to as the *a priori* and *a posteriori* signal-to-noise ratio (SNR), respectively [3]. Note that the maximum-likelihood (ML) estimate of the *a priori* SNR, i.e., $\widehat{\xi}_{ml} = \gamma - 1$ can be interpreted as an instantaneous SNR estimator of the spectral component while ξ acts as a long term estimator of the SNR.

C. THE MMSE LSA ESTIMATOR FOR LAPLACIAN SPEECH PRIOR

In [6], Hendriks et al. proposed a DFT-based LSA estimator which assumes chi-distributed speech amplitudes. However, a closed-form solution is not obtainable and numerical approximation is instead required due to the induced mathematical complexity. In contrast, we show that the DCT-based LSA estimator for Laplacian speech prior can be expressed in the closed-from. The derivation is analogous to the Gaussian case given in section II-B and the derivation given in [4].

To facilitate the development, we designate the following shorthand notations:

$$A = \frac{\sigma_D}{\sigma_X} + \frac{|y|}{\sqrt{2}\,\sigma_D} = \frac{1}{\sqrt{\xi}} + \sqrt{\frac{\gamma}{2}} \tag{29a}$$

$$B = \frac{\sigma_D}{\sigma_X} - \frac{|y|}{\sqrt{2}\,\sigma_D} = \frac{1}{\sqrt{\xi}} - \sqrt{\frac{\gamma}{2}} \tag{29b}$$

$$M = \frac{1}{\sqrt{\pi}}\left[\operatorname{erfc}(A) + e^{-2\sqrt{\frac{2\gamma}{\xi}}}\operatorname{erfc}(B)\right]^{-1} \tag{29c}$$

Substituting (8) and (10) into (6) and using [27, eq. 3.322.2, 3.462.1, 9.240], we find

$$E\left[|X|^t \mid Y\right] = M\Big\{[\mathcal{P}_1(A) - \mathcal{P}_2(A)] + e^{-2\sqrt{\frac{2\gamma}{\xi}}}[\mathcal{P}_1(B) - \mathcal{P}_2(B)]\Big\} \tag{30}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}E\left[|X|^t \mid Y\right]\Big|_{t=0} = M\left\{\left[\frac{\mathrm{d}\,\mathcal{P}_1(A)}{\mathrm{d}t} - \frac{\mathrm{d}\,\mathcal{P}_2(A)}{\mathrm{d}t}\right] + e^{-2\sqrt{\frac{2\gamma}{\xi}}}\left[\frac{\mathrm{d}\,\mathcal{P}_1(B)}{\mathrm{d}t} - \frac{\mathrm{d}\,\mathcal{P}_2(B)}{\mathrm{d}t}\right]\right\} \tag{31}$$

where

$$\mathcal{P}_1(z) = \sigma_D^t\, 2^{\frac{t}{2}}\,\Gamma(\frac{t}{2}+\frac{1}{2})\,\Phi\left(-\frac{t}{2}, \frac{1}{2}; -z^2\right) \tag{32}$$

$$\mathcal{P}_2(z) = 2z\,\sigma_D^t\, 2^{\frac{t}{2}}\,\Gamma(\frac{t}{2}+1)\,\Phi\left(\frac{1-t}{2}, \frac{3}{2}; -z^2\right) \tag{33}$$

It is easily shown that the derivative of $\mathcal{P}_1(z)$ over t yields, with the help of (21) and (24)

$$\frac{\mathrm{d}\mathcal{P}_1(z)}{\mathrm{d}t}\Big|_{t=0} = \sqrt{\pi}\left[\ln\sigma_D - \frac{C}{2} - \frac{\ln 2}{2} + \frac{z^2}{2}\mathcal{I}(-z^2)\right] \tag{34}$$

where $\mathcal{I}(\cdot)$ is defined by (27) and C is the Euler's constant [27, eq. 8.367.1]. Similarly, we differentiate each part for $\mathrm{d}\mathcal{P}_2(z)/\,\mathrm{d}t$ [see (35)], as shown at the bottom of the next page. It is known [27, eq. 9.236.1, 9.212.1] that

$$\Phi\left(\frac{1}{2}, \frac{3}{2}; -z^2\right) = \frac{\sqrt{\pi}\operatorname{erf}(z)}{2z} \tag{36}$$

$$\Phi(\alpha, \beta; z) = e^z\,\Phi(\beta-\alpha, \beta; -z) \tag{37}$$

By using (37) and [29, eq. 19, 20a], we obtain the derivative of $\Phi\left(\frac{1-t}{2}, \frac{3}{2}; -z^2\right)$ at $t = 0$

$$\frac{\mathrm{d}\Phi\left(\frac{1-t}{2}, \frac{3}{2}; -z^2\right)}{\mathrm{d}t}\Big|_{t=0} = e^{-z^2}\left[\frac{\mathrm{d}\Phi\left(\frac{t}{2}+1, \frac{3}{2}; z^2\right)}{\mathrm{d}t}\Big|_{t=0}\right]$$
$$= e^{-z^2}\times \left\{\frac{1}{2}\frac{z^2}{\left(\frac{3}{2}\right)}\sum_{r=0}^{\infty}\frac{(1)_r}{\left(\frac{5}{2}\right)_r}\frac{(z^2)^r}{r!}\,{}_2F_2\left[1, 1; 2, \left(\frac{5}{2}+r\right); z^2\right]\right\} \tag{38}$$

A numerically useful integral representation for (38) can be obtained by employing the integral representation for ${}_2F_2\left[1, 1; 2, \left(\frac{5}{2}+r\right); z^2\right]$ (see [27, eq. 7.512.12]):

$${}_2F_2\left[1, 1; 2, \left(\frac{5}{2}+r\right); z^2\right] = \frac{3}{2}\frac{\left(\frac{5}{2}\right)_r}{\left(\frac{3}{2}\right)_r}\int_0^1 (1-x)^{r+\frac{1}{2}}\,\Phi(1, 2; z^2x)\,\mathrm{d}x \tag{39}$$

with this representation the series of (38) becomes:

$$\frac{\mathrm{d}\Phi\left(\frac{1-t}{2}, \frac{3}{2}; -z^2\right)}{\mathrm{d}t}\Big|_{t=0} = e^{-z^2}\times\left\{\frac{z^2}{2}\int_0^1 (1-x)^{\frac{1}{2}}\,\Phi\left[1, \frac{3}{2}; z^2(1-x)\right]\times\Phi(1, 2; z^2x)\,\mathrm{d}x\right\} \tag{40}$$

Using the relations (36), (37), and the one corresponding to $\Phi(1, 2, z^2x)$ [27, eq. 9.212.1, 9.212.2, 9.215.1]

$$\Phi(1, 2, z^2x) = \frac{e^{(z^2x)} - 1}{z^2x} \tag{41}$$

we get, after some algebra,

$$\frac{\mathrm{d}\Phi\left(\frac{1-t}{2}, \frac{3}{2}; -z^2\right)}{\mathrm{d}t}\Big|_{t=0} = \frac{z}{4}\sqrt{\pi}\int_0^1 \operatorname{erf}(z\sqrt{1-x})\left[\frac{e^{-z^2x-1}}{-z^2x}\right]\mathrm{d}x \tag{42}$$

Combining (21), (36), (42), and (35), we arrive at

$$\frac{\mathrm{d}\mathcal{P}_2(z)}{\mathrm{d}t}\Big|_{t=0} = \sqrt{\pi}\left\{\operatorname{erfc}(z)\left(\ln\sigma_D + \frac{\ln 2}{2} - \frac{C}{2}\right) + \frac{z^2}{2}\int_0^1 \operatorname{erf}\left(z\sqrt{1-x}\right)\left[\frac{e^{-z^2x}-1}{-z^2x}\right]\mathrm{d}x\right\} \tag{43}$$

Now, by using (34) and (43), we obtain (44), as shown at the bottom of the page, from (31), where $\mathcal{G}(\cdot)$ is defined by:

$$\mathcal{G}(z) \triangleq z^2 \left\{ \mathcal{I}(-z^2) - \int_0^1 \operatorname{erf}\left(z\sqrt{1-x}\right) \left[\frac{e^{-z^2 x} - 1}{-z^2 x} \right] \mathrm{d}x \right\} - \ln(4) \tag{45}$$

where $\mathcal{I}(\cdot)$ is defined by (27) and $\operatorname{erf}(\cdot)$ denotes the error function [27, eq. 8.250.1]. On substituting (44) into (5) and using (4), we get the desired amplitude estimator

$$\begin{aligned} |\widehat{X}| = \sqrt{\frac{2}{\exp(C)\,\gamma}} \exp\Bigg[\left(\frac{1}{2}\right) \\ \times \frac{\mathcal{G}\left(\frac{1}{\sqrt{\xi}} + \sqrt{\frac{\gamma}{2}}\right) + e^{-2\sqrt{\frac{2\gamma}{\xi}}} \mathcal{G}\left(\frac{1}{\sqrt{\xi}} - \sqrt{\frac{\gamma}{2}}\right)}{\operatorname{erfc}\left(\frac{1}{\sqrt{\xi}} + \sqrt{\frac{\gamma}{2}}\right) + e^{-2\sqrt{\frac{2\gamma}{\xi}}} \operatorname{erfc}\left(\frac{1}{\sqrt{\xi}} - \sqrt{\frac{\gamma}{2}}\right)} \Bigg] |Y| \\ \triangleq G_{L-LSA}(\xi, \gamma)\,|Y| \end{aligned} \tag{46}$$

where $\operatorname{erfc}(\cdot)$ denotes the complementary error function [27, eq. 8.250.4].

D. GAIN CHARACTERISTICS OF THE PROPOSED MMSE LSA ESTIMATORS

It can be seen that the similarities in behavior between the respective gain curves of G_{N-LSA} (26), and those of the Ephraim and Malah solution [4], denoted as G_{EM-LSA} (Fig. 2). However, G_{N-LSA} always maintains a higher attenuation than the one which results from G_{EM-LSA} (Fig. 3). On the other hand, the equivalent of G_{L-LSA} (46) in the DFT domain, denoted as G_{SG-LSA} [6], has no closed-form solutions, and numerical approximation has resorted. It also appears that G_{SG-LSA} provides markedly less attenuation than the other estimators when the instantaneous SNR (ISNR) is high and ξ is low (Fig. 2). In such acoustic conditions, the amplitude of the noise component is likely to be higher than the one of the speech component, and therefore, applying G_{SG-LSA} to the noisy observation may yield a higher level of residual noise in the resultant speech. Notably, while G_{SG-LSA} offers the least attenuation, G_{L-LSA} provides the highest attenuation when the acoustic conditions are undesirable (e.g., $\xi = -15$ or $\xi = -5$ dB) and ISNR is low (Fig. 3). It implies G_{L-LSA} may offer better performance than G_{SG-LSA} in terms of noise suppression, especially for undesirable acoustic conditions. Moreover, when ISNR is large, estimators with super-Gaussian speech prior, i.e., G_{L-LSA} and G_{SG-LSA}, provide less attenuation than those ones with Gaussian speech prior, i.e., G_{N-LSA} or G_{EM-LSA} (Fig. 3). This is because super-Gaussian distributions are leptokurtic (e.g., heavy-tailed) distributions, in which high observed noisy amplitudes (i.e., high ISNRs) are considered more likely to contain clean speech components than in the Gaussian model [15]. Thus, these estimators with super-Gaussian prior justifiably gain more success in recovering the speech spectral peaks and thereby reduce the amount of the perceived speech distortion.

Fig. 4 shows the gain curves of the new estimators along with the corresponding gain curves which result from the MMSE STSA estimators derived in [19]. The behavior of these gain curves is similar to the new gain curves. However, the new gain function [which results from (26) or (46)] always gives a lower gain than the corresponding one which results from the estimator of [19]. These lower gain values imply the new estimators may reduce the residual noise level further than those estimators in [19], particularly in regions of low ISNR values.

III. DERIVATION OF POLARITY ESTIMATOR (PoE)

Our goal is to derive an explicit relationship between the noisy polarity Φ_Y and the clean polarity Φ_X giving the input parameters, i.e., the instantaneous (spectral) SNR ξ_I [30] and the instantaneous *a posteriori* SNR γ_I [31]:

$$\xi_I \triangleq \frac{|X|^2}{|D|^2} \tag{47}$$

$$c\gamma_I \triangleq \frac{|Y|^2}{|D|^2} \tag{48}$$

To achieve this, we obtain from (2)

$$\frac{\Phi_Y}{\Phi_X}\frac{|Y|}{|D|} = \frac{|X|}{|D|} + \frac{\Phi_D}{\Phi_X} \tag{49}$$

$$\begin{aligned} \left.\frac{\mathrm{d}\mathcal{P}_2(z)}{\mathrm{d}t}\right|_{t=0} = 2z \Bigg\{ \left[\left.\frac{\mathrm{d}\,\sigma_D^t}{\mathrm{d}t}\right|_{t=0} \times \Phi\left(\frac{1}{2}, \frac{3}{2}; -z^2\right) \right] + \left[\left.\frac{\mathrm{d}\,2^{\frac{t}{2}}}{\mathrm{d}t}\right|_{t=0} \times \Phi\left(\frac{1}{2}, \frac{3}{2}; -z^2\right) \right] \\ + \left[\left.\frac{\mathrm{d}\Gamma\left(\frac{t}{2} + \frac{1}{2}\right)}{\mathrm{d}t}\right|_{t=0} \times \Phi\left(\frac{1}{2}, \frac{3}{2}; -z^2\right) \right] + \left[\left.\frac{\mathrm{d}\Phi\left(\frac{1-t}{2}, \frac{3}{2}; -z^2\right)}{\mathrm{d}t}\right|_{t=0} \right] \Bigg\} \end{aligned} \tag{35}$$

$$\begin{aligned} \left.\frac{\mathrm{d}}{\mathrm{d}t} E\left[|X|^t \mid Y\right]\right|_{t=0} &= \left(\ln \sigma_D + \frac{\ln 2}{2} - \frac{C}{2}\right) + \frac{M}{2}\left\{\mathcal{G}(A) + e^{-2\sqrt{\frac{2\gamma}{\xi}}}\mathcal{G}(B)\right\} \\ &= \left(\ln |Y| + \frac{1}{2}\ln\frac{2}{\gamma} - \frac{C}{2}\right) + \frac{M}{2}\left\{\mathcal{G}(A) + e^{-2\sqrt{\frac{2\gamma}{\xi}}}\mathcal{G}(B)\right\} \end{aligned} \tag{44}$$

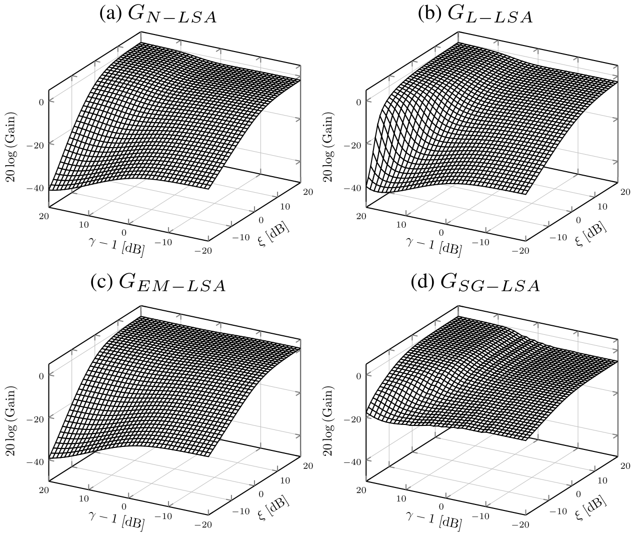

FIGURE 2. Gain curves plotted against the *a priori* SNR ξ and the instantaneous SNR $\gamma - 1$ for the MMSE LSA estimators. (a) G_{N-LSA} defined by (26), (b) G_{L-LSA} defined by (46), (c) the respective Ephraim and Malah solution, G_{EM-LSA}, as seen in [4, (30)], and (d) G_{SG-LSA} with $v = 0.3$ and $L = 15000$, as seen in [6].

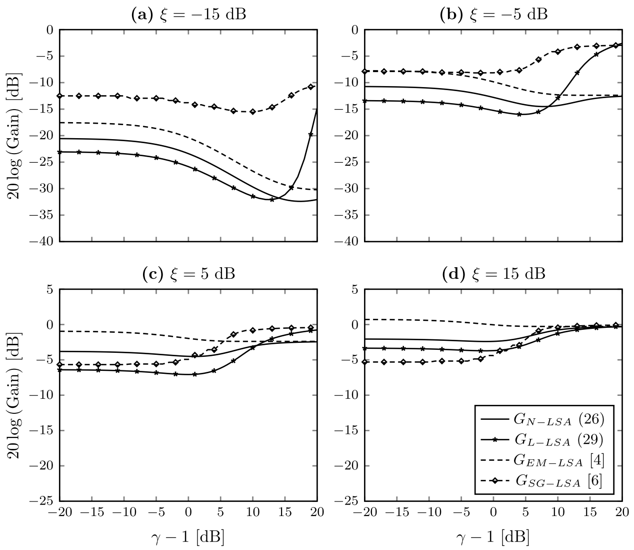

FIGURE 3. Gain curves comparison for the proposed MMSE LSA estimators for (a) $\xi = -15$ dB, (b) $\xi = -5$ dB, (c) $\xi = 5$ dB, and (d) $\xi = 15$ dB. The solid and solid-dotted lines correspond to G_{N-DCT} and G_{L-DCT} defined by (26) and (46), respectively. The corresponding curves for the Ephraim and Malah solution [4] G_{N-DFT} (complex Gaussian prior), and the Hendriks et al. solution [6] G_{SG-DFT} (with super-Gaussian prior, $v = 0.3$ and $L = 15000$), are indicated with dashed and dash-dotted lines respectively for reference.

We next consider four different cases:

- **Case 1**: $\Phi_X = \Phi_Y$, and $\Phi_X = \Phi_D$
  In this case, (49) becomes $\frac{|Y|}{|D|} = \frac{|X|}{|D|} + 1$, hence, $\frac{|Y|}{|D|} \geqslant 1$.
- **Case 2**: $\Phi_X = \Phi_Y$, and $\Phi_X = -\Phi_D$
  In this case, (49) becomes $\frac{|Y|}{|D|} = \frac{|X|}{|D|} - 1$. Since $\frac{|Y|}{|D|} \geqslant 0$, implying that $\frac{|X|}{|D|} \geqslant 1$
- **Case 3**: $\Phi_X = -\Phi_Y$, and $\Phi_X = \Phi_D$
  In this case, (49) becomes $-\frac{|Y|}{|D|} = \frac{|X|}{|D|} + 1$. Since $-\frac{|Y|}{|D|} \leqslant 0$ contradicts $\frac{|X|}{|D|} + 1 \geqslant 1$, and hence no solution exits.

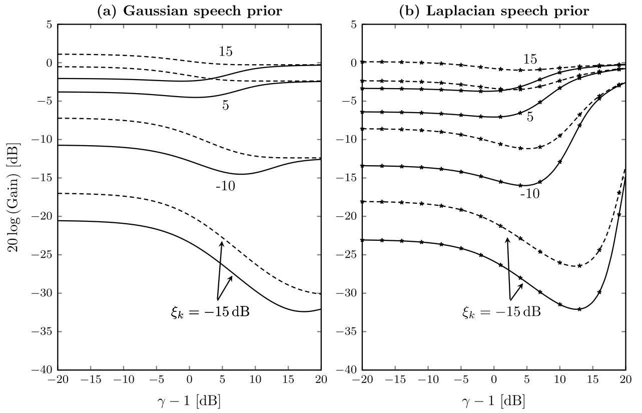

FIGURE 4. Gain curves comparison. (a) Gaussian speech prior: solid line, G_{N-DCT} (26); dashed line, the corresponding gain curves which result from the MMSE STSA estimator ([19], formula (15)). (b) Laplacian speech prior: solid-dotted line, G_{L-DCT} (46); dash-dotted line, the corresponding curves which result from MMSE STSA estimator ([19], formula (22)).

- **Case 4**: $\Phi_X = -\Phi_Y$, and $\Phi_X = -\Phi_D$
  In this case, (49) becomes $-\frac{|Y|}{|D|} = \frac{|X|}{|D|} - 1$, which implies $\frac{|X|}{|D|} < 1$

Consolidating **Case 1** to **Case 4**, we obtain that $\frac{|Y|}{|D|} = \sqrt{\gamma_I} \geqslant 1$ and $\frac{|X|}{|D|} = \sqrt{\xi_I} \geqslant 1$ are sufficient conditions for $\Phi_X = \Phi_Y$; However, for $\Phi_X = -\Phi_Y$, if and only if $\sqrt{\xi_I} < 1$ holds (Algorithm 1).

Algorithm 1 Polarity Estimation

```
if $\sqrt{\gamma_I} \geqslant 1$ then
    $\Phi_X = \Phi_Y$;
else
    if $\sqrt{\xi_I} \geqslant 1$ then
        $\Phi_X = \Phi_Y$;
    else
        $\Phi_X = -\Phi_Y$;
```

Due to the deterministic nature of the algorithm, when oracle information is given, meaning $|X|$ and $|D|$ in (47) are known as a prior, perfect recovery of the clean polarity is guaranteed (Fig. 5, dashed-line). In practice, estimates of $|X|$ and $|D|$ are used, and hence the accuracy of the algorithm might be strongly affected by the reliability of these estimates. In fact, the dominant influence on the accuracy appears to be the reliability of the noise estimate. Experiment results show that when speech STSA estimate [obtained by using (46)] and oracle noise estimate (51) were used, the accuracy dropped slightly from 100% to around 95%. However, when both parameters were blind-estimated, accuracy declines significantly (Fig. 5, dotted line). In the experiment, we find $\widehat{\gamma_I}$ using $\widehat{\gamma_I} \approx |Y|^2/\widehat{\sigma}_D^2$, where $\widehat{\sigma}_D^2$ is the noise variance estimate and can be attained using the estimator given in [19]. We then find $\widehat{\xi_I}$ using $\widehat{\xi_I} \approx |\widehat{X}|^2/\widehat{\sigma}_D^2$, which requires estimates of speech STSA $|\widehat{X}|$ and $\widehat{\sigma}_D^2$ as parameters. The proposed algorithm can be efficiently implemented using

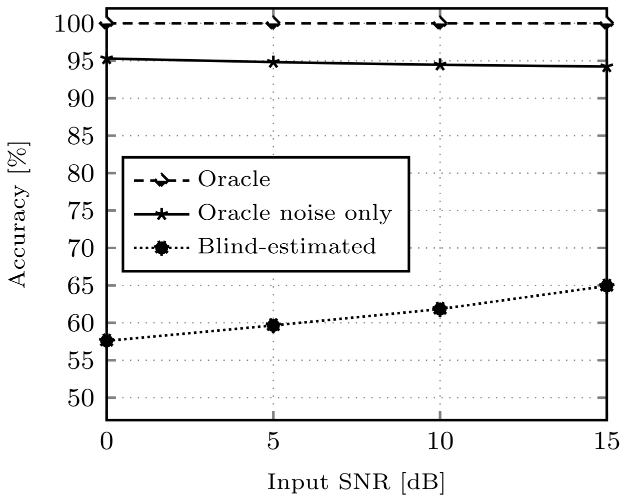

FIGURE 5. **The effect of parameter estimation on the accuracy of the polarity estimator. Dashed line: oracle information was given, meaning $|X|$ and $|D|$ in (47) were known as prior, yielding an accuracy of %100. Solid line: clean STSA estimate $|\widehat{X}|$ from (46) and oracle noise estimate from (51) were used. Dotted line: both $|X|$ and $|D|$ were blind-estimated, where noise estimate was obtained from the estimator given in [19]. Results were averaged over seven noise types (white noise, pink noise, speech noise, voice babble noise, F-16 noise, car factory noise, and car Volvo-340 noise).**

matrix operations. A Matlab pseudocode for computing $\widehat{\phi}_X$ is given in Appendix A.

Given that the performance of the polarity estimator might be strongly affected by the noise estimation, in the next section we examine the effect of polarity estimation on the perceived quality of an enhanced speech signal.

IV. IMPLEMENTATION AND PERFORMANCE EVALUATION

In this section, we evaluate the performance of proposed estimators on enhancing noisy speech. In order to draw a complete and accurate conclusion, it requires to test under all of the noise conditions, existing methods, and simulation conditions. As we intend these results to be illustrative rather than exhaustive, we limit our simulation conditions to specific noise types (stationary and non-stationary), and typical speech enhancement framework setups.

A. SPEECH CORPUS

For the evaluation of our approach, we used 40 gender-balanced utterances from the TSP speech database [32]. 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 7 kinds of additive noise. They were white noise, pink noise, speech noise, voice babble noise, F-16 noise, car factory noise, and car Volvo-340 noise from the RSG-10 database [33], the last five being real-world non-stationary noise types. After combination with the clean speech utterances from above, 40 × 7 = 280 noisy speech utterances were obtained. Each evaluation was repeated for 0, 5, 10, and 15 dB SNR conditions, respectively.

B. EXPERIMENT SETUP

For the sake of a fair performance evaluation, all the comparative STSA estimators use the same or equivalent basic setup, meaning that the analysis-modification-synthesis (AMS) setup, the *a prior* SNR estimation, and the noise PSD estimation (see Section IV-D) are equal for all methods. The experiment parameters are not optimized for any of the presented methods. They are based on heuristic knowledge and are widely accepted in literature.

The noisy corpus is processed through the AMS framework to obtain the estimate $\widehat{x}(n)$ of the underlying clean speech signal $x(n)$ (Fig. 6). At the analysis stage, the original signal is segmented into 32ms frames (512 samples length), each with an overlap of 75%.[2] The 512-point DCT is then applied to decompose the framed noisy speech into its spectral components. At the modification stage, a gain function as given in (26) or (46) is applied to each component independently. To study the effects of polarity modification on the enhanced speech, either the noisy polarity [Fig. 6 (a)] or the estimate of the clean polarity [Fig. 6 (b)] is used for reconstruction. The enhanced spectral components are then synthesized by means of inverse transform and overlap adding [34]. The Hamming window [35] is employed for both analysis and synthesis. Similar procedures were carried out when implementing the DFT-based methods with DCT and DFT the only difference and all other factors being equivalent.

The gain value can be obtained by exact calculation or by using look-up tables indexed by ξ and γ values. However, these values are in general not known a priori, they have to be estimated from the noisy observations as well. We employ an MMSE-based noise power estimator, described first in [37] for DFT and the modified version in [19] for DCT, to determine the variance of the noise samples. A "decision-directed" (DD) approach [3] is used to estimate the *a priori* SNR of the speech samples on a frame-to-frame basis:

$$\widehat{\xi}(i) = \max\left\{\alpha_n \frac{|\widehat{X}(i-1)|^2}{\widehat{\sigma}_D^2(i-1)} \right.$$
$$\left. + (1-\alpha_n)\max\left[\gamma(i)-1\,,0\right]\,,\xi_{min}\right\} \quad (50)$$

where $|\widehat{X}(i-1)|$ and $\widehat{\sigma}_D^2(i-1)$ are the estimates of the spectral amplitude and the noise variance in the past frame, respectively. The max{·} operator denotes the maximum function to ensure the positiveness of the estimator, while $\alpha_n = 0.98$ (as determined by simulations and informal listening tests in [3]) is the smoothing factor and $\xi_{min} = -25$ dB is the SNR floor value for eliminating low-level musical noise [38]. In all

[2] In speech processing, a window duration between 20-40 ms is typically used, so that the properties of the signal do not change appreciably, and the windows must overlap by 75% to avoid aliasing [36].

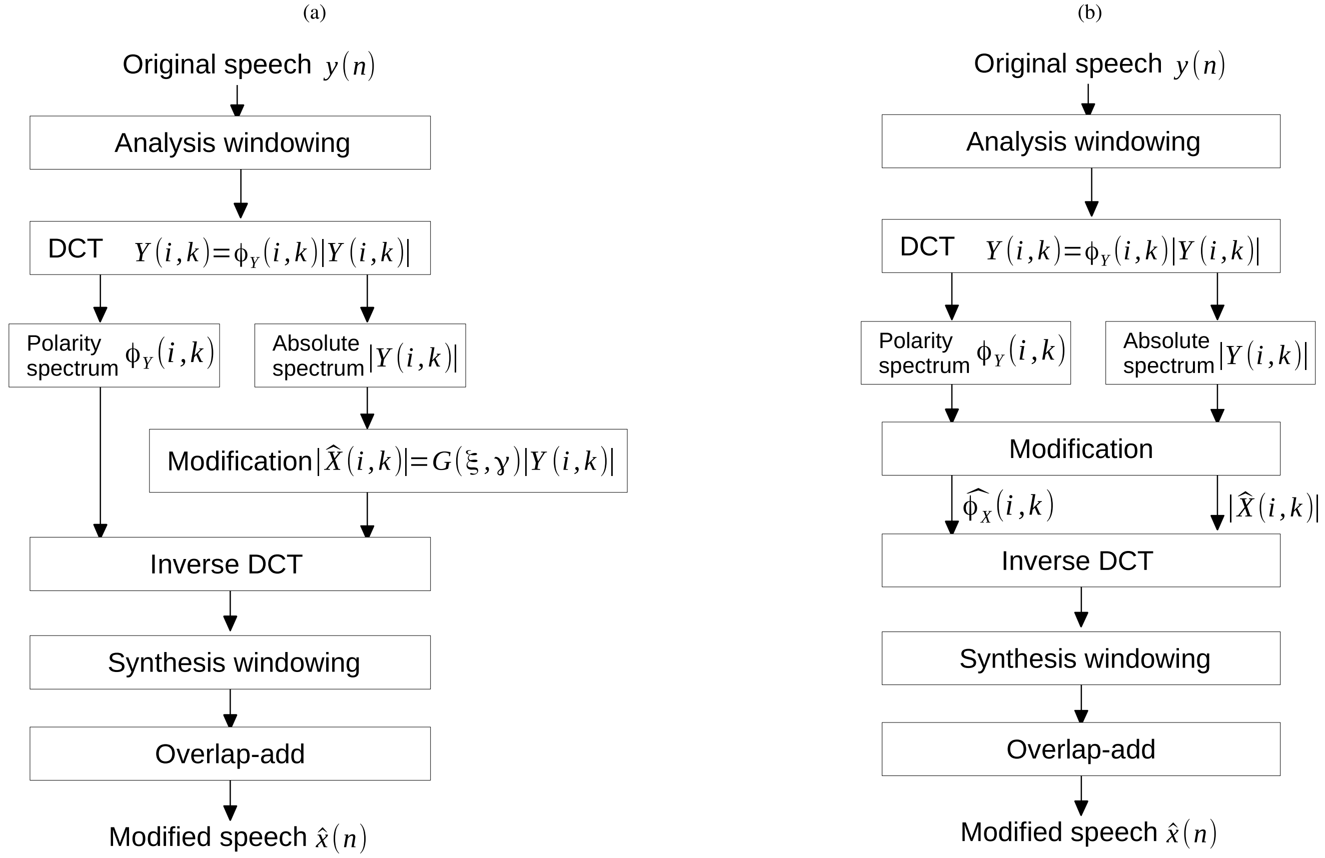
(a)
Original speech $y(n)$
Analysis windowing
DCT $Y(i,k)=\phi_Y(i,k)|Y(i,k)|$
Polarity spectrum $\phi_Y(i,k)$
Absolute spectrum $|Y(i,k)|$
Modification $|\hat{X}(i,k)|=G(\xi,\gamma)|Y(i,k)|$
Inverse DCT
Synthesis windowing
Overlap-add
Modified speech $\hat{x}(n)$
(b)
Original speech $y(n)$
Analysis windowing
DCT $Y(i,k)=\phi_Y(i,k)|Y(i,k)|$
Polarity spectrum $\phi_Y(i,k)$
Absolute spectrum $|Y(i,k)|$
Modification
$\hat{\phi}_X(i,k)$
$|\hat{X}(i,k)|$
Inverse DCT
Synthesis windowing
Overlap-add
Modified speech $\hat{x}(n)$

FIGURE 6. **Block diagrams of the analysis-modification-synthesis (AMS) procedure used in the experiments. (a) The conventional single-channel speech enhancement procedure, where the modified DCT Absolute Spectrum (AS) and the noisy Polarity Spectrum (PoS) are used for reconstruction. (b) Alternatively, both the AS and PoS are modified and used for reconstruction. The DCT representation of the noisy signal is given by $\phi_Y|Y|$. The estimates of clean AS and PoS are denoted by $|\hat{X}|$ and $\hat{\phi}_X$, respectively.**

experiments, the gain value was limited to 0.1, for perceptual reasons [18].

The Matlab software, version R2019b, was used for the experiments. The audio examples, along with the completed scripts for implementing the new estimators, as well as further experimentation and comparison with other methods, are available online at [39].

C. OBJECTIVE QUALITY AND INTELLIGIBILITY MEASURES

The performance was quantified in terms of (i) the average Segmental SNR (SegSNR) [40], which is a local SNR computed over short segments; (ii) wideband perceptual evaluation of speech quality (PESQ) [41], which is an objective score for assessing speech quality in wideband telecommunication networks; (iii) the short-time objective intelligibility (STOI) improvements [42], which has been shown to highly correlate with the intelligibility scores obtained through listening tests, and (iv) the phase deviation (PD) [43], which is a distortion metric between the noisy phase and clean phase. For SegSNR, PESQ, and STOI, we report the improvement (or gain) over the noisy input instead of the absolute values.

It should be noted that SegSNR and PESQ are conventional instrumental measures where the focus is on the spectral amplitude distortion, and hence no phase distortion is taken into account. In particular, it was reported that PESQ might overestimate the quality for methods using phase modification, where spurious harmonics are introduced leading to a buzzy quality [44]. In contrast, PD penalizes these harmonization artifacts by predicting a worse quality (note that the lower the PD score the better the estimated quality).

D. ORACLE AND BLIND NOISE PSD ESTIMATES

To examine the influence of noise estimation accuracy on the performance of the proposed estimators, we first run a set of experiments using an oracle noise estimator, which is computed as:

$$\hat{\sigma}_D^2 = |D|^2 \tag{51}$$

where $|D|^2$ is the periodogram of the noise signal. The above noise estimator was used to isolate the effect of a noise estimation algorithm. We run the second set of experiments using the noise estimator proposed in [19] and [37] for the DCT-based algorithms and the DFT-based algorithms, respectively.

E. SUBJECTIVE TESTING PROCEDURE

Subjective evaluation was carried out through a series of blind AB listening tests to obtain an accurate estimate of the perceived speech quality. The AB listening test has been widely used for speech enhancement applications, e.g., [45], [46], and [47], and its procedure has been described in Appendix B. Two utterances from the test set are used as the clean speech stimuli: sentence 3 from list 33, as uttered by male speaker MF, and sentence 4 from list 53, as uttered by female speaker FI. To produce the noisy speech stimuli, speech, and F-16 noise were mixed with the clean speech stimuli from speaker MF and FI, respectively, at an SNR level of 5 dB. The enhanced speech stimuli for each of the speech enhancement methods were produced from the noisy speech stimuli. For each utterance, all possible stimuli pair combinations were presented to the listener. A total of five English-speaking listeners (with normal hearing capability) participated. The average of the scores given by the listeners, termed as mean subjective preference (%) score, was used as an indicator for the perceived speech quality.

All procedures of listening tests were performed under the approval of Griffith University Human Research Ethics: database protocol number 2018/671.

F. SPECIFICATIONS OF COMPETITIVE METHODS

For benchmarking, we included the following algorithms in our evaluation (Table 2). To examine the effectiveness of the polarity estimator (PoE), we either combine the enhanced absolute spectrum with the noisy polarity spectrum (cases **N-LSA** and **L-LSA**) or combine the blind-estimated polarity spectrum (cases **N-PoE** and **L-PoE**) for speech reconstruction (Fig. 6). We also use the oracle polarity spectrum estimate (cases **N-PoE-O** and **L-PoE-O**) to set the upper bound on the maximum improvements achievable for the polarity estimation. Specifically, methods (1)-(3) are DFT-based algorithms while the rest are DCT-based; methods (1)-(2), (4)-(6) modify the STSA only and left the noisy polarity (or phase) intact; while the rest modify both of the STSA and spectral polarity (or phase).

By using super-Gaussian prior, the complexity of **SG-LSA** and **L-LSA** are much higher than those ones based on Gaussian prior and thereby require higher computation costs. To circumvent this issue, the gain values of **SG-LSA** and **L-LSA** were computed with high precision [48] and tabulated in look-up tables. During run-time, these tables use a pair of ξ and γ values as the index to retrieve the corresponding gain values.

G. RESULTS

The objective results were first averaged across all the utterances for a compact and general comparison (Fig. 7). For illustrative purposes, we also present the results for F-16, speech, and white noises separately (Fig. 10) under the fully blind experiment setup. Multiple comparison statistical tests were conducted according to Tukey's honestly significant difference (HSD) test [50] to assess significant differences between algorithms. Differences between scores were deemed significant if the obtained p value (level of significance) was smaller than 0.05.

TABLE 2. Specifications of the competitive methods.

	Methods	Description
(1)	**EM-LSA** [4]	Ephraim and Malah DFT MMSE LSA estimator with complex Gaussian speech prior.
(2)	**SG-LSA** [6]	Hendriks et al. DFT MMSE LSA estimator with super-Gaussian speech prior. The parameters $v = 0.3$ and $L = 15000$ were set as in [6], and the MATLAB implementation of the algorithm is available at [49].
(3)	β-**PC** [25]	Perceptual weighted β-order STSA estimator with phase compensation (PC) procedure and complex Gaussian speech prior.
(4)	**L-STSA** [19]	DCT MMSE STSA estimator with Laplacian speech prior.
(5)	**N-LSA**	DCT MMSE LSA estimator with Gaussian speech prior [proposed, Eq. (26)].
(6)	**L-LSA**	DCT MMSE LSA estimator with Laplacian speech prior [proposed, Eq. (46)].
(7)	**N-PoE**	**N-LSA** estimator + blind polarity estimation [proposed, Algor. 1].
(8)	**L-PoE**	**L-LSA** estimator + blind polarity estimation [proposed, Algor. 1].
(9)	**N-PoE-O**	**N-LSA** estimator + oracle polarity estimation.
(10)	**L-PoE-O**	**L-LSA** estimator + oracle polarity estimation.

1) OBJECTIVE RESULTS WITH THE ORACLE NOISE ESTIMATOR

With the oracle noise estimator (Fig. 7, left column), it shows **L-PoE-O** yields the highest PESQ, STOI, and SegSNR gains, as well as the lowest phase distortion (PD). When compared to the upper-bound performance achieved by **L-PoE-O**, there is only a marginal decrease in perceived quality for **L-PoE**, which yields the second highest score for all metrics. The difference in performances between **L-PoE-O** and **L-PoE** was not found to be statistically significant. This is due to the dominant influence of noise estimate on the accuracy of the PoE. With the oracle noise estimator, the accuracy has only slightly dropped to around 95% (Fig. 5, solid line), and thereby most of the clean polarity information has been recovered. Similar observations can be made between **N-PoE-O** and **N-PoE** with the Gaussian speech prior.

It also shows that the new estimators in conjunction with the PoE, e.g., **L-PoE** and **N-PoE**, scored significantly higher than those in conjunction with the noisy PoS, e.g., **L-LSA** and **N-LSA**, across most conditions. This is the result of the PoE being able to retrieve clean polarity information, which contributes to the improvement of speech quality. This effect is particularly noticeable for low SNR conditions. For instance, as the SNR decreases from 5 to 0 dB, the differences between the **L-PoE** and **L-LSA** in terms of PD, STOI, and SegSNR score have increased from 0.10 to 0.16, 1.61% to 3.25% and 1.14 dB to 1.41 dB, respectively. A similar

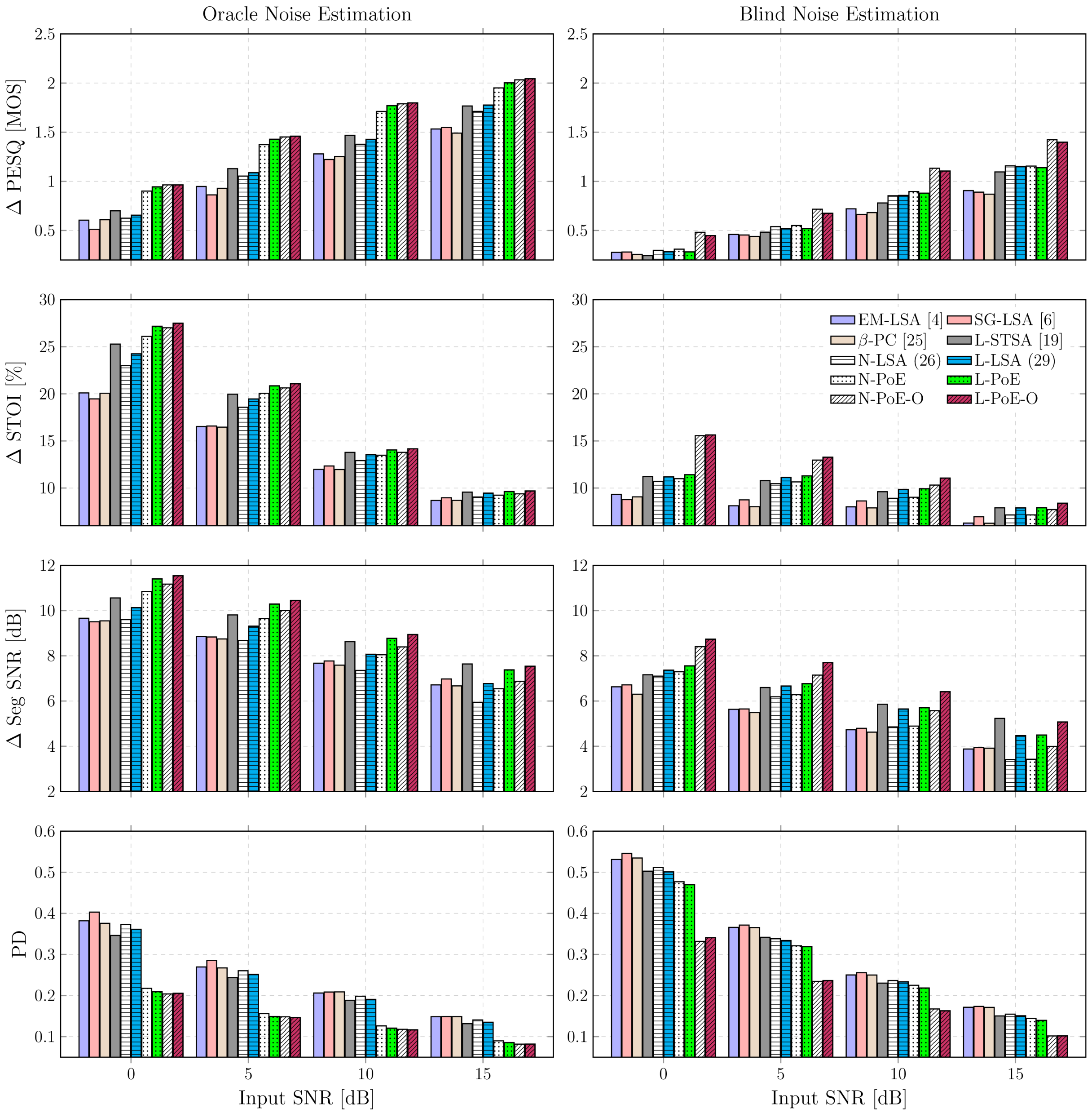

FIGURE 7. Performance comparison among various estimators (Sec. IV-F) tested using the oracle noise estimator (left column) and blind noise estimation given the noisy speech (right column). Results are shown in terms of PESQ, STOI, Segmental SNR improvements, and Phase deviation (PD) score. Results are averaged over seven noise types (white noise, pink noise, speech noise, voice babble noise, F-16 noise, car factory noise, and car Volvo-340 noise). The decision-directed approach [3] was used for the *a priori* SNR estimation.

trend can be seen between **N-PoE** and **N-LSA**. Furthermore, **L-PoE** (or **N-PoE**) on average achieved a 0.4 higher PESQ score than **L-LSA** (or **N-LSA**), for all tested SNR conditions.

Regarding the speech priors, we note that estimators with Laplacian prior, e.g., **L-PoE** and **L-LSA**, generally performed better than those with Gaussian prior, e.g., **N-PoE** and **N-LSA**. Notably, the former group performed significantly better in terms of SegSNR (i.e., residual noise level) and STOI (0 dB and 5 dB) than the latter. This is probably a result of lower gain, particularly in regions of low ISNR values, which results in lower residual noise level; and higher gain at high ISNR values, which reduces the amount of speech distortion (Fig. 3, solid-dotted line). However, in terms of PESQ, STOI (10 dB and 15 dB), and PD scores, there was no statistically significant difference between the two groups.

It is interesting to see that **L-STSA** obtained slightly higher objective scores than **L-LSA**. This is probably because **L-STSA** offers less attenuation at low ISNRs [Fig. 4 (b), dashed-dotted line], which preserves a few more speech spectral components, at the expense of a larger number of spurious spectral peaks. Nevertheless, these spurious peaks contribute more to the musical character of the residual noise rather than to the perceived speech quality, as indicated by the listening tests (see Sec. IV-G4). On the other hand, **L-LSA** suppresses some of the weaker spectral components, but at the same

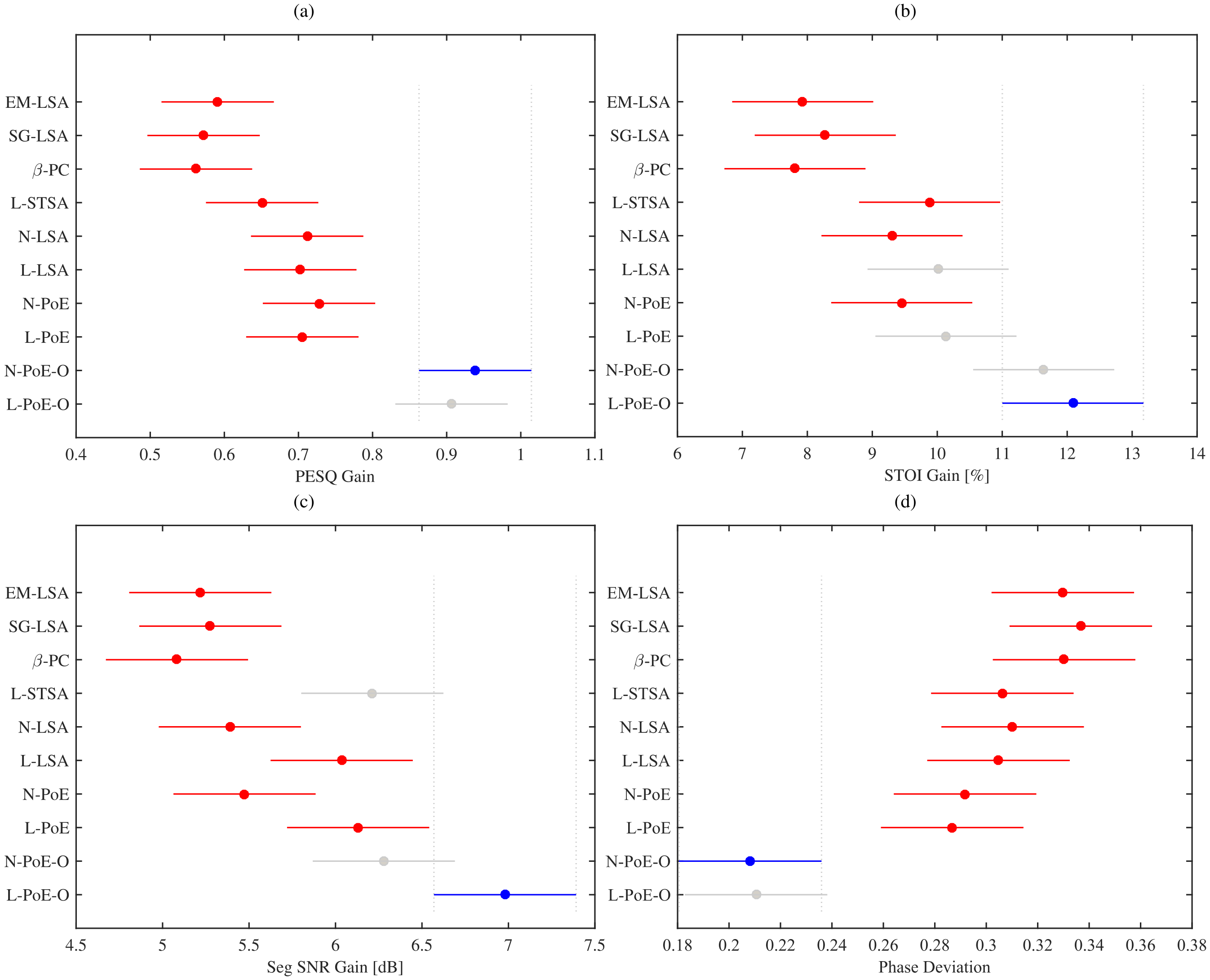

FIGURE 8. **Results obtained from comparative statistical analysis of (a) PESQ, (b) STOI, (c) SegSNR gains, and (d) Phase deviation for the full blind scenario (Table 3). Multiple-paired comparisons (Tukey's HSD) were conducted to assess significant differences between algorithms. Differences between scores were deemed significant if the obtained *p* value (level of significance) was smaller than 0.05. The means and the comparison intervals are represented by the circles and the bars, respectively. The algorithm with the highest score averaged across all noise conditions is highlighted in blue. The red bar indicates the statistically significant difference between the algorithm with the highest score and the denoted algorithm. Algorithms that do not have significantly different scores appear in grey.**

time, the fewer spurious spectral peaks reduce the amount of speech distortion [Fig. 4 (b), solid-dotted line]. However, the differences in performance were not statistically significant.

More importantly, it can be seen that the DCT-based algorithms, e.g., **L-PoE**, **L-LSA**, and **N-LSA**, generally score higher than their DFT-based counterparts, e.g., β-**PC**, **SG-LSA**, and **EM-LSA**, for all SNRs (except at 10-15 dB, where **N-LSA** obtained the lowest SegSNR gains). This is mainly because PoS has a higher JND of perception in noise distortion and preserves speech quality better than the PhS (Fig. 1).

2) OBJECTIVE RESULTS WITH THE NOMINATED NOISE ESTIMATOR

Experiment results for the blind case are reported in Table 3 with 95% confidence intervals. With the nominated noise estimator (Fig. 7, right column), **N-PoE-O** yields the highest PESQ and PD gains (except at 10 dB SNR, where **L-PoE-O** gives the best PD score). While **L-PoE-O** yields the highest STOI and SegSNR gains (except at 15 dB SNR, where **L-STSA** gives the highest SegSNR gain). The statistical analysis results (Fig. 8) show that incorporating accurate polarity estimation in the STSA estimator, e.g., **L-PoE-O**, can potentially improve the performance significantly. We also notice **L-PoE** no longer significantly improves the speech quality when compared to the upper-bound performance given by **L-PoE-O**. This is because the accuracy of the PoE has declined to around 60% (Fig. 5, dotted line). Similar observations can be made for **N-PoE-O** and **N-PoE** with Gaussian speech prior.

Despite the massive decline of PoE accuracy, **L-PoE** has slightly higher PESQ, STOI, and SegSNR gains than **L-LSA**

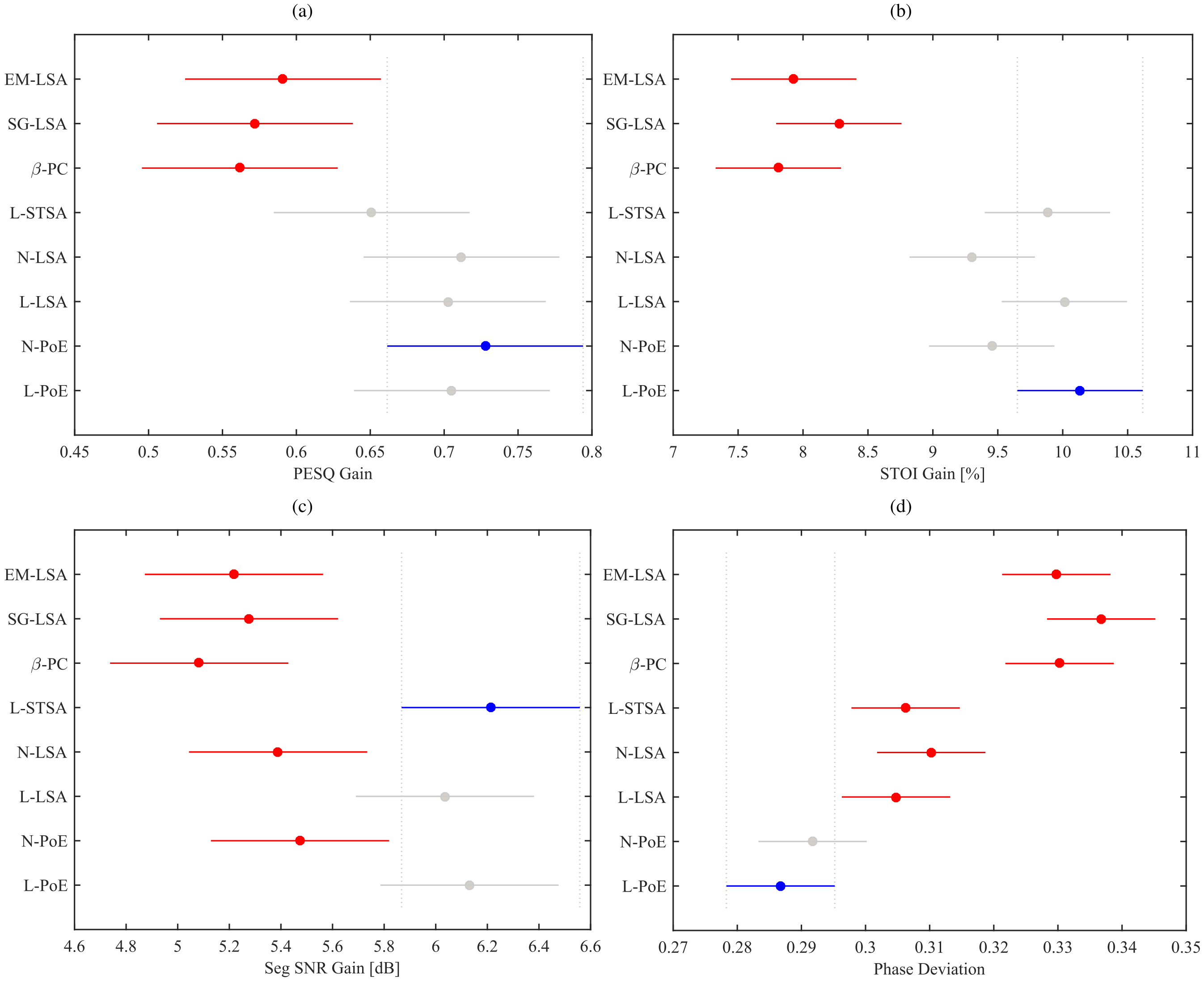

FIGURE 9. **Results obtained from comparative statistical analysis of (a) PESQ, (b) STOI, (c) SegSNR gains, and (d) Phase deviation for the full blind scenario (Table 3). Note that estimators with oracle polarity estimation (i.e., L-PoE-O and N-PoE-O) are excluded from the statistical test. Multiple-paired comparisons (Tukey's HSD) were conducted to assess significant differences between algorithms. Differences between scores were deemed significant if the obtained *p* value (level of significance) was smaller than 0.05. The means and the comparison intervals are represented by the circles and the bars, respectively. The algorithm with the highest score averaged across all noise conditions is highlighted in blue. The red bar indicates the statistically significant difference between the algorithm with the highest score and the denoted algorithm. Algorithms that do not have significantly different scores appear in grey.**

for all tested SNRs (except at 15 dB SNR, **L-LSA** has higher PESQ gain, see Table 3). Results obtained from comparative statistical analysis show there were no statistically significant improvements between **L-PoE** and **L-LSA** in terms of these objective measures (Fig. 9a-9c). Nevertheless, **L-PoE** consistently attained significantly better PD scores than **L-LSA** (Fig. 9d), which predicts improvement in speech intelligibility. Similar observations can be made for **N-PoE** and **N-LSA** with Gaussian speech prior. This result signifies that using the PoE to replace the noisy PoS for reconstruction is still beneficial, given that the accuracy of the PoE is above 60%. Considering the recent advances in noise estimation, the accuracy of the PoE can be improved by utilizing a more accurate noise estimator such as [51]. Consequently, the performance gap between the oracle case, e.g., **L-PoE-O**, and the blind case, e.g., **L-PoE**, can be effectively reduced.

Compared to the phase-aware STSA estimator β**-PC** [25], the proposed method **L-PoE**, utilizing Laplacian prior and blind polarity estimate, leads to an average improvement of up to 0.18 in PESQ and 1.05 dB in SegSNR for perceived speech quality, as well as 2.32% in STOI and 0.04 in PD scores for speech intelligibility (over 4 SNR conditions and 7 noise types, Table 3). Figure 9 shows β**-PC** performed significantly worse than the proposed methods, e.g., **L-PoE** and **L-LSA**. This was surprising at first, but a close analysis indicated that β**-PC** was sensitive to the accuracy of the phase estimate. Furthermore, in [25] the β**-PC** algorithm used a more advanced *a prior* estimator, and hence, the experimental results reported in [25] do not necessarily represent a fair

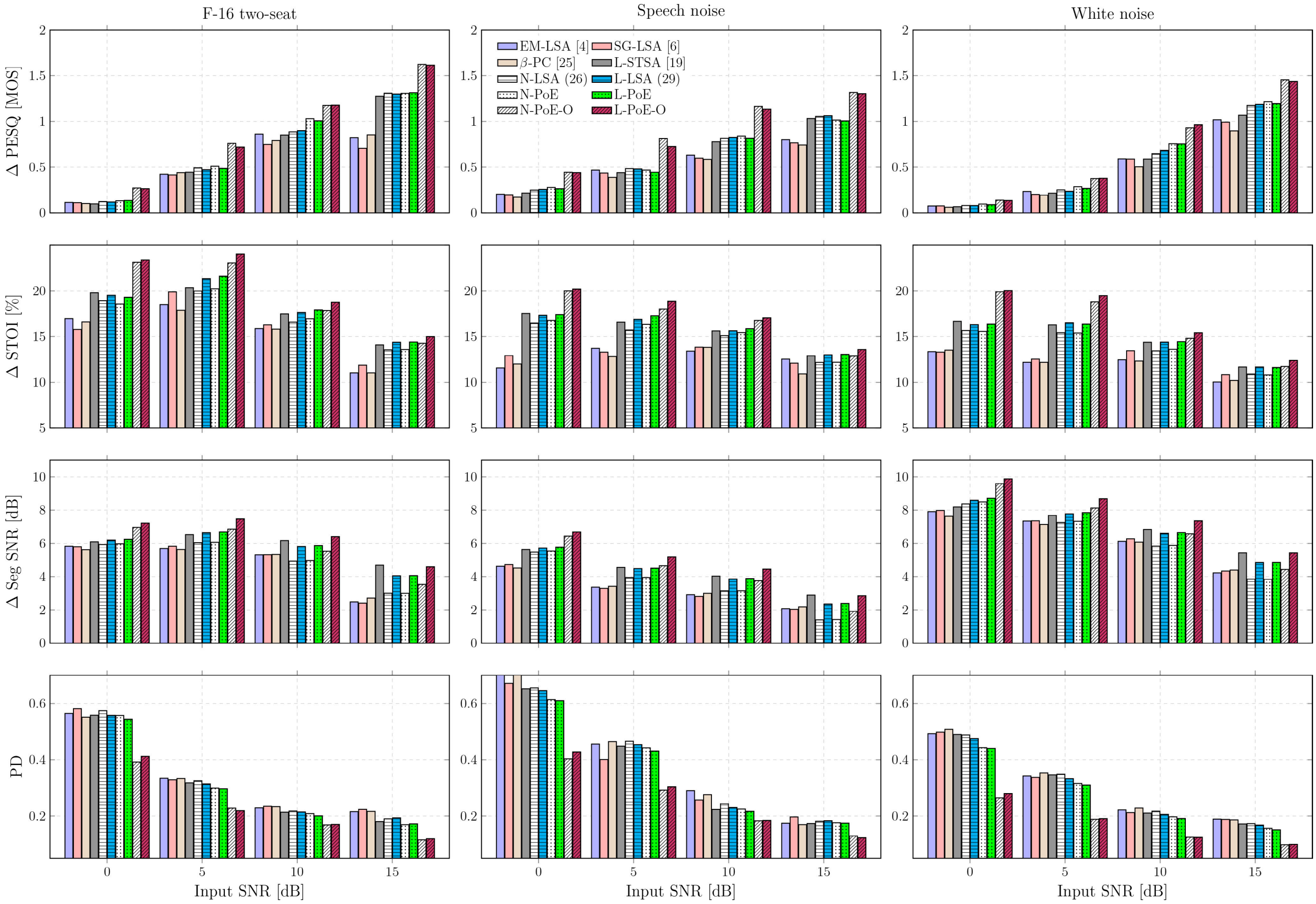

FIGURE 10. Performance comparison in terms of PESQ, STOI, Segmental SNR improvements, and Phase deviation (PD) score, between various estimators. Results are illustrated for three noise types: F-16 noise (left column), speech noise (middle column), and white noise (right column). The nominated MMSE noise estimator introduced in [19] and [37] was used for the DCT-based methods and DFT-based methods, respectively. The decision-directed approach [3] was used for the *a priori* SNR estimation.

performance comparison. Similar to the oracle case, in the blind case the DCT-based algorithms performed consistently better than the DFT-based algorithms in terms of overall quality, across all conditions (except at 15 dB **N-LSA** and **N-PoE** obtained the lowest SegSNR gains). Figure 9 also shows that there was no single best algorithm, but rather that several algorithms performed equally well across all conditions.

For illustrative purposes, we also present the results for the F-16, the speech, and the white noise separately (Fig. 10). We observe that the SegSNR gains were higher for the white noise because the non-stationary F-16 and the speech noise (average long-term speech spectrum) are harder to track by the noise estimation algorithm. Speech background noise is a particularly tough condition for speech enhancement since it exhibits characteristics similar to the target speech. For the speech noise, the PESQ gains (at 10 and 15 dB) were generally lower, and the PD scores (at 0 dB and 5 dB) were generally higher than those obtained for the other two noises. Furthermore, the majority of the algorithms provided higher STOI gains for the F-16 noise than for the other two noises. Finally, for the F-16 and the white noise, **L-PoE** (**N-PoE**) obtained higher PESQ gains than **L-LSA** (**N-LSA**) at all SNR levels and the improvement was significant at 10 dB SNR; whereas for the speech noise, it was the opposite at 5 and 15 dB. In this case, there was no statistically significant difference between **L-PoE** (**N-PoE**) and **L-LSA** (**N-LSA**) across all SNR conditions.

3) SPECTROGRAM ANALYSIS

The enhanced speech spectrograms produced by various speech enhancement algorithms are also analyzed (Fig. 11 and 12). Notably, the proposed estimators incorporating oracle polarity information restore the lower frequency regions of speech onsets very well [pictured in (k)-(l) in Fig. 11 and 12]. The speech onsets (i.e., transients, highlighted by the dashed circles) are known to have the highest contribution to speech intelligibility [44]. This impact on speech intelligibility of the reconstructed speech signal has been captured by the instrumental measures as well (Fig. 10). Moreover, DCT-based estimators are able to reduce more residual noise with a less or equal amount of speech distortion than their DFT-based counterparts. For instance, **L-LSA** can preserve the formant peaks better in low-frequency bands than **SG-LSA** (highlighted by the solid circles in Fig. 11). As a result, the DCT-based estimators give better perceived speech quality since the speech components with weak energies are

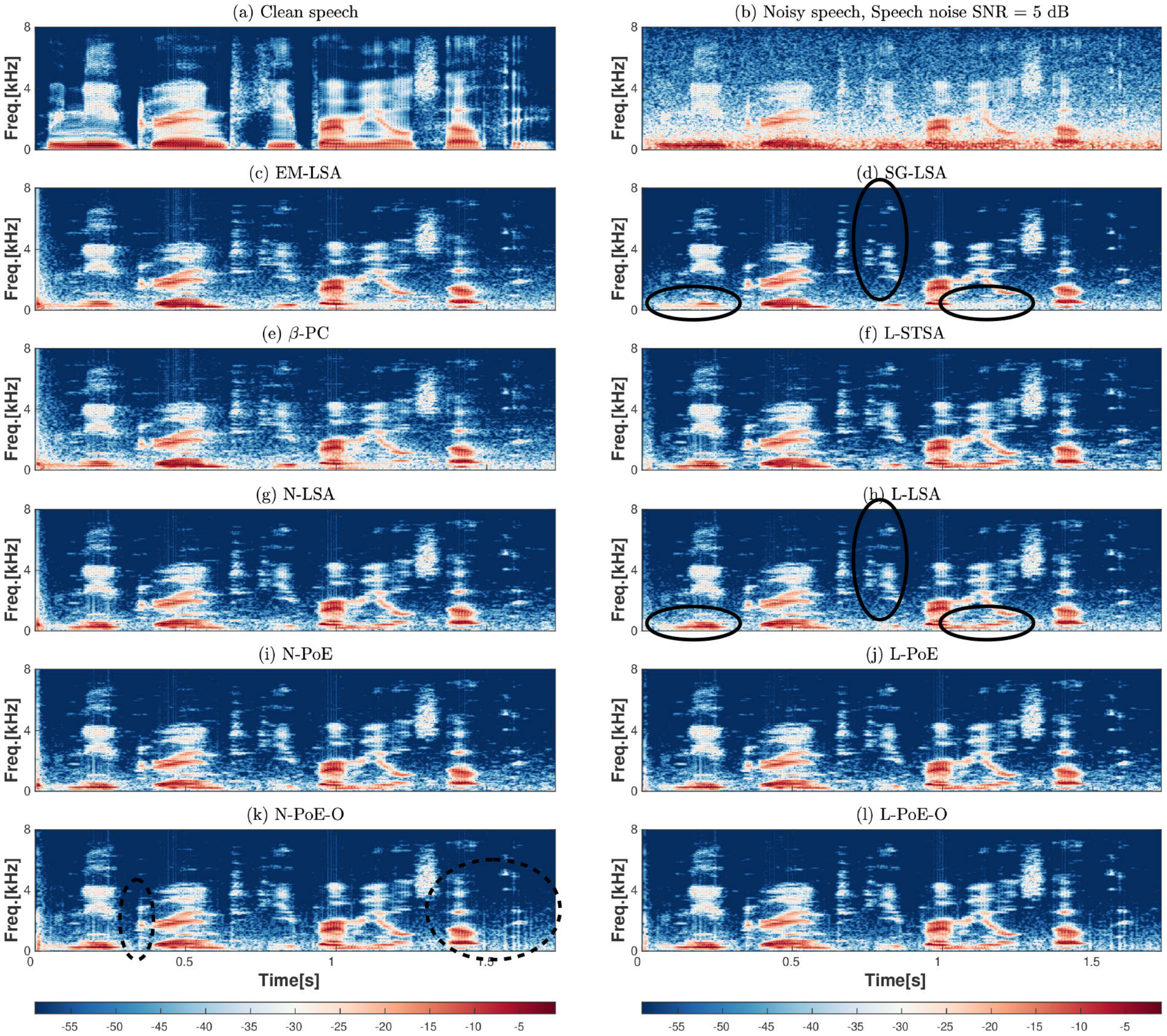

FIGURE 11. Spectrograms of (a) the clean sentence, (b) the sentence corrupted by speech noise at 5 dB, and (c)-(l) the enhanced speech produced by the corresponding speech enhancement algorithm (see Section IV-F). The sentence 'We need grain to keep our mules healthy.' (utterance MF33_03), was taken from the TSP speech database [32]. The nominated MMSE noise estimator introduced in [19] and [37] was used for the DCT-based methods and the DFT-based methods, respectively. The decision-directed approach [3] was used for the *a priori* SNR estimation.

retained and yield a smaller amount of speech distortion. Therefore, speech intelligibility is improved (predicted by the STOI and PD scores in Fig. 10). Note that Fig. 11 and 12 are representative images for most utterances, which show similar characteristics.

4) SUBJECTIVE TEST RESULTS

Finally, the subjective test results for two non-stationary noise conditions (i.e., speech noise and F-16 noise) are reported in Fig. 13 and 14. The human listening tests can reliably quantify the character of speech quality, or estimate the speech quality achievable by an algorithm. It reveals that **L-PoE-O** and **N-PoE-O** were widely preferred by the listeners over other methods, apart from the clean speech. The enhanced speech obtained by **L-PoE-O** and **N-PoE-O** suffers much less residual noise, while no difference in the speech itself was noticed.

L-PoE, utilizing the Laplacian prior and blind polarity estimate, is found to be the next most preferred method, followed by **N-PoE**, **N-LSA**, and **L-LSA**. Although the PESQ, STOI, and SegSNR scores of **L-PoE** and **L-LSA** are very similar, it was reported that **L-PoE** appears to have less residual noise and speech distortion than **L-LSA**. This result further highlights the effect of polarity estimation on the perceived speech quality. The utterance modified by **β-PC**, which utilize the phase compensation technique, were much less preferred by the listeners than those enhanced by the proposed methods. The listeners also reported that the phase-aware enhanced speech suffers from some reverberations, resulting in garbled noise. These artifacts can be predicated as a degraded perceived speech quality (e.g., PESQ and SegSNR) or intelligibility score (e.g., STOI and PD) as seen in Fig. 10. The listening test results show that incorporating an erroneous phase estimate can strongly influence the performance of

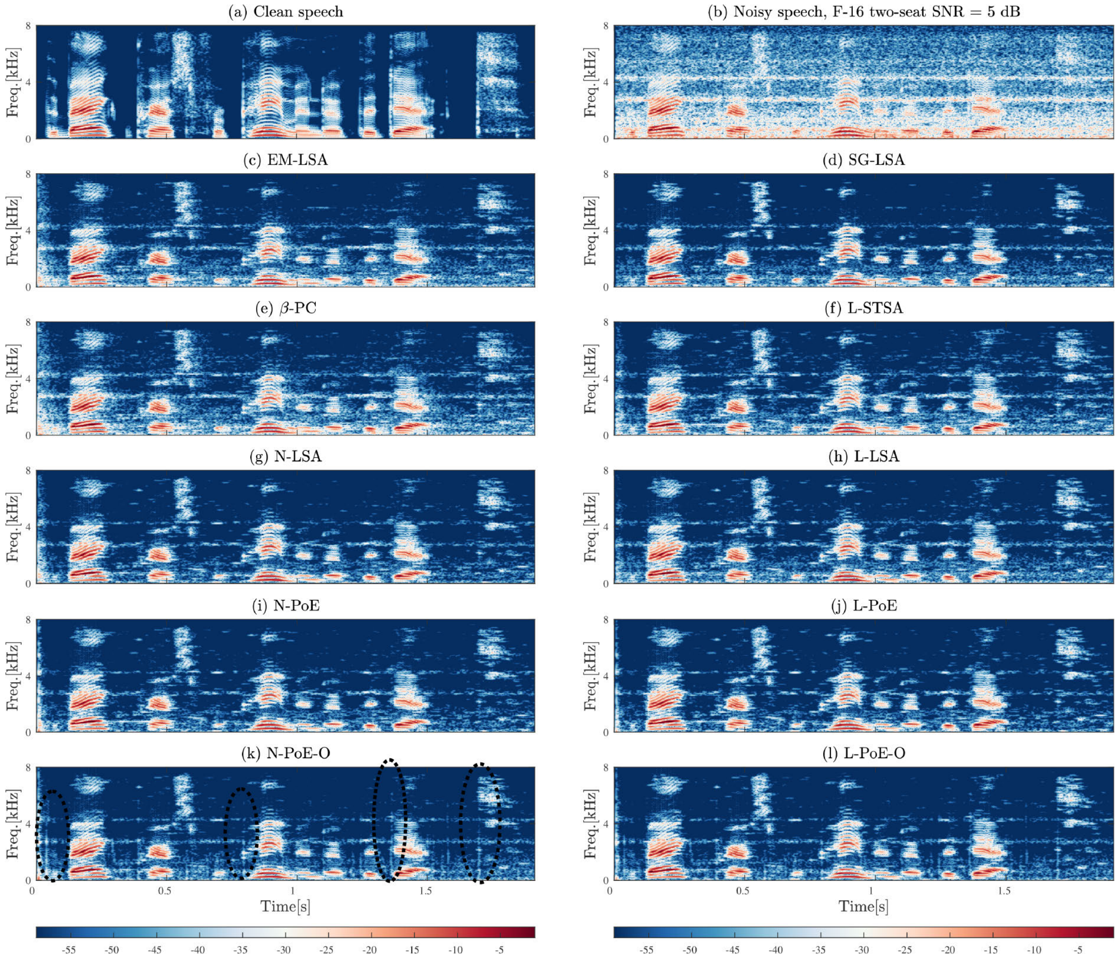

FIGURE 12. Spectrograms of (a) the clean sentence, (b) the sentence corrupted by non-stationary F-16 noise at 5 dB, and (c)-(l) the enhanced speech produced by the corresponding speech enhancement algorithm (see Section IV-F). Note that the perceivable differences between the various estimators are generally small since the maximum suppression was limited to 0.1 for all methods. The sentence 'The bank pressed for payment of the debt.' (utterance FI53_04), was taken from the TSP speech database [32]. The nominated MMSE noise estimator introduced in [19] and [37] was used for the DCT-based methods and the DFT-based methods, respectively. The decision-directed approach [3] was used for the *a priori* SNR estimation.

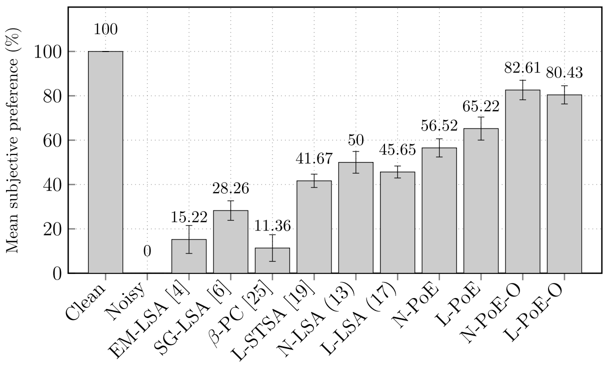

FIGURE 13. The mean subjective preference score (%) comparison for each speech enhancement method. The male utterance (MF33_03) corrupted with 5 dB speech noise (averaged long-term speech spectrum) was used for the subjective tests. The error bars indicated the standard deviation of the scores.

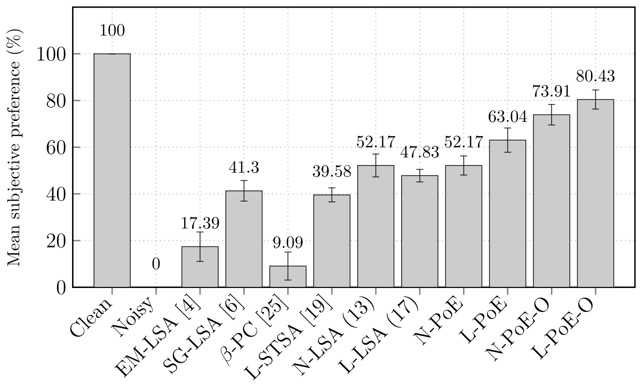

FIGURE 14. The mean subjective preference score (%) comparison for each speech enhancement method. The female utterance (FI53_04) corrupted with 5 dB non-stationary F-16 noise was used for the subjective tests. The error bars indicated the standard deviation of the scores.

phase-aware estimators; however, a much less accurate polarity estimate doesn't lead to a noticeable decrease in perceived quality and can potentially improve speech intelligibility.

Another interesting comparison is that of **L-STSA** and **L-LSA**. We found that the enhanced speech obtained by both estimators sounds very similar, with the exception that with the first estimator, the residual noise sounds a little more

TABLE 3. Performance comparison, in terms of PESQ, STOI, SegSNR, and PD improvements (lower the PD score the better-estimated quality), between various estimators tested using the nominated *a priori* SNR and MMSE noise estimators. These scores are averaged over all noise types for the fully blind scenario and reported with 95% confidence intervals. For each SNR and metric, the best performance is highlighted by boldface letters.

Noise	Method	0dB	5dB	10dB	15dB
PESQ [MOS]	**EM-LSA**	0.278	0.460	0.720	0.906
	SG-LSA	0.281	0.454	0.663	0.890
	β-PC	0.256	0.440	0.682	0.869
	L-STSA	0.245	0.483	0.780	1.096
	N-LSA	0.297	0.539	0.853	1.158
	L-LSA	0.283	0.520	0.856	1.151
	N-PoE	0.310	0.550	0.896	1.155
	L-PoE	0.283	0.520	0.879	1.139
	N-PoE-O	**0.481**	**0.716**	**1.133**	**1.424**
	L-PoE-O	0.448	0.675	1.105	1.398
STOI [%]	**EM-LSA**	9.318	8.114	8.005	6.278
	SG-LSA	8.778	8.746	8.629	6.951
	β-PC	9.061	8.018	7.896	6.263
	L-STSA	11.225	10.793	9.609	7.901
	N-LSA	10.701	10.462	8.908	7.142
	L-LSA	11.189	11.125	9.843	7.893
	N-PoE	10.994	10.651	9.019	7.149
	L-PoE	11.413	11.289	9.926	7.906
	N-PoE-O	15.563	12.966	10.314	7.711
	L-PoE-O	**15.636**	**13.271**	**11.057**	**8.387**
SegSNR [dB]	**EM-LSA**	6.630	5.634	4.731	3.877
	SG-LSA	6.717	5.649	4.793	3.946
	β-PC	6.301	5.495	4.624	3.913
	L-STSA	7.164	6.598	5.857	**5.233**
	N-LSA	7.102	6.189	4.854	3.411
	L-LSA	7.365	6.663	5.646	4.469
	N-PoE	7.297	6.284	4.891	3.424
	L-PoE	7.555	6.768	5.705	4.496
	N-PoE-O	8.406	7.148	5.574	3.990
	L-PoE-O	**8.736**	**7.701**	**6.411**	5.073
PD	**EM-LSA**	0.531	0.366	0.250	0.172
	SG-LSA	0.546	0.371	0.256	0.174
	β-PC	0.535	0.365	0.250	0.171
	L-STSA	0.503	0.342	0.230	0.150
	N-LSA	0.512	0.338	0.236	0.155
	L-LSA	0.501	0.334	0.233	0.151
	N-PoE	0.477	0.321	0.225	0.144
	L-PoE	0.470	0.319	0.218	0.140
	N-PoE-O	**0.332**	**0.234**	0.167	**0.102**
	L-PoE-O	0.341	0.236	**0.163**	**0.102**

musical (less uniform). The **L-LSA** results in lower residual noise levels than **L-STSA**, although the latter is slightly more successful in recovering the weaker speech spectral components. Note that the perceivable differences between the various estimators are generally small since the maximum suppression was limited to 0.1 for all methods.

V. CONCLUSION

In this article, we demonstrate the advantage of DCT representation and derive STSA estimators which minimize the mean-square error of the log-spectra. A novel polarity estimator (PoE) is also derived to assess the usefulness of polarity estimation for improved speech enhancement. Along with the proposed STSA estimator, the PoE has been incorporated into the speech enhancement system at the speech resynthesis (reconstruction) stage. When the PoE is used to replace the noisy polarity spectrum (PoS), it shows consistent improvement in perceived speech quality and intelligibility. For comparison purposes, we also include the oracle scenario, where the oracle clean polarity spectrum is used for reconstruction together with the enhanced absolute spectrum. Our objective and subjective results show that accurate PoS estimates have the potential to significantly improve speech enhancement performance. This outcome highlights the usefulness of clean polarity information in signal reconstruction and is interpreted as the upper-bound performance for the polarity spectrum estimation.

Our results also show that the dominant influence on the performance of the PoE is the reliability of the noise estimate and thus, with an accurate noise estimator, the PoS can be sufficiently recovered. When the oracle noise information and blind-estimated STSA are used, the PoE recovers around 95% of clean PoS. Moreover, when the accuracy drops to around 60%, using the PoE does not lead to a noticeable decrease in perceived quality and potentially improves speech intelligibility, as indicated by listening tests. Comparing the outcome of oracle experiments to those of blind experiments, we observe that using the oracle noise estimator results in considerable improvements relative to the blind case. Thus, we believe the proposed algorithms can still benefit from more precise noise estimates.

The proposed methods demonstrate superior performance in enhancing noisy speech, compared with their counterparts based on the DFT. This should be attributed to the fact that the DCT polarity spectrum has a higher JND of perception in noise distortion and preserves speech quality better than the DFT phase spectrum. Compared to the state-of-the-art DFT-based phase-aware system [25], the proposed method utilizing Laplacian prior and PoE, leads to an average improvement of up to 0.18 in PESQ and 1.05 dB in SegSNR for perceived speech quality, as well as 2.32% in STOI and 0.04 in PD scores for speech intelligibility in a blind speech

Algorithm 2 Matlab Implementation for Computing $\widehat{\Phi}_X$ Given Φ_Y, $\widehat{\xi}_I$, and $\widehat{\gamma}_I$

```
Input: Noisy polarity spectrogram $\Phi_Y \in \mathrm{R}^{L\times N}$;
       $\widehat{\xi}_I \in \mathrm{R}^{L\times N}$; $\widehat{\gamma}_I \in \mathrm{R}^{L\times N}$
Output: Clean polarity spectrogram estimate
       $\widehat{\Phi}_X \in \mathrm{R}^{L\times N}$
Function PoE($\Phi_Y$, $\widehat{\xi}_I$, $\widehat{\gamma}_I$)
    $\widehat{\Phi}_X = \Phi_Y$;
    gamma_logic = sqrt($\widehat{\gamma}_I$) > 1;
    xi_logic = sqrt($\widehat{\xi}_I$) > 1;
    step_1 = gamma_logic == *false*;
    step_2 = step_1 & xi_logic;
    step_3 = gamma_logic | step_2;
    $\widehat{\Phi}_X$(~step_3) = $\widehat{\Phi}_X$(~step_3)*(−1);
    return $\widehat{\Phi}_X$
```

enhancement setup. Our subjective results confirmed better performance in quality and intelligibility.

The possibilities for incorporating the additional clean polarity information are many. We have used independently obtained STSA and PoS estimates for speech reconstruction. Therefore, the proposed estimators do not comprise any polarity information in their derivation. Future work will be towards incorporating polarity information in STSA estimation and developing a more precise DCT-based noise estimator to exploit the full potential of DCT-based STSA estimation.

APPENDIX A
MATLAB PREUDOCODE for IMPLEMENTING the POLARITY ESTIMATOR

See Algorithm 2.

APPENDIX B
SUBJECTIVE TESTING PROCEDURE

In this appendix, we describe the procedure used to obtain the subjective quality scores in Fig. 13 and 14. These tests were done in the form of AB listening tests [52], in which listeners were asked to select a preferred stimulus for each stimuli pair. The listeners were presented with three labeled options after listening to each stimuli pair. The first and second options were used to indicate a preference for the corresponding stimulus, while the third option was used to indicate that the stimuli sounded the same. Pair-wise scoring was employed, with a score of +1 awarded to the preferred version and +0 to the other. For a similar preference response, both were awarded a score of +0.5. The participants were allowed to re-listen to stimuli if required. Five English speakers participated in all the subjective experiments. In the main listening tests, one clean stimulus was always paired with a modified stimulus. Each stimuli pair occurred twice in the playlist as the order of the stimuli pair was switched. This avoided any bias associated with listening order. In each test, stimuli pairs were played back to the participants in randomized order.

Two utterances (one from a male speaker and one from a female speaker) from the test set described in Sec. IV-A were used. Each utterance was modified as described in Sec. IV-E for the required SNR. Thus, a total of 132 modified utterances were generated for the subjective test, and since each stimuli pair was also played in reverse order, each participant scored 264 stimuli pairs. Each listening test is conducted in a separate session, in a quiet room using closed circumaural headphones at a comfortable listening level.

REFERENCES

[1] P. C. Loizou, "Introduction," in *Speech Enhancement: Theory and Practice*, 2nd ed. Boca Raton, NW, USA: CRC Press, 2013, ch. 1, pp. 1–2.

[2] P. C. Loizou, "Statistical-model-based methods," in *Speech Enhancement: Theory and Practice*, 2nd ed. Boca Raton, NW, USA: CRC Press, 2013, ch. 7, pp. 209–263.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech Signal Process.*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 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 Trans. Audio, Speech, Language Process.*, vol. 15, no. 6, pp. 1741–1752, Aug. 2007.

[6] R. C. Hendriks, R. Heusdens, and J. Jensen, "Log-spectral magnitude MMSE estimators under super-Gaussian densities," in *Proc. 10th Annu. Conf. Int. Speech Commun. Assoc.*, 2009, pp. 1–4.

[7] D. Wang and J. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-30, no. 4, pp. 679–681, Aug. 1982.

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

[9] K. K. Paliwal and L. Alsteris, "Usefulness of phase spectrum in human speech perception," in *Proc. 8th Eur. Conf. Speech Commun. Technol. (Eurospeech)*, Sep. 2003, pp. 1–4.

[10] L. D. Alsteris and K. 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.

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

[12] R. Chappel, B. Schwerin, and K. Paliwal, "Phase distortion resulting in a just noticeable difference in the perceived quality of speech," *Speech Commun.*, vol. 81, pp. 138–147, Jul. 2016.

[13] S. Shi, A. Busch, K. Paliwal, and T. Fickenscher, "On the use of discrete cosine transform polarity spectrum in speech enhancement," in *Proc. 28th Eur. Signal Process. Conf. (EUSIPCO)*, Jan. 2021, pp. 421–425.

[14] A. Papoulis and U. Pillai, *Probability, Random Variables and Stochastic Processes*, 4th ed. New York, NY, USA: McGraw-Hill, Nov. 2001.

[15] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP J. Adv. Signal Process.*, vol. 2005, no. 7, Dec. 2005, Art. no. 354850.

[16] B. Chen and P. C. Loizou, "A Laplacian-based MMSE estimator for speech enhancement," *Speech Commun.*, vol. 49, no. 2, pp. 134–143, 2007.

[17] I. Andrianakis and P. R. White, "Speech spectral amplitude estimators using optimally shaped gamma and chi priors," *Speech Commun.*, vol. 51, no. 1, pp. 1–14, Jan. 2009.

[18] R. Martin, "Speech enhancement based on minimum mean-square error estimation and SuperGaussian priors," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 845–856, Sep. 2005.

[19] S. Shi, K. Paliwal, and A. Busch, "On DCT-based MMSE estimation of short time spectral amplitude for single-channel speech enhancement," *Appl. Acoust.*, vol. 202, Jan. 2023, Art. no. 109134. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003682X22005084

[20] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. London, U.K.: Pearson, 2006.

[21] S. Gazor and W. Zhang, "Speech probability distribution," *IEEE Signal Process. Lett.*, vol. 10, no. 7, pp. 204–207, Jul. 2003.

[22] A. Aroudi, H. Veisi, H. Sameti, and Z. Mafakheri, "Speech signal modeling using multivariate distributions," *EURASIP J. Audio, Speech, Music Process.*, vol. 2015, no. 1, pp. 1–14, Dec. 2015.

[23] A. Aroudi, H. Veisi, and H. Sameti, "Speech enhancement based on hidden Markov model with discrete cosine transform coefficients using Laplace and Gaussian distributions," in *Proc. 11th Int. Conf. Inf. Sci., Signal Process. Their Appl. (ISSPA)*, Jul. 2012, pp. 304–309.

[24] A. Aroudi, H. Veisi, and H. Sameti, "Hidden Markov model-based speech enhancement using multivariate Laplace and Gaussian distributions," *IET Signal Process.*, vol. 9, no. 2, pp. 177–185, 2015.

[25] N. Saleem, M. I. Khattak, A. Nawaz, F. Umer, and M. K. Ochani, "Perceptually weighted β-order spectral amplitude Bayesian estimator for phase compensated speech enhancement," *Appl. Acoust.*, vol. 178, Jul. 2021, Art. no. 108007.

[26] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions*. New York, NY, USA: Wiley, 1994.

[27] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*. New York, NY, USA: Academic, 2007.

[28] M. Abramowitz, I. A. Stgun, and R. H. Romer, *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*. MD, USA: American Association of Physics Teachers, 1988.

[29] L. U. Ancarani and G. Gasaneo, "Derivatives of any order of the confluent hypergeometric function ${}_1F_1(a, b, z)$ with respect to the parameter a or b," *J. Math. Phys.*, vol. 49, no. 6, 2008, Art. no. 063508.

[30] P. C. Loizou, "Spectral-subtractive algorithms," in *Speech Enhancement: Theory and Practice*, 2nd ed. Boca Raton, NW, USA: CRC Press, 2013, ch. 5, p. 103.

[31] P. C. Loizou, "Spectral-subtractive algorithms," in *Speech Enhancement: Theory and Practice*, 2nd ed. Boca Raton, NW, USA: CRC Press, 2013, ch. 5, p. 111.

[32] P. Kabal, *TSP Speech Database*, Database Version, vol. 1. Montreal, QC, Canada: McGill Univ., 2002, pp. 2–9.

[33] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, Jul. 1993.

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

[35] P. C. Loizou, "Discrete-time signal processing and short-time Fourier analysis," in *Speech Enhancement: Theory and Practice*, 2nd ed. Boca Raton, NW, USA: CRC Press, 2013, ch. 2, p. 29.

[36] L. R. Rabiner and R. W. Schafer, *Theory and Applications of Digital Speech Processing*, vol. 64. Upper Saddle River, NJ, USA: Pearson, 2011, ch. 7.

[37] T. Gerkmann and R. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[38] I. Cohen, "Relaxed statistical model for speech enhancement and a priori SNR estimation," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 870–881, Sep. 2005.

[39] S. Shi. *MATLAB Implementation for DCT-Based LSA Estimation and Spectral Polarity Estimation*. Accessed: Mar. 13, 2023. [Online]. Available: https://github.com/SisiShi18/DCT_MMSE_LSA_EST

[40] P. C. Loizou, "Objective quality and intelligibility measures," in *Speech Enhancement: Theory and Practice*, 2nd ed. Boca Raton, NW, USA: CRC Press, 2013, ch. 11, p. 480.

[41] P. C. Loizou, "Objective quality and intelligibility measures," in *Speech Enhancement: Theory and Practice*, 2nd ed. Boca Raton, NW, USA: CRC press, 2013, ch. 11, pp. 502–503.

[42] 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 Process.*, vol. 19, no. 7, pp. 2125–2136, Feb. 2011.

[43] A. Gaich and P. Mowlaee, "On speech quality estimation of phase-aware single-channel speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 216–220.

[44] P. Mowlaee, J. Kulmer, J. Stahl, and F. Mayer, *Single Channel Phase-Aware Signal Processing in Speech Communication: Theory and Practice*. Hoboken, NJ, USA: Wiley, 2016.

[45] P. C. Loizou, "Evaluating performance of speech enhancement algorithms: Listening tests," in *Speech Enhancement: Theory and Practice*, 2nd ed. Boca Raton, NW, USA: CRC Press, 2013, ch. 10, pp. 459–462.

[46] T. Fingscheidt, S. Suhadi, and S. Stan, "Environment-optimized speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 4, pp. 825–834, May 2008.

[47] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 1, pp. 7–19, Jan. 2015.

[48] A. LLC. *Multiprecision Computing Toolbox for MATLAB*. Accessed: Mar. 13, 2023. [Online]. Available: http://www.advanpix.com/

[49] R. Hendriks. *Toolbox for Log-Spectral Magnitude MMSE Estimators Under Super-Gaussian Densities*. Accessed: Mar. 13, 2023. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/25431-toolbox-for-log-spectral-magnitude-mmse-estimators-under-super-gaussian-densities

[50] H. Abdi and L. J. Williams, "Tukey's honestly significant difference (HSD) test," *Encyclopedia Res. Des.*, vol. 3, no. 1, pp. 1–5, 2010.

[51] M. Kim and J. W. Shin, "Improved speech enhancement considering speech PSD uncertainty," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 1939–1951, 2022.

[52] S. So and K. K. Paliwal, "Modulation-domain Kalman filtering for single-channel speech enhancement," *Speech Commun.*, vol. 53, no. 6, pp. 818–829, Jul. 2011.

SISI SHI was born in Jinan, China. She received the B.Eng. degree (Hons.) from Griffith University, Brisbane, in 2016, where she is currently pursuing the Ph.D. degree with the Signal Processing Laboratory. Her research interests include digital speech processing, speech enhancement algorithms, and deep learning.

KULDIP K. PALIWAL was born in Aligarh, India, in 1952. He received the B.S. degree from Agra University, Agra, India, in 1969, the M.S. degree from Aligarh Muslim University, Aligarh, in 1971, and the Ph.D. degree from Bombay University, Mumbai, India, in 1978. He has been carrying out research in the area of speech processing, since 1972. He was with a number of organizations, including the Tata Institute of Fundamental Research, Mumbai; the India Norwegian Institute of Technology, Trondheim, Norway; the University of Keele, U.K.; AT&T Bell Laboratories, Murray Hill, NJ, USA; AT&T Shannon Laboratories, Florham Park, NJ, USA; and Advanced Telecommunication Research Laboratories, Kyoto, Japan. Since July 1993, he has been a Professor with the School of Microelectronic Engineering, Griffith University, Brisbane, Australia. His current research interests include speech recognition, speech coding, speaker recognition, speech enhancement, face recognition, image coding, pattern recognition, and artificial neural networks. He has published more than 300 articles in these research areas. He is currently a fellow of the Acoustical Society of India. He has served as a Founding Member of the IEEE Signal Processing Society's Neural Networks Technical Committee, from 1991 to 1995, and the Speech Processing Technical Committee, from 1999 to 2003. He received the IEEE Signal Processing Society's Best (Senior) Paper Award for his paper on LPC quantization, in 1995. He was the General Co-Chair of the Tenth IEEE Workshop on Neural Networks for Signal Processing (NNSP2000). He has co-edited two books *Speech Coding and Synthesis* (Elsevier) and *Speech and Speaker Recognition: Advanced Topics* (Kluwer). He was an Associate Editor of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, from 1994 to 1997 and from 2003 to 2004. He is on the editorial board of the *IEEE Signal Processing Magazine*. He also served as an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS, from 1997 to 2000. He served as the Editor-in-Chief for the *Speech Communication Journal* (Elsevier), from 2005 to 2011.

ANDREW BUSCH received the B.Eng. and Ph.D. degrees in electrical and electronic engineering from the Queensland University of Technology, Brisbane, in 1998 and 2004, respectively. Since 2005, he has been a Lecturer, a Senior Lecturer, and an Associated Professor successively with Griffith University, Brisbane. He is currently the Deputy Head of School (Learning and Teaching) with the School of Built Environment and Engineering and the Director of the Griffith Sciences Partnerships Office. He is an active researcher in machine vision, medical imaging, and agricultural applications of machine vision and medical devices.