Applied Acoustics 202 (2023) 109134

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

On DCT-based MMSE estimation of short time spectral amplitude for single-channel speech enhancement

Sisi Shi [a,*], Kuldip Paliwal [a], Andrew Busch [b]

[a] Signal Processing Laboratory, Griffith School of Engineering, Griffith University, Nathan QLD 4111, Australia
[b] School of Engineering, Griffith University, Nathan QLD 4111, Australia

ARTICLE INFO

Article history:
Received 13 April 2022
Received in revised form 13 October 2022
Accepted 16 November 2022

Keywords:
Discrete Cosine transform (DCT)
Minimum mean-square error (MMSE) estimator
Speech enhancement
Super-Gaussian speech modelling
Speech presence uncertainty (SPU)

ABSTRACT

This paper proposes Discrete Cosine Transform (DCT) based speech enhancement algorithms. These algorithms utilize minimum mean square error (MMSE) estimator of clean short-time spectral amplitude, which respectively uses Gaussian, Laplace and Gamma probability density functions (PDF) as speech priors. We consider the noise process is additive and Gaussian. The proposed estimators are closed-form solutions, whereas the conventional Discrete Fourier Transform (DFT) based estimators derived under super-Gaussian speech priors have no closed-form solutions. We also examine the estimators with the Speech Presence Uncertainty (SPU) that addresses the speech or silence problem with probability. Compared to the alternative approaches, such as the Ephraim and Malah or the Erkelens et.al MMSE-STSA estimators, the proposed methods demonstrate superior performance in terms of Segmental SNR (SegSNR), Perceptual Evaluation of Speech Quality (PESQ), short-time objective intelligibility measure (STOI), and mean subjective preference score, while exhibiting an equal or lower complexity.

1. Introduction

Single-channel speech enhancement algorithms aim to improve the quality, and preferably intelligibility of corrupted speech signals [1]. They can therefore be used to reduce listener fatigue and to improve the performance of speech processing systems such as cochlear implants or speech recognition systems. This paper focuses on the minimum mean square error (MMSE) estimators of short-time spectral amplitude (STSA) due to its low computational complexity and good performance in various noise conditions [2]. The proposed algorithms are formulated in the short-time Discrete Cosine Transform (DCT) [3] domain and implemented under the Analysis-Modification-Synthesis (AMS) framework [4].

In a standard AMS framework, the noisy speech is processed within short-time segments, which typically are 20 to 40 ms in duration due to the non-stationary nature of speech signals. An orthogonal transform, e.g., the DCT or the Discrete Fourier Transform (DFT), is applied to decompose the framed noisy speech into its spectral components, i.e., transform expansion coefficients. Due to the major importance of spectral amplitude in both speech quality and intelligibility relative to the spectral phase [5], the majority of DFT-based algorithms aim to enhance the magnitude spectrum (MS) but not the phase spectrum (PhS) of corrupted speech, e.g., [6–9]. The enhanced signal is then synthesized by means of inverse transform and overlap adding. In this paper the short-time modifier is implied when referring to the spectral domain, DFT, DCT, and their corresponding spectra unless otherwise stated.

Although DFT is the most commonly used transform for speech enhancement, it is not necessarily optimal. Studies show that DCT is closer to the optimal Karhunen–Loeveis Transform (KLT) and offers a higher effectiveness in decorrelating signals compared to the DFT [21]. Since the DCT is a closer approximate of the KLT than the DFT, it is extensively used for data compression. Its outstanding energy compaction property is very useful for speech enhancement as well. For example, most of the speech energy is concentrated in only a few transform coefficients, whilst the noise energy is typically evenly distributed over the whole domain. This makes it easier to separate the noise energy from the noisy speech. DCT also produces about twice the independent spectral components of DFT, as half of the DFT coefficients are complex conjugates. Hence, the feasibility of using DCT for speech enhancement have been examined through methods such as the Wiener filter [11,12], the dual-gain Wiener (DGW) filter [13] and the dual-MMSE (DMMSE) [16] estimator. The latter two methods use bilateral gains that deal

* Corresponding author.
E-mail addresses: sisi.shi@alumni.griffithuni.edu.au (S. Shi), k.paliwal@griffith.edu.au (K. Paliwal), a.busch@griffith.edu.au (A. Busch).

https://doi.org/10.1016/j.apacoust.2022.109134
0003-682X/

with the constructive and destructive interference of the speech and noise DCT coefficients, respectively. Nonetheless, all dual gain estimators require a separate polarity algorithm [13] to determine the state of interference, and yet erroneous polarity estimation introduces undesirable artifacts [16]. These studies assume the speech and noise DCT coefficients are statistically independent Gaussian random variables. However, it has been remarked that clean speech components in the decorrelated domains are more accurately described by super-Gaussian distributions such as Laplacian (double-sided Exponential) or double-sided Gamma distributions, and noise components can be appropriately modeled by Gaussian distributions [7,22,23]. Thus, improved MMSE spectral estimators have been derived by assuming that DCT speech coefficients are super-Gaussian distributed and noise coefficients are either Gaussian distributed, i.e., [18–20], or Laplacian distributed, i.e., [17]. Methods introduced in [19,20] utilize Hidden Markov Model (HMM) parameter estimation and multivariate distribution for signal modeling. Here, we focus on uni-variate distributions, the derived closed-form solutions can be easily extended to a multivariate model. Notably, the estimators with DCT speech coefficients modeled as super-Gaussian distribution yields better performance than Gaussian model based estimators. Hitherto DCT-based speech enhancement algorithms concentrate on enhancing each spectral coefficients, and therefore they are not optimal for enhancing the STSA (refer to Table 1).

Despite theoretical advantages of the DCT, most speech enhancement methods still prefer the DFT, which has readily available STSA estimators [24]. Related work initialized by Ephraim and Malah [6] models the speech and noise DFT coefficients as independent circular-complex Gaussian random variables, partially due to mathematical convenience. The assumption of Gaussian priors implies the amplitudes and phases of DFT coefficients are statistically independent. Consequently, the STSA estimator can be elegantly reduced to a closed-form expression. It also takes into account the Speech Presence Uncertainty (SPU) in the noisy observations and further reduced residual noise [6]. Later, many STSA estimators have been proposed incorporating super-Gaussian priors for speech [8,9,15]. However, utilizing super-Gaussian speech models complicates the derivation of the estimators since the amplitudes and phases of the DFT coefficients are no longer independent. Some approximations for the conditional distribution of the clean speech amplitude and the Bessel function must be made due to the intractability of the closed-form solution [8,9]. This paper will show that this issue can be resolved by using a real-valued transform such as DCT. On the other hand, most STSA-estimation-based estimators use noisy phase for speech reconstruction thereby introducing an upper bound on the maximum improvement in speech quality [25]. Several methods have incorporated the phase estimation to alleviate this issue, but the performance of these methods remains sub-optimal [26–28]. A joint MMSE estimator of clean speech amplitude and phase was derived in [27] using the harmonic model-based method in [26] as prior phase information. It shows improved PESQ scores at the expense of degraded speech intelligibility [27,29]. This is due to the buzziness in the phase-enhanced signal as reported in [26,28–30]. These artifacts mainly stem from the spurious harmonics introduced particularly at high frequencies [26,30]. Consequently, speech distortion outweighs the achievable noise suppression especially at high SNRs leads to limitations on the effectiveness of this method [30]. Later, [28] derived an improved phase-aware MMSE STSA estimator where the cost function includes both a weighting factor and a power law. Unlike [27], it has been suggested in [28] using the maximum *a posteriori* phase estimator (MAP) [31] or the geometry-based method [32] to obtain the prior phase information. When compared to the phase-unaware STSA-estimation solutions, although [28] improves the perceived quality in terms of PESQ measure, it degrades speech intelligibility considerably in low SNRs in terms of STOI (Fig. 8, [28]).

Recently, the use of DCT Polarity Spectrum (PoS) in the context of STSA-estimation-based speech enhancement has been explored in [33]. For this, a theoretical analysis showed that the optimal estimate of the clean PoS is the noisy PoS under the Gaussian distribution assumptions and the constrained MMSE criterion. To verify this result experimentally, the effect of using the noisy PoS for signal resynthesis as compared to using the noisy PhS is evaluated through objective measures (i.e., PESQ [34]) and human listening tests. In these experiments, clean speech corpora were degraded with Gaussian white noise, at different segmental SNRs. The noisy speech corpora were modified and reconstructed in two transform domains separately: for DCT, the ideally filtered amplitudes were combined with the noisy sign components and similarly for DFT, the ideally filtered DFT amplitudes were combined with the noisy DFT phases. Thus the effects on the perceived speech quality are a result from the changes in PoS or PhS only. Results show the DCT PoS is better able to conserve the speech quality than the DFT PhS for the same level of global distortion. To examine this, we conduct both objective and subjective listening tests and discuss the results in Section 2.

In this paper, we derive DCT-based MMSE STSA estimators assuming that speech and noise coefficients are modeled by super-Gaussian and Gaussian distributions, respectively. They are MMSE estimators of the spectral amplitudes, rather than the MMSE estimators of the spectral coefficients (already reported previously, see Table 1), and these estimators were derived in closed-form without making any approximations. We also derive the estimators under the speech presence uncertainty. To determine the variance of the noise, we modified the noise estimator presented in [35], so that it can be used in combination with the DCT-based algorithms (see Appendix D). We chose this noise estimator due to its popularity as a baseline method at this present time. Moreover, the *a priori* SNR is estimated using the Decision-Directed (DD) [6] approach. The performance of our new estimators are compared to other DCT-based approaches such as [11,18,13,17], as well as some of the well known DFT-based MMSE STSA estimators, e.g., [6,9]. Additionally, it is investigated how the proposed system performs compared to the State-Of-The-Art

Table 1
Classification of the MMSE estimators with respect to transform domain. The noise is assumed to be additive and Gaussian. * indicate estimators for which no exact closed-form solutions exist.

Complex DFT		DCT	
Coefficient	**Spectral Amplitude**	**Coefficient**	**Spectral Amplitude (proposed)**
Wiener filter [10] super-Gaussian prior [7] generalized Gamma [14]	G_{EM}, complex Gaussian [6] complex Laplacian prior [8]* G_{L-FSA}, generalized Gamma [9]*,[15]*	G_W, Wiener filter [11,12] DGW, dual-gain Wiener [13] dual-gain MMSE, Gaussian prior [16] NBLG, dual-gain MMSE, Laplacian prior [17] G_{L-CSC}, super-Gaussian [18] multivariate Laplacian [19,20]	G_N, Gaussian, refer to (15) G_L, Laplacian, refer to (22) G_G, Gamma, refer to (27)

(SOTA) DFT-based phase-aware systems. This is of interest since the proposed STSA estimators rely on no prior knowledge of the DCT polarity compared to phase-aware STSA estimators, which rely on prior knowledge of the DFT phase.

This paper is organized as follows. In Section 2, the effect of using noisy polarity spectrum for signal resynthesis is analysed through objective and subjective measures. In Section 3, we explain the statistical models and assumptions used. Section 4 derives the optimal MMSE estimators of DCT spectral amplitudes, while Section 5 examines the new estimators under speech presence uncertainty. Section 6 evaluates the proposed estimators in various noise conditions, showing similar or improved performance in enhancing speech compared to competing algorithms. Finally, in Section 8, we discuss the results and draw conclusions.

2. The effect of using noisy polarity spectrum for speech resynthesis

In this section we explore the relevance of DCT polarity spectrum in the context of STSA-estimation-based speech enhancement. To achieve this, we use the approach described in [Sec.III, B-(b)] [33] to create polarity-only (PO) and phase-only (PhO) stimuli. The PO (or PhO) stimuli was generated by adding a controlled level of distortion into the polarity spectrum (or the phase spectrum), while keeping its spectral amplitudes fixed from the clean input. Thus the effects on the perceived speech quality are a result from the changes in PoS (or PhS) only. The distortion was added with respect to the Segmental SNR [SegSNR, as defined in Section 6.3, (41)].

The Perceptual Estimation of the Speech Quality (PESQ) [34] metric was employed as an objective speech quality measure. The results of testing using the modified signals with specified SegSNR, are shown in Fig. 1. It exhibits 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 SegSNR values.

For a more reliable indication of the stimuli quality, we further conducted three subjective tests, at three distinct SegSNR values: −5, 0, and 5 dB. The mean subjective preference (%) scores was used as the subjective quality measure. Mean subjective preference (%) scores were determined from a series of AB listening test [36]. Details about the subjective testing procedure are given in Appendix A. The results along with standard error bars are illustrated in Fig. 2. As expected, the clean speech stimuli achieved the highest subjective preference, while noisy speech stimuli were never preferred among other stimuli types. Explicitly, the PO stimuli achieved higher preference score compared to the PhO stimuli for all three tests.

Experimental results suggest 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 [33]. This allows PoS to have a higher degree of distortion tolerance than the PhS. Hence, the approximation of the clean PoS by its noisy counterpart can be considered superior, with a significantly lower SNR when compared to DFT-based methods. Therefore, speech enhancement can be achieved by combining the noisy polarity spectrum with accurately estimated DCT spectral amplitudes. In the successive three sections, we will develop optimal MMSE estimators of DCT spectral amplitudes.

3. Signal Models in the DCT domain

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:

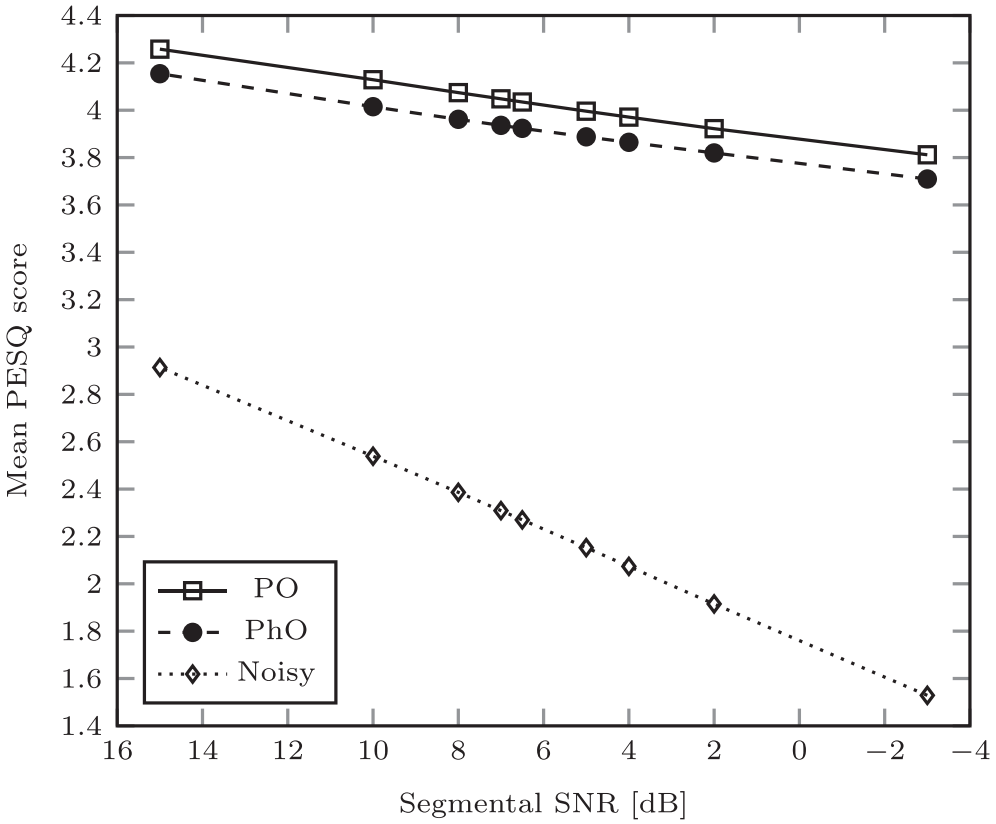

Fig. 1. Perceived quality estimation for Polarity-Only (PO, solid line), Phase-Only (PhO, dashed line) and noisy (dotted line) stimuli as a function Segmental SNR.

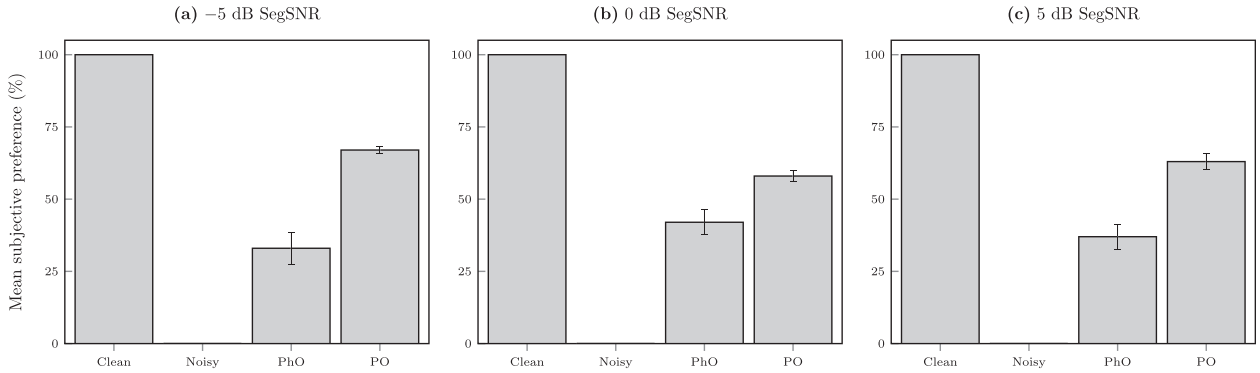

Fig. 2. Mean preference scores (with standard error bars) for four stimuli types at: (a) −5 dB, (b) 0 dB, and (c) 5 dB Segmental SNR (SegSNR).

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

The short-time DCT analysis of the observed speech signal, $y(n)$, is given by

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

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

$$cm_k = \begin{cases} \sqrt{\frac{1}{L}} & \text{for } k = 0 \\ \sqrt{\frac{2}{L}} & \text{for } k \neq 0 \end{cases} \tag{3}$$

n, k and i are the discrete time, frequency and frame index, respectively. $w(n)$ is the analysis window function of length N_w. N_s and L are the length of the frame shift and frequency analysis, respectively.

Let $Y(i,k) \triangleq \phi_Y(i,k)|Y(i,k)|$, $X(i,k)$, $D(i,k)$ denote the DCT spectral coefficients of the noisy $y(n)$, the clean speech $x(n)$, and the noise signal $d(n)$, respectively. We assume that $X(i,k)$ and $D(i,k)$ are statistically independent with zero mean. For better readability, the frame index i and the frequency index k are subsequently omitted and consider a single-DCT coefficient at a given time–frequency instant. Equation (1) can be represented in the DCT domain as:

$$\phi_Y|Y| = \phi_X|X| + \phi_D|D| \tag{4}$$

with Y given by (2). For this study, 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). Our task is to estimate the modulus $|X|$ from the degraded signal Y. The optimal solution can be computed as a gain function multiplied by the noisy DCT STSA:

$$|\widehat{X}| \triangleq G(\cdot,\cdot)|Y| \tag{5}$$

We use capital letters and its corresponding lower case letters to denote the random variable (R.V.) and its realization, respectively, and a hat symbol to denote its estimate, i.e.,$|\widehat{X}|$. It has been shown that in [33], the noisy PoS is the best estimate of the original PoS under the constrained MMSE criterion. Therefore, we combine the enhanced AS, $|\widehat{X}|$, with the noisy PoS, ϕ_Y, to get the final estimate of the spectral component

$$\widehat{X} \cong \phi_Y|\widehat{X}| \triangleq G(\cdot,\cdot)Y \tag{6}$$

The Gaussian assumption holds for the distribution of 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. Then, the Gaussian, the Laplacian, and the Gamma priors for the clean speech coefficients are defined as follows [37,18]:

- Gaussian speech prior

$$p(X) = \frac{1}{\sqrt{2\pi}\sigma_X}\exp\left(-\frac{X^2}{2\sigma_X^2}\right) \tag{8}$$

- Laplacian speech prior

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

- Double-sided Gamma speech prior

$$p(X) = \frac{\sqrt[4]{3}}{2\sqrt{2\pi\sigma_X}}|X|^{-\frac{1}{2}}\exp\left(-\frac{\sqrt{3}|X|}{2\sigma_X}\right) \tag{10}$$

where σ_X and σ_X^2 are the standard deviation and variance of the clean DCT coefficients, respectively. Note that the speech PDFs used in our paper are different from [7], despite of using the same citation of [37]. The PDFs of the imaginary and real parts of the DFT coefficient as listed in [7], use the variance of the complex DFT coefficient in the expression.

4. MMSE Estimation of DCT Spectral Amplitude

This section derives MMSE estimators of clean DCT STSA when the speech prior is modeled by a Gaussian, Laplacian or Gamma PDF, and the noise is Gaussian distributed. With the assumption that the DCT spectral coefficients are statistically independent, the MMSE estimator is obtained by computing the conditional expectation [6,9,15]:

$$|\widehat{X}| = E\{|X||Y\} = \int_0^\infty |x|p(|x||Y)\mathrm{d}|x| \tag{11}$$

where $E\{\cdot\}$ denotes the expectation operator. Since $|X|$ is a two-to-one function of X in that both $+x$ and $-x$ map to $|x|$, the probabilistic relationship between X and $|X|$ can be expressed as:

$$p(|X| = x) = p(X = |x|) + p(X = -|x|) \tag{12}$$

Utilizing this property, we can evaluate the expectation integral in (11) directly from the PDF of $p(X|Y)$:

$$|\widehat{X}| = \int_{-\infty}^{\infty} |x| p(x|Y) \mathrm{d}x \tag{13a}$$

$$= \frac{\int_{-\infty}^{\infty} |x| p(Y|x) p(x) \mathrm{d}x}{\int_{-\infty}^{\infty} p(Y|x) p(x) \mathrm{d}x} \tag{13b}$$

As given by (4), the noisy coefficient Y is the sum of two independent random variables, which imply that the conditional PDF of Y given X is

$$p(Y|X) = \frac{1}{\sqrt{2\pi}\sigma_D} \exp\left[\frac{-(Y-X)^2}{2\sigma_D^2}\right] \tag{14}$$

4.1. MMSE STSA estimator for Gaussian speech Prior

For the Gaussian speech prior, the MMSE estimator of the clean speech spectral amplitudes is given by (the derivation is comparable with [6], see Appendix B)

$$|\widehat{x}| = \frac{\xi}{1+\xi}\left\{\sqrt{\frac{2}{\pi v}}\exp\left(-\frac{v}{2}\right) + \mathrm{erf}\left(\sqrt{\frac{v}{2}}\right)\right\}|Y| \triangleq G_N(\xi,\gamma)\,|Y| \tag{15}$$

where $\mathrm{erf}(\cdot)$ denotes the error function [eq.8.250.1][38] and v is defined by

$$v \triangleq \frac{\xi}{1+\xi}\gamma \tag{16}$$

where ξ and γ are the *a priori* and *a posteriori* signal-to-noise (SNR), respectively and are defined by [6]

$$\xi \triangleq \frac{E\left\{|X|^2\right\}}{E\left\{|D^2\right\}|} = \frac{\lambda_X}{\lambda_D} = \frac{\sigma_X^2}{\sigma_D^2} \tag{17}$$

$$\gamma \triangleq \frac{|Y|^2}{E\left\{|D|^2\right\}} = \frac{|Y|^2}{\sigma_D^2} \tag{18}$$

It is interesting to note that for high *a priori* SNR, i.e., $\xi \gg 1$, the gain function in (15) approximates the Wiener filter gain function which is given by [39]

$$G_W = \frac{\xi}{1+\xi} \tag{19}$$

and it is independent of the *a posteriori* SNR, γ. Therefore, G_N can be interpreted as the Wiener filter gain multiplied by a modification factor. Moreover, the estimator (15) is derived based on the assumption that the speech and noise variance are known. As these values are in general not known a priori, they have to be estimated from the noisy observations as well. We employ a MMSE-based noise power estimator (see Appendix D) to determine the variance of the noise samples and a "decision-directed" (DD) approach [6] to estimate the *a priori* SNR of the speech samples on a frame-to-frame basis:

$$\widehat{\xi}(i) = \max\left\{\alpha_n \frac{|\widehat{X}(i-1)|^2}{\widehat{\sigma}_D^2(i-1)} + (1-\alpha_n)\max\left[\gamma(i)-1\,,0\right],\, \xi_{min}\right\} \tag{20}$$

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 $\max\{\cdot\}$ operator denotes the maximum function to ensure the positiveness of the estimator, while $\alpha_n = 0.98$ (was determined by simulations and informal listening tests in [6]) is the smoothing factor and $\xi_{min} = -25\,\mathrm{dB}$ is the SNR floor value for eliminating low-level musical noise [40].

The relation between G_N and G_{EM} [6] can be observed from their respective gain curves. The closeness of the gain curves, which correspond to the same value of ξ, implies that G_N and G_{EM} are nearly equivalent (Fig. 3). However, despite their similarity in behavior, G_N may yield better perceived speech quality than G_{EM}. Because G_N or G_{EM} only enhances the STSA of the noisy observation and leaves its polarity or phase spectrum unmodified; and using the noisy polarity spectrum for signal resynthesis has fewer consequences than using noisy phase spectrum (Section 2). On the other hand, we can also observe that both G_N and G_{EM} converges to the Wiener filter gain G_W (19) at high SNRs, however, G_W may result in over-attenuation of the weak signals in low SNR conditions.

4.2. MMSE STSA estimator for Laplacian speech Prior

Now we consider the DCT coefficients of the clean speech obey a Laplacian distribution. In analogy to the derivation developed in [7], we designate the following shorthand notations:

$$L_+ = \frac{\sigma_D}{\sigma_X} + \frac{|y|}{\sqrt{2}\,\sigma_D} = \frac{1}{\xi'} + \sqrt{\frac{\gamma}{2}} \tag{21a}$$

$$L_- = \frac{\sigma_D}{\sigma_X} - \frac{|y|}{\sqrt{2}\,\sigma_D} = \frac{1}{\xi'} - \sqrt{\frac{\gamma}{2}} \tag{21b}$$

$$M_+ = \exp\left(L{+}^2\right)\cdot \mathrm{erfc}(L_+) \tag{21c}$$

$$M_- = \exp\left(L{-}^2\right)\cdot \mathrm{erfc}(L_-) \tag{21d}$$

where $\mathrm{erfc}(\cdot)$ denotes the complementary error function [eq.8.250.4] [38]. After substituting (9) and (14) into (13b) and using [eq.3.322.2,3.462.1] [38], we obtain the optimal estimator:

$$|\widehat{X}| = \frac{\frac{2}{\sqrt{\pi}} - (L_+M_+ + L_-M_-)}{M_+ + M_-}\sqrt{2}\,\sigma_D = \left\{\frac{\frac{2}{\sqrt{\pi}} - (L_+M_+ + L_-M_-)}{(M_+ + M_-)}\sqrt{\frac{2}{\gamma}}\right\}|Y| \triangleq G_L(\xi',\gamma)\,|Y| \tag{22}$$

and the DD approach is used to estimate ξ':

$$\widehat{\xi}'(i) = \max\left\{\alpha_l \frac{|\widehat{X}(i-1)|}{\widehat{\sigma}_D(i-1)} + (1-\alpha_l)\sqrt{\max\left[\gamma(i)-1\,,0\right]},\, \xi'_{min}\right\} \tag{23}$$

with constant weighting factor $\alpha_l = 0.91$ and SNR floor value $\xi'_{min} = -12.5$ dB. These values were determined empirically via simulations and informal listening tests. Noting that $|\widehat{X}|$ is an even symmetric function of Y, enables us to interchange y with $|y|$ in (21).

Note that the equivalent of G_L in the DFT domain has no closed-form solutions [8,9,15]. To solve this problem, the authors in [8] approximated the join PDF of the DFT spectral amplitudes and phase with a simplified expression. Alternatively, [9] combined two kinds of approximations for the modified Bessel function of the first kind to obtain numerically stable results, for a desirable range of SNRs. Finally, numerical integration was resorted in [15] to compute (11). Fig. 4 shows the resulting gain curve of the DFT MMSE spectral amplitude estimator [9] G_{L-FSA} matches closely to the DCT estimator G_L for low *a priori* SNRs (-20 dB $\leqslant \xi \leqslant$-10 dB). As ξ increases, the value of G_{L-FSA} is up to 5 dB greater than G_L (-10 dB $\leqslant \xi \leqslant$5 dB). However, G_{L-FSA} no longer offers any noise attenuation when ξ is higher than 5 dB. This disadvantage is likely a consequence of using the approximations for the Bessel functions. Moreover, the DCT MMSE spectral coefficient estimator with Laplacian prior [18] G_{L-CSC} is almost identical to its DFT counterpart [14] G_{L-FSC}, for all SNR conditions. It can be seen that, the

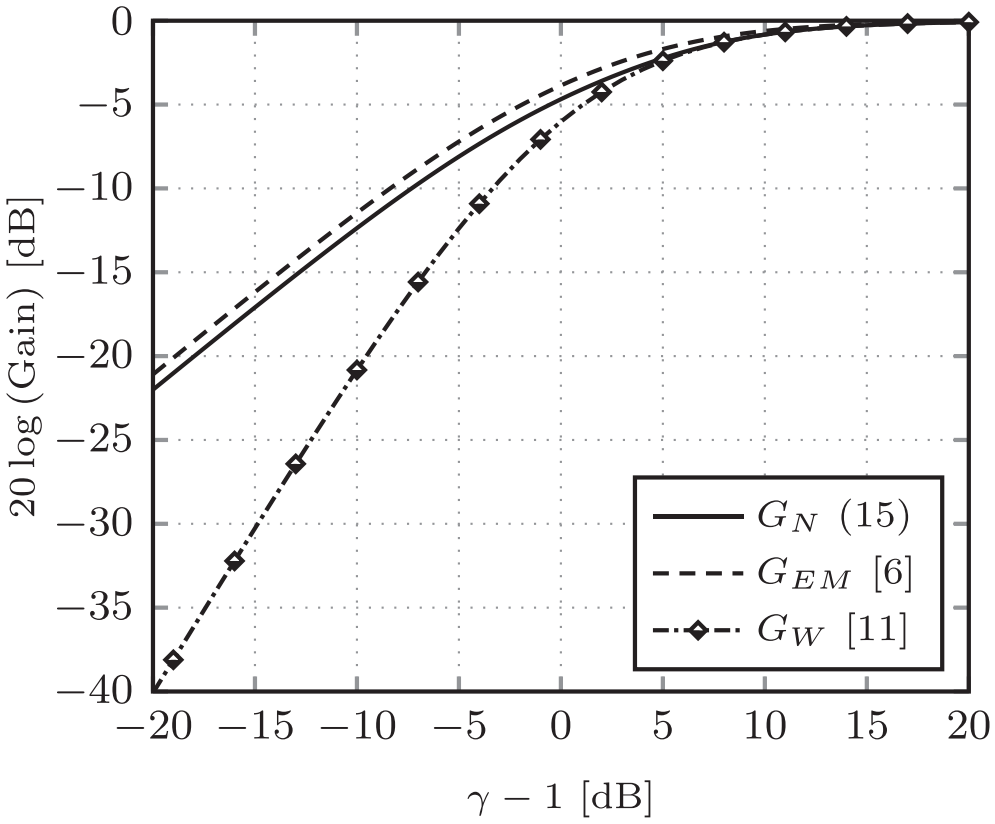

Fig. 3. Gain curves (with $\xi = \gamma - 1$) describing: DCT MMSE spectral amplitude estimator with Gaussian prior $G_N(\xi, \gamma)$ defined by (15), indicated with solid line; the Ephraim and Malah solution [6] G_{EM} (complex Gaussian speech prior), indicated with dashed line; Wiener filter solution G_W defined by (19) (Gaussian speech prior, linear filter), indicated with dash-dotted line.

spectral coefficient estimators always suppress more noise than the corresponding spectral amplitude estimators.

4.3. MMSE STSA estimator for double-sided Gamma speech Prior

Analogous to the Laplacian case and the derivation given in [7], we define

$$A_+ = \frac{\sqrt{3}\,\sigma_D}{2\,\sigma_X} + \frac{|y|}{\sigma_D} = \frac{\sqrt{3}}{2}\,\xi' + \sqrt{\gamma} \tag{24a}$$

$$A_- = \frac{\sqrt{3}\,\sigma_D}{2\,\sigma_X} - \frac{|y|}{\sigma_D} = \frac{\sqrt{3}}{2}\,\xi' - \sqrt{\gamma} \tag{24b}$$

and obtain the MMSE STSA estimator for Gamma speech prior by substituting (10) and (14) into (13b) [eq.3.462.1] [38]:

$$|\widehat{X}| = \frac{\sigma_D}{2}\left\{\frac{\exp\left(\frac{A_+^2}{4}\right)\mathscr{D}_{-\frac{3}{2}}(A_+) + \exp\left(\frac{A_-^2}{4}\right)\mathscr{D}_{-\frac{3}{2}}(A_-)}{\exp\left(\frac{A_+^2}{4}\right)\mathscr{D}_{-\frac{1}{2}}(A_+) + \exp\left(\frac{A_-^2}{4}\right)\mathscr{D}_{-\frac{1}{2}}(A_-)}\right\} \tag{25}$$

where $\mathscr{D}_p(z)$ denotes the parabolic cylinder function defined as [Th.9.240] [38]

$$\mathscr{D}_p(z) = 2^{\frac{p}{2}} e^{-\frac{z^2}{4}}\left\{\frac{\sqrt{\pi}}{\Gamma\left(\frac{1-p}{2}\right)}\Phi\left(-\frac{p}{2}, \frac{1}{2}; \frac{z^2}{2}\right) - \frac{\sqrt{2\pi}z}{\Gamma\left(-\frac{p}{2}\right)}\Phi\left(\frac{1-p}{2}, \frac{3}{2}; \frac{z^2}{2}\right)\right\} \tag{26}$$

where $\Gamma(z)$ denotes the gamma function [Th.8.310.1] [38] and $\Phi(\alpha, \gamma; z)$ is the confluent hypergeometric function [Th.9.210.1] [38]. We again interchange y with $|y|$ in (24) due to even symmetry of (25) and use the same approach as given in (23) to estimate ξ'.

Nevertheless, the computation of (25) for a wide dynamic range is not trivial, and numerical problems may result when the arguments are large. To improve numerical stability, we rewrite (25) in terms of the modified Bessel functions (see Appendix C)

$$\begin{aligned}|\widehat{X}| &= (2\sqrt{\gamma})^{-1}\left\{\frac{\exp\left(\frac{A_+^2}{4}\right)S + \exp\left(\frac{A_-^2}{4}\right)U}{\exp\left(\frac{A_+^2}{4}\right)T + \exp\left(\frac{A_-^2}{4}\right)V}\right\}|Y| \\ &\triangleq G_G(\xi', \gamma)\,|Y|\end{aligned} \tag{27}$$

where

$$S = A_+^{\frac{3}{2}}\left[K_{\frac{3}{4}}\left(\frac{A_+^2}{4}\right) - K_{\frac{1}{4}}\left(\frac{A_+^2}{4}\right)\right] \tag{28a}$$

$$T = A_+^{\frac{1}{2}} K_{\frac{1}{4}}\left(\frac{A_+^2}{4}\right) \tag{28b}$$

$$U = \frac{\pi}{\sqrt{2}}\,|A_-|^{\frac{3}{2}}\left\{I_{-\frac{3}{4}}\left(\frac{A_-^2}{4}\right) + I_{\frac{1}{4}}\left(\frac{A_-^2}{4}\right) - \phi_{A_-}\left[I_{\frac{3}{4}}\left(\frac{A_-^2}{4}\right) + I_{-\frac{1}{4}}\left(\frac{A_-^2}{4}\right)\right]\right\} \tag{28c}$$

$$V = \frac{\pi}{\sqrt{2}}\,|A_-|^{\frac{1}{2}}\left[I_{-\frac{1}{4}}\left(\frac{A_-^2}{4}\right) - \phi_{A_-} I_{\frac{1}{4}}\left(\frac{A_-^2}{4}\right)\right] \tag{28d}$$

where $\phi_{A_-} \triangleq \mathrm{sgn}(A_-)$, $I_{\pm\nu}(z)$ and $K_\nu(z)$ denote the modified Bessel functions of the first and second kind, respectively [eq.9.6.1] [42].

Similar to the Laplacian case, there is no closed-form solution for the equivalent DFT-based MMSE spectral amplitude estimator when the Gamma PDF is used [9,15]. Fig. 5 illustrates the gain

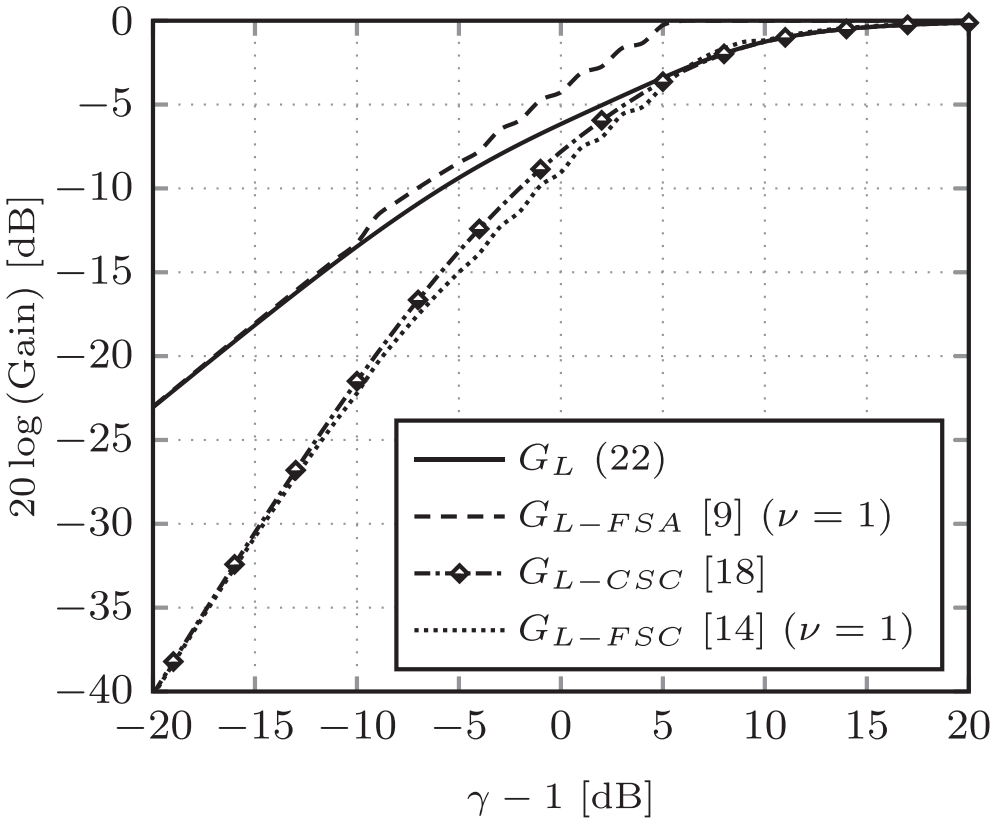

Fig. 4. Gain curves (with $\xi = \gamma - 1$) describing: DCT MMSE spectral amplitude estimator with Laplacian prior G_L defined by (22), indicated with solid line; DFT MMSE spectral amplitude estimator [9] G_{L-FSA} with $\nu = 1$ and $K = 20$ (i.e., the Taylor series of the modified Bessel function was truncated after 20 terms), indicated with dashed line; DCT MMSE spectral coefficient estimator with Laplacian speech prior [18] G_{L-CSC}, indicated with dashed-dotted line; DFT MMSE spectral coefficient estimator [14] G_{L-FSC} with $\nu = 1$ and $K = 20$, indicated with dotted line. MATLAB implementations of the algorithms presented in [9,14] are available at [41].

characteristics for the Gamma speech model. We find a behaviour similar to the case of Laplacian speech prior. Due to inaccurate approximations, G_{G-FSA} saturates at around 7 dB *a priori* SNR and thus, no longer provides any noise attenuation. It is worth noting that the difference in attenuation between the exact DCT spectral amplitude estimator G_G and the approximated DFT spectral amplitude estimator [9] G_{G-FSA} is generally small (within 1 dB) for low *a priori* SNRs. This suggests that G_G (or G_L) can be used as a computationally simpler alternative to the DFT MMSE estimator G_{G-FSA} (or G_{L-FSA}) without introducing the saturation problem at high *a priori* SNRs.

4.4. Gain characteristics of the proposed MMSE STSA estimators

Gain curves for the proposed MMSE STSA estimators are shown in Fig. 6, along with the respective curves of the well-known Ephraim and Malah (EM) algorithm [6] as a comparison. The gain functions are plotted against the *a priori* SNR, ξ, and the instantaneous SNR (ISNR), $\gamma - 1$, to describe the whole variations of the gain characteristics. Note that the relationship between ξ and ξ' can be easily obtained as

$$\xi' = \sqrt{\xi} \tag{29}$$

As can be observed in Fig. 6, the gain curves of G_N are almost identical with those obtained by the Ephraim and Malah solution. The coinciding of the gain curves indicates that (15) and the EM MMSE STSA estimator are nearly equivalent.

Fig. 7 illustrates the gain curves for various *a priori* SNR values: $\xi = -5$, 5, 10, and 15 dB. The Wiener filter gain given by (19) and the EM solution are included for reference. For desirable acoustic conditions (e.g., $\xi = 15$ or 5 dB), it demonstrates the new estimators converge to the Wiener filter for large values of ISNR. The gain curves of G_N follows those of G_{EM} closely while maintain slightly higher attenuation. It is interesting to note the dissimilarities in behavior between the G_N and G_G (or G_L) when the ISNR is small. The latter delivers more attenuation than the Wiener filter. This is due to the narrow peak of the *a priori* speech distribution, which shifts spectral estimates downward. On the other hand, for undesirable acoustic conditions (e.g., $\xi = -15$ or -5 dB), the estimators generally provide decreased attenuation than the Wiener filter when the ISNR decreases. This counter-intuitive behaviour was shown in [43] to help reduce the musical noise effect. Furthermore, when the ISNR is large, G_G (or G_L) provides significantly less attenuation than the Wiener filter. Due to the the leptokurtic (e.g., heavy-tailed) speech prior, it is highly likely that speech is present in this case. As a result, the estimators gain more success in recovering the speech spectral peaks and thus, reduce the amount of the perceived speech distortion.

5. Speech Presence Uncertainty Weighting

The above estimators were derived under the assumption that the speech is surely present in the noisy observation. However, speech is frequently absent during portions of silence and in voiced speech when most speech energy concentrated in multiples of the fundamental frequency. Therefore, improved speech enhancement was found when the estimators utilize the uncertainty of signal presence in the noisy spectral components [10,6,11,7,8]. Under this model, the appearance of the signal in the noisy spectral components is assumed to be statistically independent, along with the

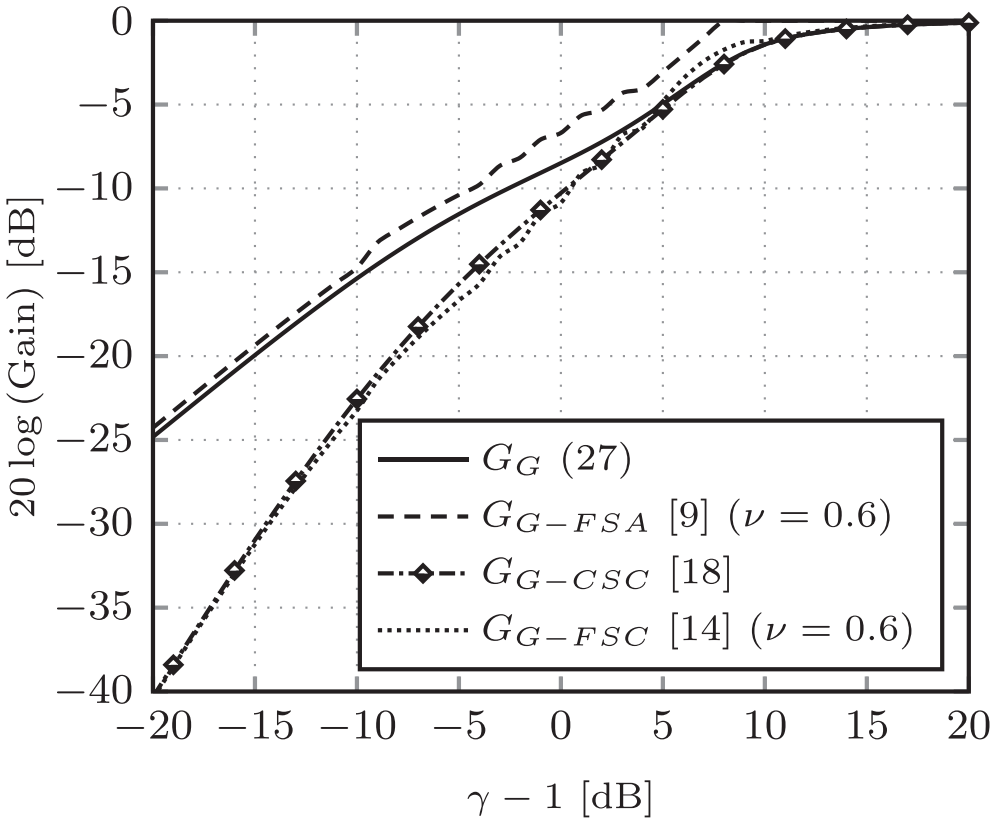

Fig. 5. Gain curves (with $\xi = \gamma - 1$) describing: DCT MMSE spectral amplitude estimator with Gamma prior G_G defined by (27), indicated with solid line; DFT MMSE spectral amplitude estimator [9] G_{G-FSA} with $v = 0.6$ and $K = 20$ (i.e., the Taylor series of the modified Bessel function was truncated after 20 terms), indicated with dashed line; DCT MMSE spectral coefficient estimator with Gamma speech prior [18] G_{G-CSC}, indicated with dashed-dotted line; DFT MMSE spectral coefficient estimator [14] G_{G-FSC} with $v = 0.6$ and $K = 20$, indicated with dotted line. MATLAB implementations of the algorithms presented in [9,14] are available at [41].

statistical independence assumption of the spectral components in Section 3, the MMSE estimator that incorporates speech presence uncertainty (SPU) is given by [44,45]

$$|\widehat{X}| = E\{|X||Y, H_1\}\, p(H_1|Y) \tag{30}$$

The *a posteriori* speech presence probability (SPP) can be obtained using Bayes' rule [44,6]

$$p(H_1|Y) = \frac{\Lambda}{\Lambda + 1} \tag{31}$$

where Λ denotes the generalized likelihood ratio

$$\Lambda = \frac{p(Y|H_1)}{p(Y|H_0)}\,\mu \tag{32}$$

with $\mu \triangleq (1-q)/q$, and $q = p(H_0)$ is the *a priori* probability for speech absence. H_1 and H_0 represent the hypotheses of speech presence and absence, respectively. $E\{|X||Y, H_1\}$ is the MMSE STSA estimator as given in (15), (22) or (27) when the speech signal is present in the noisy spectral component. Under hypothesis H_0, $Y = D$, and since the noise is Gaussian with zero mean and variance σ_D^2, it follows that

$$p(Y|H_0) = \frac{1}{\sqrt{2\pi}\sigma_D} \exp\left(-\frac{Y^2}{2\sigma_D^2}\right) \tag{33}$$

Under hypothesis H_1, $Y = X + D$, we have

$$p(Y|H_1) = \int_{-\infty}^{\infty} p(Y|x)p(x)\mathrm{d}x \tag{34}$$

Depending on the model for the speech PDF, we substitute one of (8), (9) or (10) in conjunction with (14) into the above equation to obtain the likelihood for speech activity. On substituting (33) and (34) into (32), we get an expression for the likelihood ratio. Specifically, using the Gaussian speech model for the spectral component we find

$$\Lambda_N(\xi, \gamma) = \frac{\exp(v/2)}{\sqrt{1+\xi}}\,\mu \tag{35}$$

Therefore, The amplitude estimator (30) can be written as

$$\begin{aligned} |\widehat{X}| &= \frac{\Lambda_N(\xi,\gamma)}{\Lambda_N(\xi,\gamma)+1} G_N(\xi,\gamma)\,|Y| \\ &\triangleq G_N^{SPU}(\xi,\gamma)\,|Y| \end{aligned} \tag{36}$$

where $G_N(\cdot,\cdot)$ is defined by (15). Similarly, for the Laplacian speech model

$$\begin{aligned} |\widehat{X}| &= \frac{\Lambda_L(\xi',\gamma)}{\Lambda_L(\xi',\gamma)+1} G_L(\xi',\gamma)\,|Y| \\ &\triangleq G_L^{SPU}(\xi',\gamma)\,|Y| \end{aligned} \tag{37}$$

where $G_L(\cdot,\cdot)$ is defined by (22) and

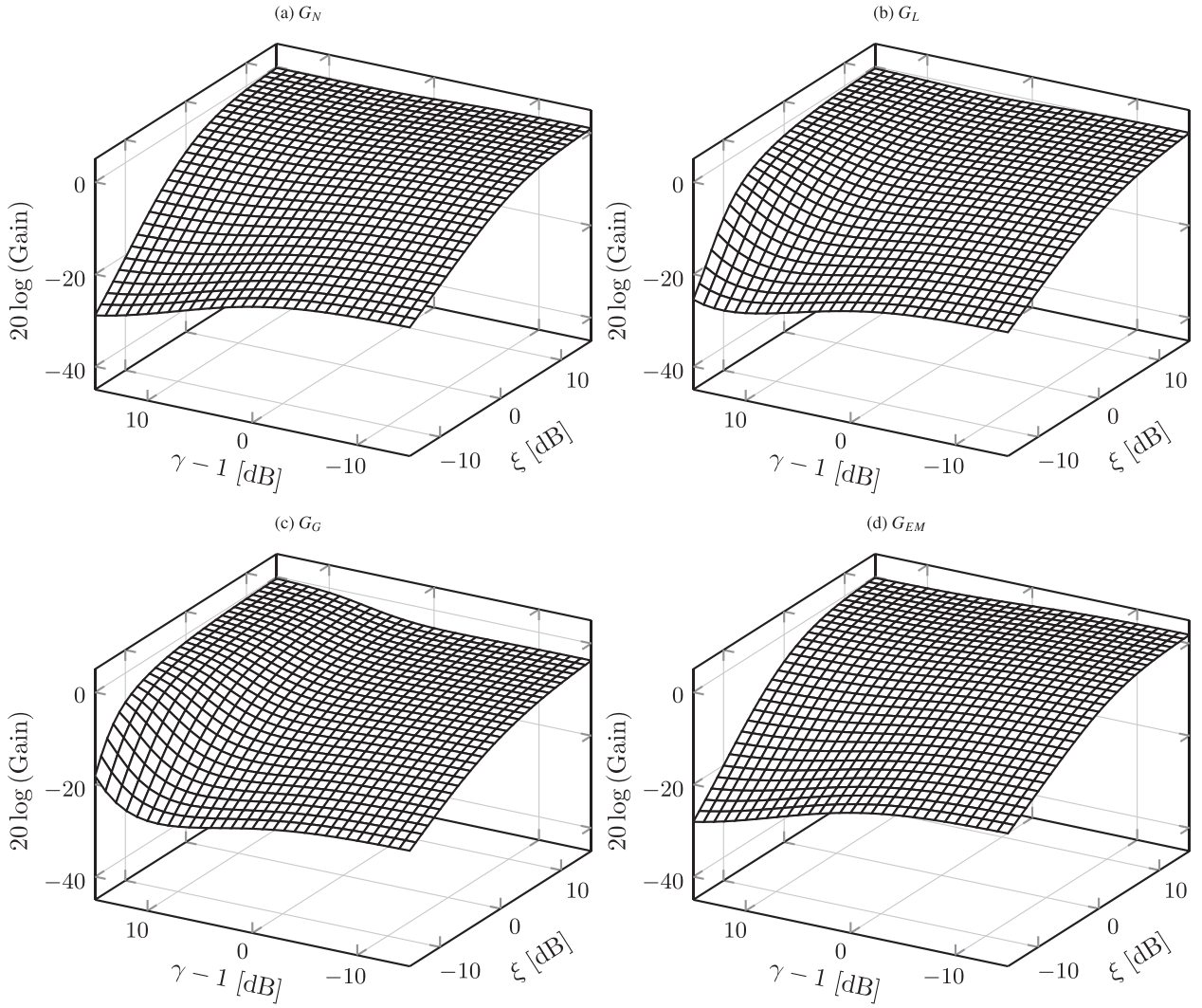

Fig. 6. Gain curves plotted against the *a priori* SNR ξ and the instantaneous SNR $\gamma - 1$ for the MMSE STSA estimators: (a) G_N (Gaussian speech prior) defined by (15), (b) G_L (Laplacian speech prior) defined by (22), (c) G_G (Gamma speech prior) defined by (27), and (d) the Ephraim and Malah solution G_{EM} (complex Gaussian speech prior), as seen in [(14)] [6].

$$\Lambda_L(\xi', \gamma) = \frac{\sqrt{\pi}}{2\,\xi'}\,(M_+ + M_-)\,\mu \tag{38}$$

For the Gamma speech model

$$\begin{aligned} |\widehat{X}| = &\ \frac{\Lambda_G(\xi', \gamma)}{\Lambda_G(\xi', \gamma) + 1} G_G(\xi', \gamma)\,|Y| \\ \triangleq &\ G_G^{SPU}(\xi', \gamma)\,|Y| \end{aligned} \tag{39}$$

where $G_G(\cdot, \cdot)$ is defined by (27) and

$$\Lambda_G(\xi', \gamma) = \frac{\sqrt[4]{3}}{4\sqrt{\pi\,\xi'}} \left\{ \exp\left(\frac{A_+^2}{4}\right) T + \exp\left(\frac{A_-^2}{4}\right) V \right\} \mu \tag{40}$$

Note that the original definition of ξ or ξ' was unconditional, whereas ξ or ξ' now provides the conditional SNR of the spectral component, assuming that speech is present. Nevertheless, we use the same estimate as in (20) or (23) for the resulting estimator as it is preferable in practice [46].

Gain curves which result from G_N^{SPU}, G_L^{SPU} and G_G^{SPU} are illustrated in Fig. 8, along with the corresponding curves of the Ephraim and Malah MMSE STSA estimator with SPU weighting [(30)] [6] G_{EM}^{SPU}, as a comparison. The gain functions for $q = 0.2$ are plotted against the *a priori* SNR and the ISNR. It shows G_N^{SPU} of (36), follows the Ephraim and Marlah solution closely and consistently, with slightly less attenuation in regions of high *a priori* SNR. Fig. 9 shows the gain curves for different *a priori* SNR values: $\xi = -15, -5, 5$, and 15 dB. The respective curves of Wiener filter with SPU weighting [11] and of G_{EM}^{SPU} are included for reference. It can be seen that G_{EM}^{SPU} yields around 5 dB higher attenuation than G_N^{SPU} as ISNR decreases and *a priori* SNR is high (i.e., $\xi = 15$ dB). It is interesting to compare these gain curves with those gain curves corresponding to the same ξ value as depicted in Fig. 7. We found that the estimators with SPU weighting generally provide more attenuation than the MMSE STSA estimators given in Section 4. For favorable acoustic conditions (i.e., $\xi = 5$, 15 dB), again the new estimators converges to the Wiener filter as ISNR increase. As ISNR decreases, G_N^{SPU} gives increased attenuation for which case $q = 0.2$, whereas G_N has decreased attenuation for which case $q = 0$. For unfavorable

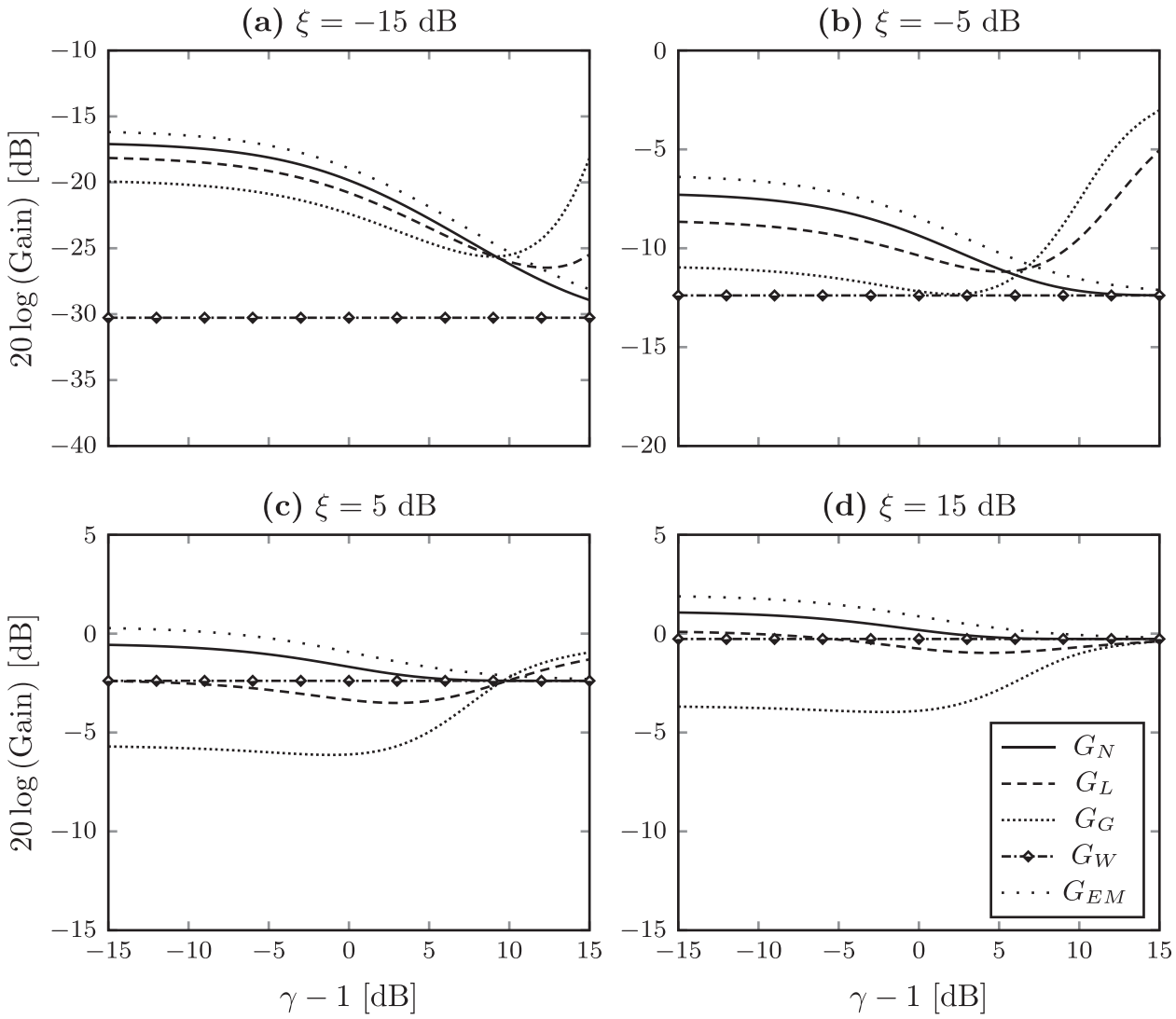

Fig. 7. Gain curves comparison for the proposed MMSE STSA estimators for (a) $\xi = -15$ dB, (b) $\xi = -5$ dB, (c) $\xi = 5$ dB and (d) $\xi = 15$ dB. The solid, dashed and dotted lines correspond to G_N (Gaussian speech prior), G_L (Laplacian speech prior) and G_G (Gamma speech prior), respectively. The corresponding Wiener filter solution given by (19) (Gaussian speech prior, linear filter), and the Ephraim and Malah solution [(14)] [6] G_{EM} (complex Gaussian speech prior), are respectively plotted with dash-dotted and loosely dotted lines for comparison.

acoustic conditions (i.e., $\xi = -5$, -15 dB), the gain curves of G_L^{SPU} shows increased attenuation compared with those of G_L. These variations in gain are a result of favoring the hypothesis of signal absence in such situations.

6. Implementation and Performance Evaluation

6.1. Test set

For the evaluation of our approach, we used 40 gender-balanced utterances from the TSP speech database [47]. These recordings were filtered with a linear phase, low-pass FIR filter and down-sampled to 16 kHz. 7 kinds of additive noise sources were applied to simulate noisy conditions. 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 [48], the last five being real-world non-stationary noise types. After combination with the clean speech utterances from above, $40 \times 7 = 280$ noisy speech utterances were obtained. Each evaluation was repeated for 0, 5, 10, 15 dB SNR conditions, respectively. The results were first averaged across all the utterances for a compact and general comparison as seen in Section 6.7 and 6.8. The objective and subjective test results for two non-stationary noise conditions (i.e., voice babble noise and F-16 noise) are reported in Section 6.9 and 6.11, respectively. The enhanced speech spectrograms produced by various speech enhancement algorithms are also analyzed in Section 6.10.

6.2. Experiment setup

For spectral analysis and synthesis, we employed a Hamming window of duration 20 ms with a 75% overlap between successive frames, corresponding to a window length of $N_w = 320$ and a window shift of $N_s = 80$. The frequency analysis length was $L = 2N_w = 640$. In implementing the new MMSE STSA estimators (given in Section 4 and 5), the gain value can be obtained by exact calculation or by using look up tables indexed by the *a priori* and the *posteriori* SNR values. The completed scripts for implementing the new estimators, as well as the enhanced utterances are available at https://github.com/SisiShi18/DCT_MMSE_STSA_EST.

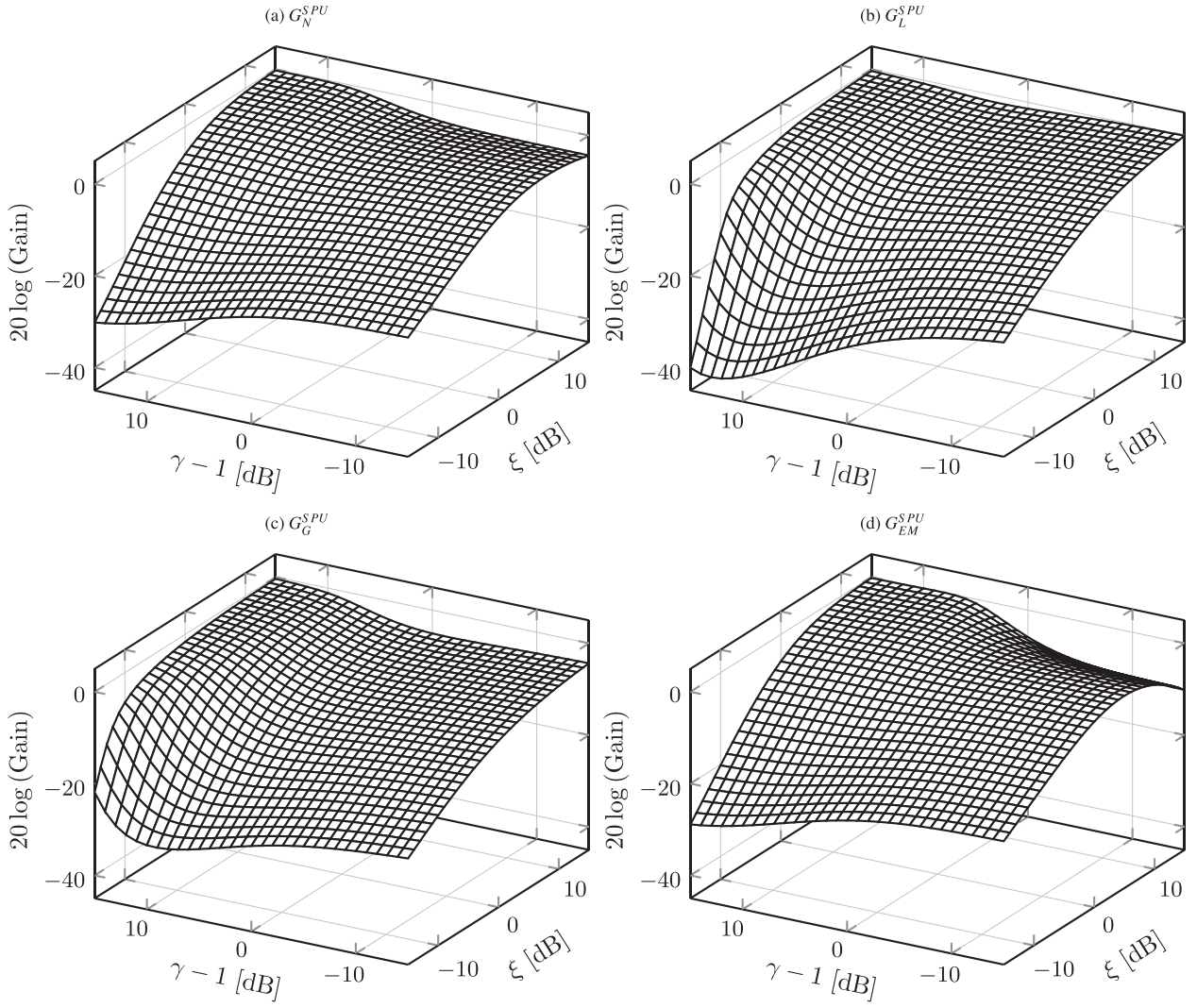

Fig. 8. Gain curves for the MMSE STSA estimators incorporating speech presence uncertainty (SPU) with $q = 0.2$. (a) G_N^{SPU} (Gaussian speech prior) defined by (36), (b) G_L^{SPU} (Laplacian speech prior) defined by (37), (c) G_G^{SPU} (Gamma speech prior) defined by ()39, and (d) the respective Ephraim and Malah solution G_{EM}^{SPU} (complex Gaussian speech prior), as seen in [(30)] [6].

6.3. Objective quality and intelligibility measures

The performance was measured in terms of: (i) the average segmental SNR (SegSNR), which is a local SNR computed over short segments; (ii) wideband perceptual evaluation of speech quality (PESQ) [49], which is an objective score for assessing speech quality in wideband telecommunication networks; and (iii) the short-time objective intelligibility (STOI) improvements [50], which has been shown to highly correlate with the intelligibility scores obtained through listening tests [50].

The SegSNR is defined as

$$\text{SegSNR} = \frac{10}{M}\sum_{i=0}^{M-1}\log_{10}\frac{\sum_{n=i\cdot N_w}^{i\cdot N_w+N_w-1} x^2(n)}{\sum_{n=i\cdot N_w}^{i\cdot N_w+N_w-1}\left[x(n)-\hat{x}(n)\right]^2} \tag{41}$$

where M is the number of signal segments that contain speech. $x(n)$ and $\hat{x}(n)$ are the clean and the enhanced speech signal, respectively. Speech pauses were excluded for summation in (41) by taking only frames with $-10\,\text{dB} < \text{SegSNR} < 35\,\text{dB}$.

6.4. Subjective evaluation

The objective measures predict the speech quality without assessing the phase distortion presented in the enhanced speech signal. In Section 2, we demonstrate that the DCT polarity spectrum is more capable of conserving the speech quality than the DFT phase spectrum giving the same amount of global distortion. Therefore, subjective evaluation was carried out through a series of blind AB listening tests [36] to obtain an accurate estimate of the perceived speech quality. The same subjective testing procedure described in Appendix A was used. Two utterances from the test set are used as the clean speech stimuli: sentence 9 from list 62, as uttered by male speaker MK, and sentence 9 from list 7, as uttered by female speaker FB. To produce the noisy speech stimuli, F-16 and voice babble noise were mixed with the clean speech stimuli from speaker MK and FB, respectively, at an SNR level of 5 dB. The enhanced speech stimuli for each of the speech enhancement

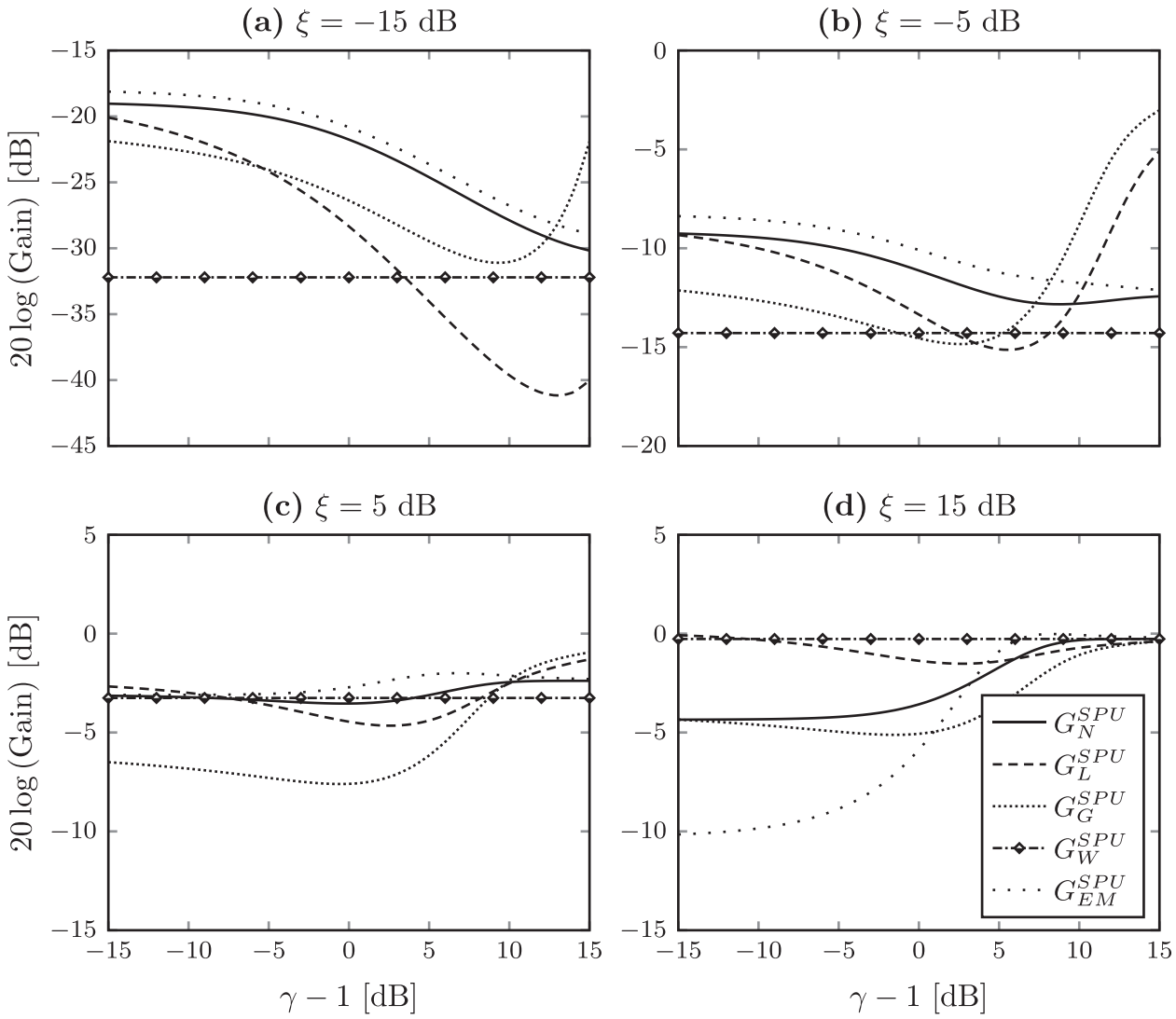

Fig. 9. Gain curves for the proposed MMSE STSA estimators under speech presence uncertainty (SPU) for (a) $\xi = -15$ dB, (b) $\xi = -5$ dB, (c) $\xi = 5$ dB and (d) $\xi = 15$ dB, with $q = 0.2$. The solid, dashed and dotted lines correspond to G_N^{SPU} (Gaussian speech prior) defined by (36), G_L^{SPU} (Laplacian speech prior) defined by (37), and G_G^{SPU} (Gamma speech prior) defined by (39), respectively. The corresponding curves for the modified Wiener filter [11] (Gaussian speech prior, linear filter), and the Ephraim and Malah solution [(30)] [6] G_{EM}^{SPU} (complex Gaussian speech prior), are respectively indicated with dash-dotted and loosely dotted line for reference.

methods was produced from the noisy speech stimuli. For each utterance, all possible stimuli pair combinations were presented to the listener. Each participant listened to a total of 220 stimuli pair combinations. 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.

6.5. Specifications of the competitive methods

For benchmarking, we included the following algorithms (refer to Table 1) in our evaluation: the Wiener filter in DCT domain without and with SPU (case G_W and G_W^{SPU}, respectively) [11]; the Laplacian-based MMSE DCT spectral coefficient estimator [18] implemented without and with SPU (cases G_{L-CSC} and G_{L-CSC}^{SPU}, respectively); dual gain Wiener filter (case DGW) [13]; non-linear bilateral Laplacian gain (case NBLG) estimator [17]; the Ephraim and Malah estimator [6] without and with SPU (cases G_{EM} and G_{EM}^{SPU}, respectively); MMSE DFT spectral amplitude estimator [9,41], with $\nu = 1$ (case G_{L-FSA}).

6.6. 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

$$\widehat{\sigma}_D^2 = |D|^2 \tag{42}$$

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 Appendix D and in [35] for the DCT-based algorithms and the DFT-based algorithms, respectively. We report the improvement (or gain) over the noisy input instead of the absolute value for all the measures.

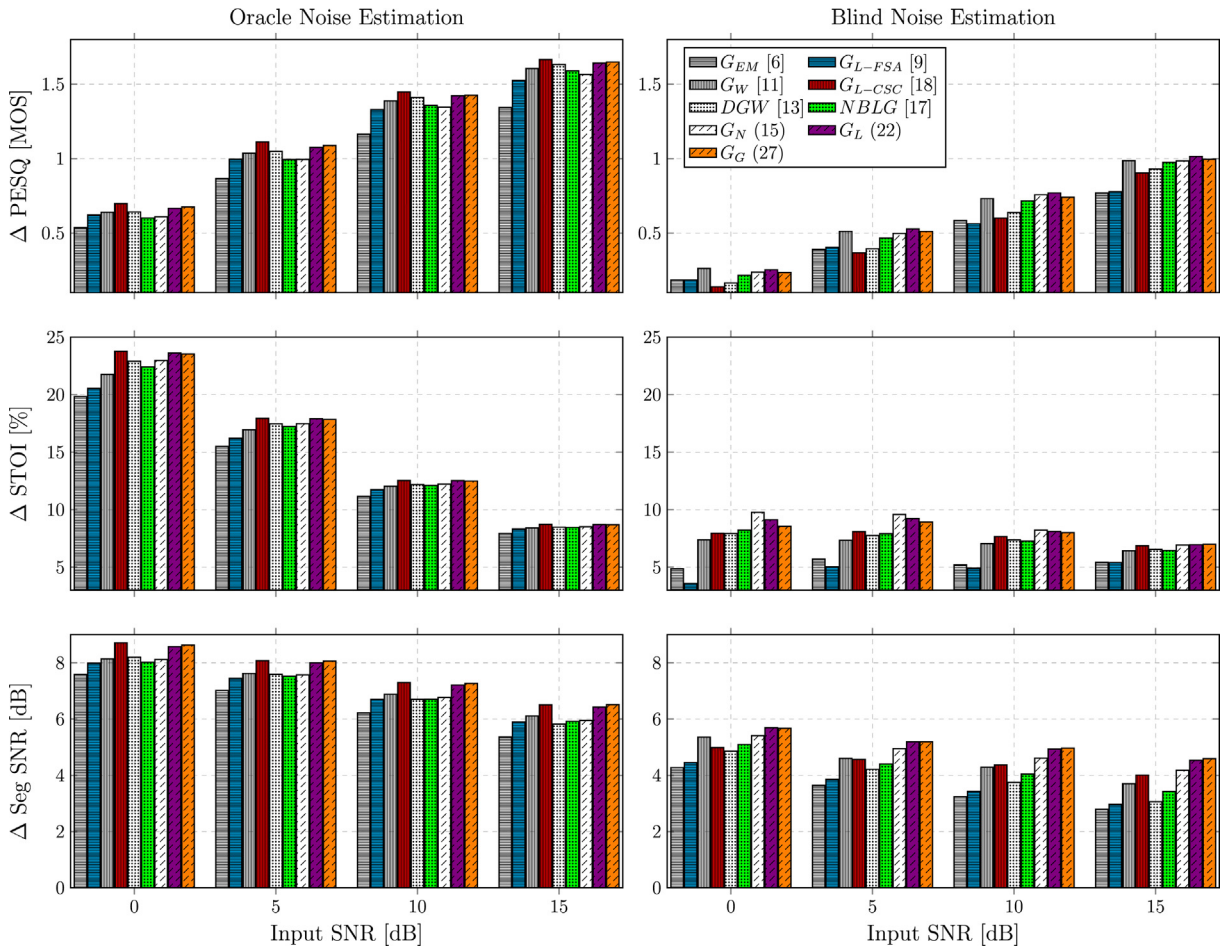

Fig. 10. Performance comparison among various estimators tested using the oracle noise estimator (left column) and blind noise estimation given the noisy speech (right column). Results are shown in terms of PESQ, STOI and Segmental SNR improvements and averaged over seven noise types (white noise, pink noise, speech noise, voice babble noise, F-16 noise, car factory noise and car Volvo-340 noise).

6.7. Objective test results without SPU

Fig. 10 shows the performance of different algorithms without SPU weighting. With the oracle noise estimator (left column), G_{L-CSC}, G_L and G_G yield the highest PESQ, STOI and SegSNR gains. On the contrary, with the nominated noise estimators (right column) the proposed estimators outperform the benchmark algorithms for all SNRs. In particular, the PESQ gain is maximum for G_L, closely followed by G_G, G_N, and G_W. As compared to the oracle case, G_L result in significant higher PESQ scores than G_{L-CSC} (about 0.2 higher for input SNRs above 5 dB). As the input SNR increases, the margin of improvement offered by the proposed estimators over the DFT-based algorithms (i.e., G_{EM} and G_{L-FSA}) increases. Moreover, G_L, G_G and G_N result in higher STOI gain than the other algorithms for input SNR in the range 0 dB to 15 dB. Contrarily to the PESQ scenario, the margin of improvement in case of STOI decreases with increase in input SNR. G_L and G_N give about 1% and 2% higher STOI gain over G_{L-CSC} for SNRs less than 10 dB, respectively. Finally, G_L and G_G yield the highest SegSNR gain for all SNRs. G_L gives about 1 dB higher SegSNR gain than G_{L-CSC} for all SNRs.

6.8. Objective test results with SPU

Fig. 11 compares the performance of various algorithms with the SPU weighting. For the oracle case, we found that the SPU approach improves the PESQ, STOI, and SegSNR especially in low SNR conditions as compared to the results in Fig. 10. Listening tests confirm that better perceived quality is achieved during speech pauses. When the blind noise estimate is used, the proposed estimators consistently yield a higher PESQ, STOI and SegSNR gain than other methods. Especially, the STOI and SegSNR score predict intelligibility and quality improvement for G_L^{SPU} and G_G^{SPU} over G_L and G_G for input SNR in the range 0 dB to 5 dB. Whereas an intelligibility drop is suggested for G_{EM}^{SPU} as compared to G_{EM} within the SNR range 10 to 15 dB. It implies the DCT based proposed estimators are less affected by the signal distortions than the benchmark methods.

6.9. Objective test results for real world non-stationary noises

The objective perceived quality scores attained by each speech enhancement algorithms over two non-stationary noise conditions are given in Table 2 and 3. It can be seen that the proposed method G_L (Laplacian PDF), is able to produce enhanced speech at a higher quality than other methods for coloured F-16 noise at all SNR levels [except for SNR at 0 dB, where the proposed G_G (Gamma PDF) performs best]. For the voice babble noise, the proposed G_N (Gaussian PDF) and G_L obtained the highest PESQ and SegSNR scores, respectively. The objective intelligibility scores attained by each method are given in Table 4. The proposed G_L is able to produce more intelligible enhanced speech than other methods for all conditions

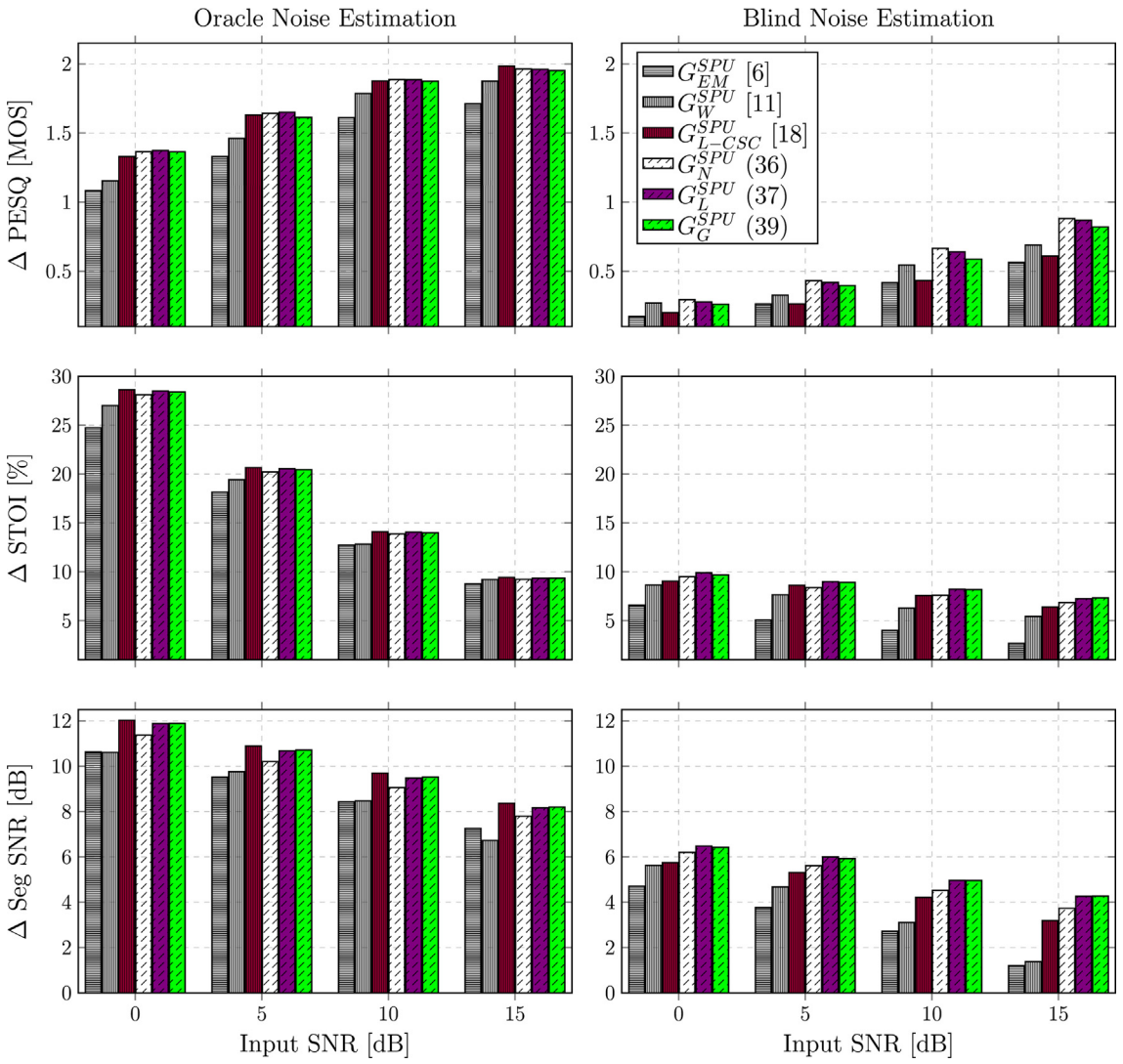

Fig. 11. Performance comparison among various estimators with speech presence uncertainty (SPU) tested using the oracle noise estimator (left column) and blind noise estimation given the noisy speech (right column). Results are shown in terms of PESQ, STOI and Segmental SNR improvements and averaged over seven noise types (white noise, pink noise, speech noise, voice babble noise, F-16 noise, car factory noise and car Volvo-340 noise).

Table 2
Performance comparison, in terms of **PESQ score gains**, between various estimators tested using the nominated MMSE noise estimator.

Noise	Method	0 dB	5 dB	10 dB	15 dB
F-16 two seat	G_{EM}	0.071	0.212	0.567	0.770
	G_{L-FSA}	0.084	0.228	0.575	0.709
	G_W	0.121	0.375	0.689	0.990
	G_{L-CSC}	0.069	0.278	0.537	0.896
	DGW	0.071	0.282	0.545	0.923
	NBLG	0.098	0.315	0.628	0.989
	G_N	0.123	0.377	0.694	1.076
	G_L	**0.152**	**0.402**	**0.723**	**1.108**
	G_G	0.140	0.385	0.696	1.086
Voice babble	G_{EM}	0.039	0.150	0.383	0.377
	G_{L-FSA}	0.026	0.110	0.285	0.188
	G_W	0.044	0.161	0.271	0.259
	G_{L-CSC}	0.024	0.117	0.211	0.295
	DGW	0.054	0.162	0.336	0.438
	NBLG	0.067	0.208	0.452	0.528
	G_N	**0.075**	**0.244**	**0.465**	**0.627**
	G_L	0.063	0.219	0.372	0.546
	G_G	0.052	0.191	0.306	0.498

Table 3
Performance comparison, in terms of **Segmental SNR gains**, between various estimators tested using the nominated MMSE noise estimator.

Noise	Method	0 dB	5 dB	10 dB	15 dB
F-16 two seat	G_{EM}	3.974	4.333	2.184	0.917
	G_{L-FSA}	4.012	4.385	2.229	0.861
	G_W	5.258	4.997	3.558	1.901
	G_{L-CSC}	5.278	5.408	4.327	2.687
	DGW	4.843	4.609	3.172	1.282
	NBLG	4.945	4.608	3.044	0.725
	G_N	5.584	5.473	4.516	3.052
	G_L	5.980	**5.750**	**4.779**	**3.135**
	G_G	**6.009**	5.743	4.751	3.132
Voice babble	G_{EM}	3.433	2.342	1.595	0.640
	G_{L-FSA}	3.508	2.395	1.642	0.716
	G_W	4.179	2.931	2.860	1.465
	G_{L-CSC}	4.086	3.050	2.394	2.034
	DGW	3.942	2.801	1.868	1.194
	NBLG	4.066	2.892	1.962	1.506
	G_N	4.330	3.292	2.787	2.090
	G_L	**4.558**	**3.399**	**2.878**	**2.236**
	G_G	4.516	3.360	2.863	2.229

Table 4
Performance comparison, in terms of **STOI score gains**, between various estimators tested using the nominated MMSE noise estimator.

Noise	Method	0 dB	5 dB	10 dB	15 dB
F-16 two seat	G_{EM}	0.096	0.086	0.063	0.043
	G_{L-FSA}	0.083	0.091	0.067	0.049
	G_W	0.120	0.129	0.092	0.078
	G_{L-CSC}	0.118	0.128	0.096	0.085
	DGW	0.114	0.124	0.092	0.080
	NBLG	0.121	0.126	0.093	0.079
	G_N	0.139	0.131	0.101	0.086
	G_L	**0.146**	**0.143**	**0.107**	**0.089**
	G_G	0.143	0.142	0.106	**0.089**
Voice babble	G_{EM}	0.014	0.035	0.059	0.046
	G_{L-FSA}	0.012	0.041	0.061	0.049
	G_W	0.044	0.077	0.074	0.062
	G_{L-CSC}	0.046	0.084	**0.087**	0.067
	DGW	0.049	0.079	0.082	0.065
	NBLG	0.051	0.079	0.079	0.063
	G_N	**0.054**	0.085	0.082	0.079
	G_L	0.049	0.089	**0.087**	**0.081**
	G_G	0.046	**0.090**	0.086	0.080

(except for voice babble at 0 dB and 5 dB, where the proposed G_N and G_G gives the best scores, respectively).

6.10. Spectrogram analysis

This section analyzes the enhanced speech spectrograms produced by each of the speech enhancement algorithms for the stimuli set described in Section 6.4. Specifically, Fig. 12 (a) shows the spectrogram of the clean speech (male utterance, MK62_09). F-16 noise at an SNR level of 0 dB is used to create the noisy speech in Fig. 12 (b). The enhanced speech produced by the proposed methods are shown in Fig. 12 (h)-(j). It can be seen that the enhanced speech produced by G_L and G_G contains the least amount of residual background noise than the other methods.

The spectrogram of the clean speech (female utterance, FB07_09) is shown in Fig. 13 (a). The clean speech is corrupted by voice babble noise at an SNR level of 0 dB to generate the noisy speech shown in Fig. 13 (b). This is a particularly tough condition for speech enhancement since the background noise exhibits characteristics similar to the speech produced by the target speaker.

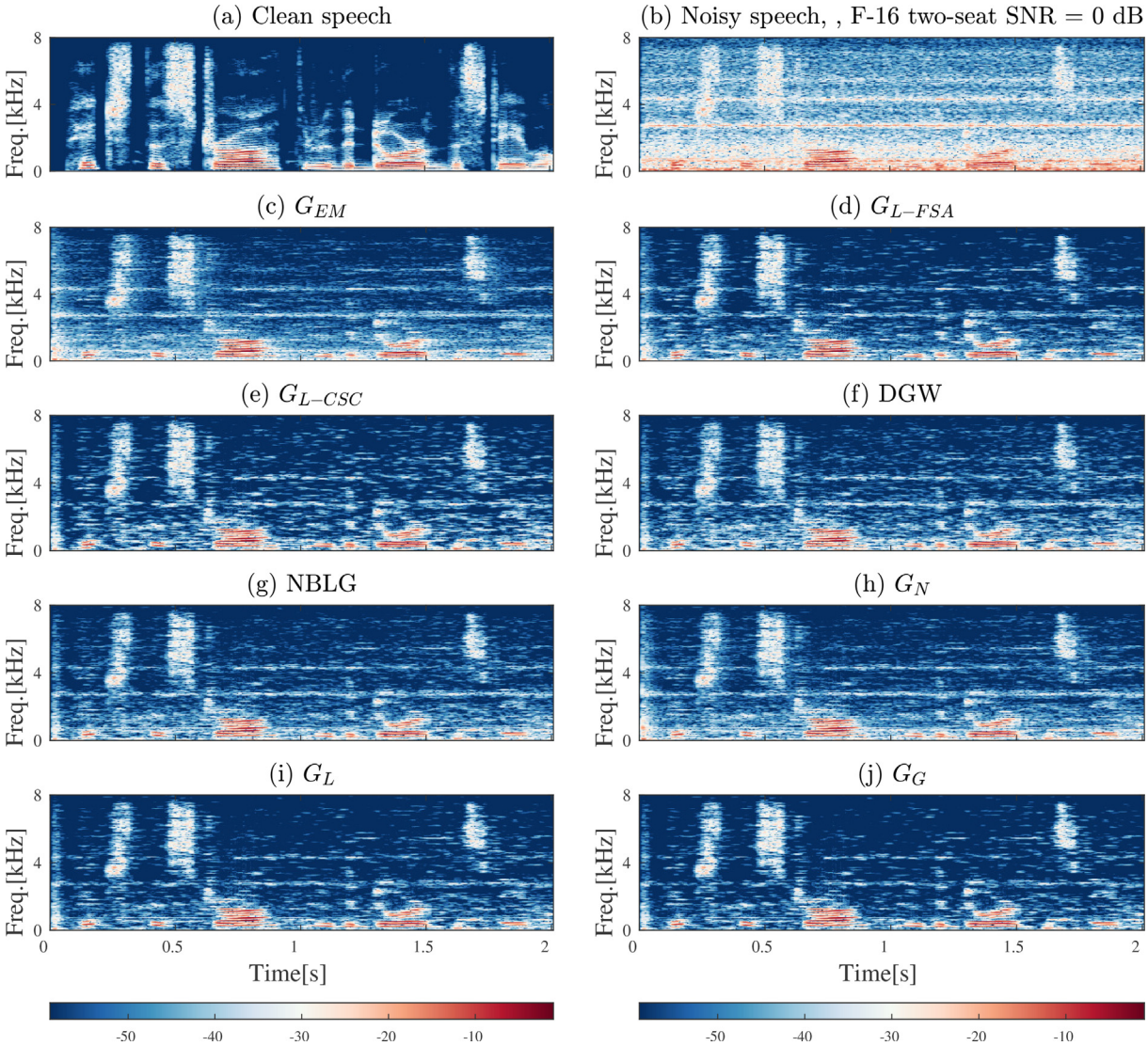

Fig. 12. Spectrograms of (a) the clean sentence, (b) the sentence corrupted by non-stationary F-16 noise at 0 dB, and (c)-(j) enhanced speech produced by corresponding speech enhancement algorithm (see Table 1). The sentence 'Pitch the straw through the door of the stable' (utterance MK62_09), was taken from the TSP speech database [47]. The nominated MMSE noise estimator introduced in Appendix D and in [35] were used for the DCT-based methods and DFT-based methods, respectively. The decision-directed approach [6] was used for the *a priori* SNR estimation.

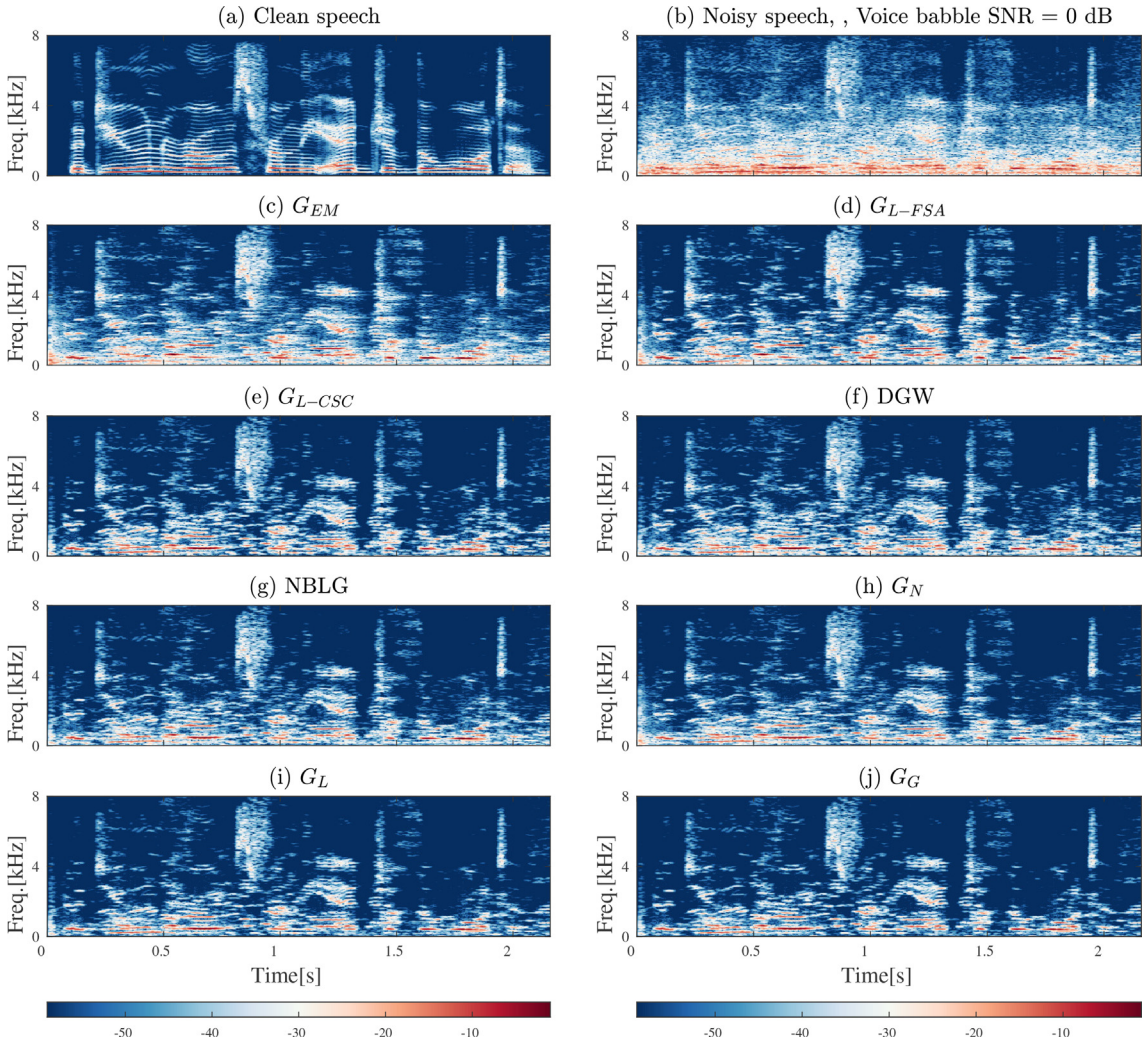

Fig. 13. Spectrograms of (a) the clean sentence, (b) the sentence corrupted by voice babble noise at 0 dB, and (c)-(j) enhanced speech produced by corresponding speech enhancement algorithm (see Table 1). The sentence 'The dune rose from the edge of the water' (utterance FB07_09), was taken from the TSP speech database [47]. The nominated MMSE noise estimator introduced in Appendix D and in [35] were used for the DCT-based methods and DFT-based methods, respectively. The decision-directed approach [6] was used for the *a priori* SNR estimation.

Moreover, the background noise is more predominant than the target speech in the band from 0 to 4 kHz. The proposed estimators utilising the Laplacian and Gamma PDF (case G_L and G_G, respectively), are able to reduce most of the residual noise with less or equal amount of speech distortion [Fig. 13 (i) and (j)]. Note that Fig. 12 and 13 are representative images for most utterances, which show similar characteristics. In both of the noise conditions, G_L (or sometimes G_G) gives best results, when compare to other approaches.

The objective measures and the spectrogram analysis predict the speech quality without assessing the phase distortion presented in the enhanced speech signal. Prior work presented in Section 2 indicates that the DCT polarity spectrum is more robust to noise distortion than the DFT phase spectrum. Thus, in the next section, we carry out human listening tests [36] to obtain an accurate estimate of the perceived speech quality.

6.11. Subjective test results

The mean subjective preference scores (%) for each algorithm are shown in Fig. 14 and 15. The F-16 noise experiment in Fig. 14 reveals that the proposed method G_G, utilising the Gamma PDF, was widely preferred (85%) by the listeners over the competing methods, apart from the clean speech (100%). G_L, utilising the Laplacian PDF, is found to be the next most preferred method (75%), following by G_N (62.5%) and DGW (57.5%). These results support the objective quality scores as seen in Table 2–4, even though G_L obtained marginal higher objective scores than G_G.

The subjective listening test results for the voice babble noise condition is shown in Fig. 15. It shows that the proposed method utilising the Laplacian PDF G_L, achieves a better preference score (77.5%) than other methods, except for the clean speech (100%). As contrary to the previous experiment, G_G was the next most preferred method (67.5%), followed by DGW (65.5%) and G_{L-CSC} (62.5%). These results indicate the enhanced speech generated by the proposed methods, G_L and G_G, exhibits better perceived speech quality among all other methods for two real world noise sources.

From previous results, it can be seen that the proposed methods outperformed their DFT-based counterparts. As shown in Section 2, this is mainly because the consequences of using noisy DCT polarity spectrum is less severe than using DFT noisy phase spectrum for signal reconstruction. Nonetheless, previous evaluations were made within the class of methods which essentially rely on no prior knowledge of phase or polarity. In the next section we inves-

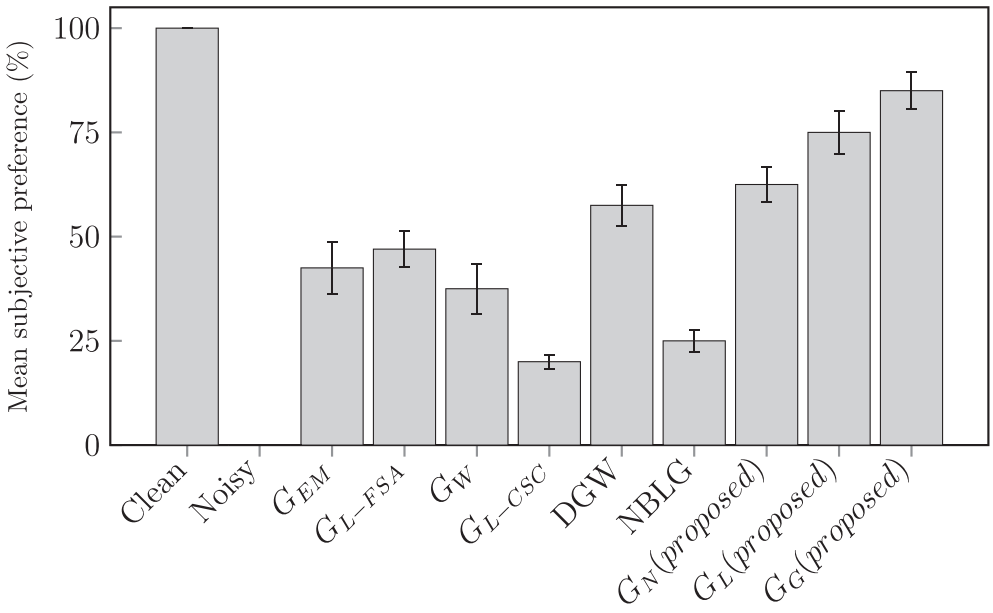

Fig. 14. The mean subjective preference score (%) comparison for each speech enhancement method. The male utterance (MK62_09) corrupted with 5 dB non-stationary F-16 noise was used for the subjective tests. The error bars indicated the standard deviation of the scores.

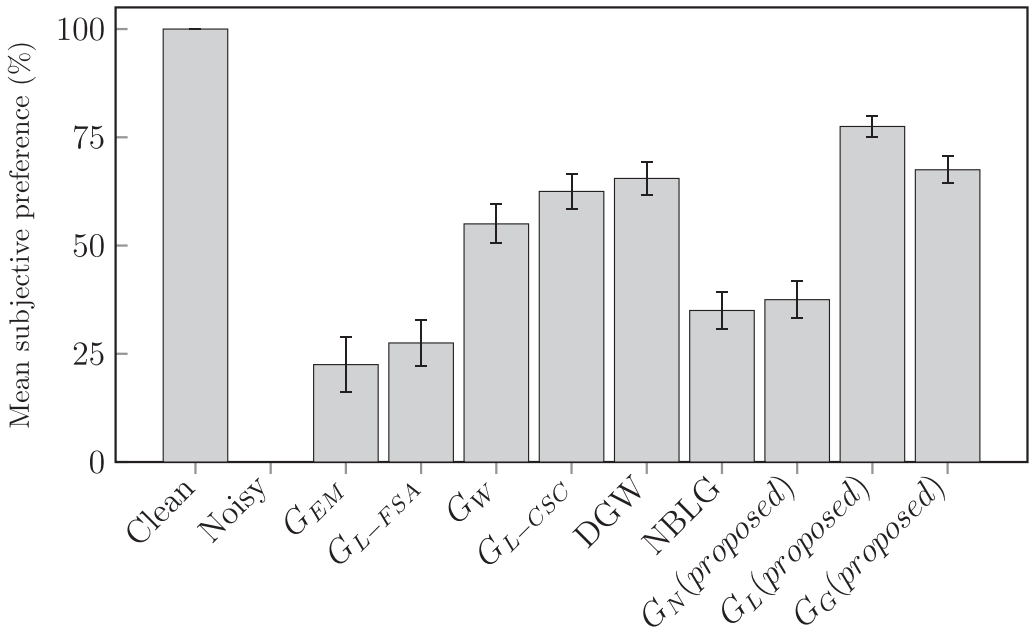

Fig. 15. The mean subjective preference score (%) comparison for each speech enhancement method. The female utterance (FB07_09) corrupted with 5 dB voice babble noise was used for the subjective tests. The error bars indicated the standard deviation of the scores.

tigate how the proposed methods performs compared to the State-Of-The-Art (SOTA) phase-aware STSA estimators, which rely on prior knowledge of the DFT phase.

6.12. Compare to SOTA Phase-aware STSA estimators

Phase-aware STSA estimators have shown the potential to improve the speech enhancement (SE) performance given that the phase spectrum is accurately estimated [27,28,30]. Nevertheless, phase estimation is still a challenging task mainly due to the unavailability of a useful DFT phase structure [29]. The model-based method, e.g., [26,27], requires an extra fundamental frequency estimator together with a voice activity detector. An erroneous signal defection leads to buzzyness and reduced quality compared to the level of the input noisy phase [26,27,29]. The MAP phase estimator [31] doesn't require a voice activity detector but suffers from spectral leakage due to inaccurate fundamental frequency estimations [29]. Moreover, the complexity associated with DFT phase-aware SE systems is much larger than that associated with the proposed DCT-based SE system. Both phase-aware STSA estimators in [27,28] have no closed-form solutions and thus numerical integration was required.

Given the difficulty of accurate phase estimation and complexity associated with DFT phase-aware SE systems, it is of interest to understand the performance gain of this type of systems as compared to the simpler proposed system. For this, we compare the

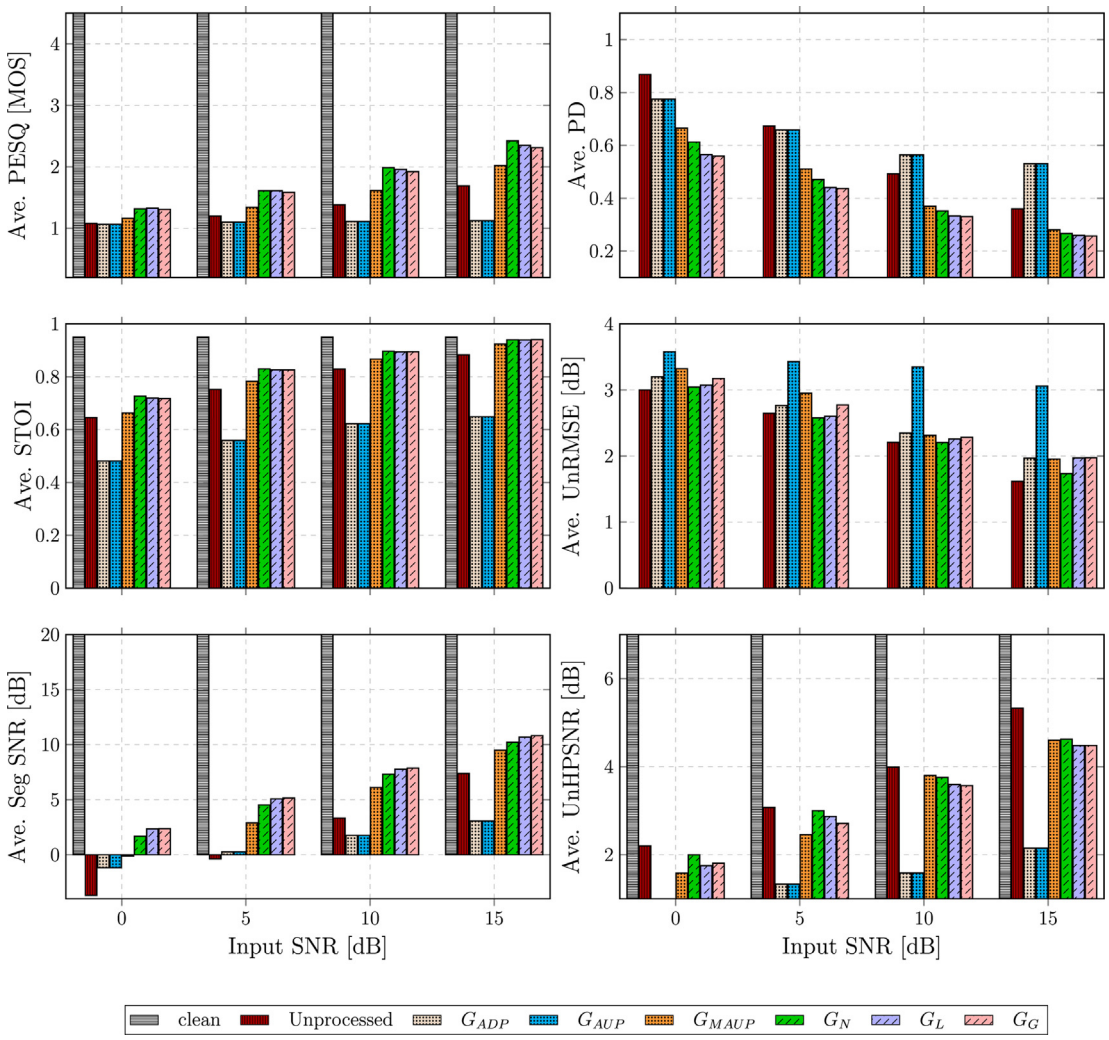

Fig. 16. Performance comparison between the state of the art phase-aware estimators (i.e., ADP [51], AUP [27], and MAUP [28]) and the propose estimators [i.e., G_N (15), G_L (22), and G_G (27)] using blind *a prior* SNR and noise PSD estimation. The clean speech and noisy (unprocessed) speech are also included as the upper bound and lower bound of the performance, respectively. Results are shown in terms of conventional (i.e., PESQ, STOI and Segmental SNR) and phase-aware (i.e., PD, UnRMSE, and UnHPSNR) instrumental metrics. Scores of the metrics 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).

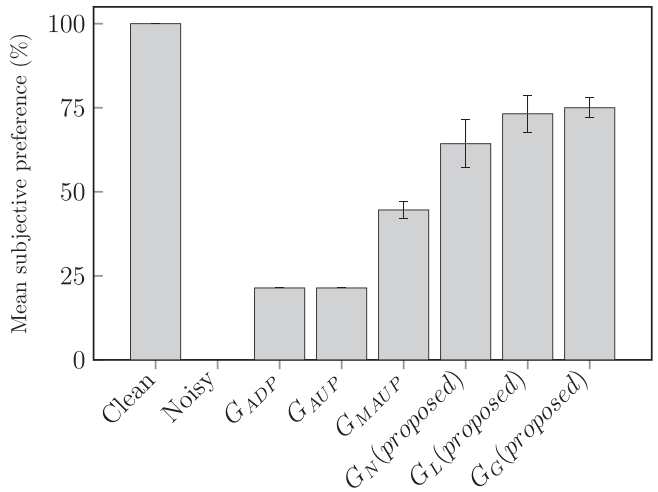

Fig. 17. The mean subjective preference score (%) comparison for each speech enhancement method. The female utterance (FF36_08) corrupted with 5 dB voice babble noise was used for the subjective tests. The error bars indicated the standard deviation of the scores.

proposed estimators with three SOTA phase-aware STSA estimators: the speech amplitude estimator given deterministic phase information (ADP) [51], the speech amplitude estimator given uncertain phase information (AUP) [27] with $\beta = \mu = 0.5$, and the modified AUP (MAUP) where the cost function includes both a power law and a weighting factor [28]. We also include the clean speech and noisy (unprocessed) speech as the upper bound and lower bound of the performance, respectively. Along with the conventional metrics such as PESQ, STOI, and SegSNR, three new phase-aware speech quality metrics are added to further evaluate the impact of phase modification: the phase deviation (PD), which is a distortion metric between the noisy phase and clean phase [52]; the unwrapped root mean square estimation error (UnRMSE) and unwrapped harmonic phase SNR (UnHPSNR), which both measured in decibels and focused on qualifying the estimation error of harmonic phase [53]. All the comparative estimators use the same or equivalent basic setup, meaning that the AMS setup, the *a prior* SNR estimation, and the noise PSD estimation are equal for all methods. For the ADP or the AUP estimator, the phase information was obtained by the model-based phase estimator [26], and for the MAUP estimator, the MAP phase estimator [31] was used. The Matlab implementations of above phase estimators are provided by PhaseLab toolbox [54].

The results obtained from the objective (Fig. 16) and and human listening tests (Fig. 17) show that the proposed estimators are providing better performance than the SOTA phase-aware estimators. The human listening tests can reliably quantify the character of speech quality, or estimate the speech quality achievable by an algorithm. It was reported that the phase-aware enhanced speech suffer from some buzzyness (Fig. 17). As a result, the utterance modified by the phase-aware methods (G_{ADP} %21.4, G_{AUP} %21.4, and G_{MAUP} %44.6) were much less preferred by the listeners than those enhanced by the proposed methods (G_G %75, G_L %73.2 and G_N %64.3). This buzzy quality was caused by the artifacts in the harmonic structure of the resultant speech, which reported earlier in [26,27,29]. These artifacts can be predicated as a degraded perceived speech quality (e.g., PESQ, SegSNR and PD) or intelligibility score (e.g., STOI, UnRMSE and UnHPSNR) (Fig. 16). The following observations can also be made:

- For high SNRs, the phase estimation is even disadvantageous as the required accuracy is very high. PD or SegSNR predicts highest perceived quality scores for the clean and lowest for the noisy at low SNRs, e.g., 0 and 5 SNRs; however, at high SNRs, e.g., 10 and 15 SNRs, it predicts the lowest for ADP and AUP (Fig. 16). Note that, PD (or UnRMSE) penalizes the harmonization artifacts by predicting a worse quality and thus the lower the score the better estimated quality.
- The phase-aware measure, UnRMSE, predicts the worst quality for AUP, where the phase structure and estimates are distorted. Despite that both ADP and AUP use the same model-based phase estimator, ADP obtained better UnRMSE scores than AUP. UnHPSNR predicts the worst results for both ADP and AUP, however, it predicts high quality for MAUP at high SNRs, e.g., 10 and 15 dB. It is important to note that the noisy phase at strong signal components, which mostly occur at low frequencies, is still similar to the clean phase. Therefore a reasonable value is predicted for the unprocessed speech, still lower than the clean speech as the upper bound.

Both subjective and objective results reveal that an accurate phase estimate is indispensable in order to benefit from the additional phase information. Incorporating an erroneous phase estimate strongly influence the performance of phase-aware estimators. On the other hand, Vary [25] derived the maximum phase deviation (defined as the difference between the clean and noisy spectral phase), $\hat{\phi}_{dev,max} \approx 0.679$ radians, roughly corresponding to the threshold of perception in phase distortion for an instantaneous SNR (ISNR) of 6 dB. Markedly, it was demonstrated that this critical threshold in DCT polarity spectrum is 0 dB ISNR [33], which is about 6 dB lower than the threshold in DFT phase spectrum. This means when the ISNR is above 0 dB, leaving the DCT polarity unmodified has no effect on perceived speech quality, and the effects solely come from modifying the DCT amplitude; however, modifying the DFT amplitude alone might not achieve the same improvement and thus, an accurate DFT phase estimation might be required.

7. Directions for future research

After many years in the shadow of DFT-based speech enhancement, the DCT STSA-estimation based approach is now burgeoning: with still many aspects to explore, it is an exciting research area that is likely to lead to breakthroughs and push speech enhancement forward. In our opinion, three main directions to follow are:

- employing perceptually motivated optimization criteria: such as the mean-square error of the log STSA [55], the weighted euclidean distortion measure [56], and the β-order distortion measure [57];
- developing a more precise *a prior* SNR estimator: although the presented SE methods all strongly depend on the reliability of the *a priori* SNR estimate, to the authors' best knowledge, all of the advanced *a priori* SNR estimators are formulated in the DFT domain [2];
- incorporating polarity information for DCT STSA estimation: since DCT coefficient is real, the issue caused by phase wrapping in the complex DFT analysis doesn't apply. Moreover, the DCT polarity components have only two possible states and thus, are mathematically easier to model and calculate than the DFT phases.

8. Conclusions

In this paper, we have derived estimators for single-channel speech enhancement in the DCT domain. Our approach is to optimally estimate the short-time spectral amplitude (STSA) of the DCT coefficients due to its major importance in speech perception. To achieve this, we have derived the MMSE STSA estimators, which are based on super-Gaussian speech priors and Gaussian noise model. The optimal STSA estimators were fine-tuned by incorporating speech presence uncertainty in the noisy spectral components. Moreover, we have derived the optimal MMSE estimator of the DCT noise PSD, to be used in conjunction with the new estimators. When applied to the TSP speech database [47] and simulated on a wide range of noisy conditions, both objective scores and subjective listening tests show that the proposed estimators not only provide better perceptual quality, but also introduce less distortions in the enhanced speech signal, relative to alternative methods. We found that the DCT-based MMSE algorithms generally perform better than their DFT counterparts, especially when blind noise estimation is used. Specifically, although having almost identical gain curves as described in Section 4.4, the proposed G_N (Gaussian PDF) outperformed G_{EM} ([6], complex Gaussian PDF) in both the objective and subjective tests. Similar to the Gaussian prior scenario, the objective and subjective scores suggests that the proposed G_L (Laplacian PDF) achieves a higher perceptual quality than G_{L-FSA} ([9], for $\gamma = 1$, complex Laplacian PDF). Subjective listening tests reveal that the residual noise level in the G_N and G_L estimation is lower than that in the G_{EM} and G_{L-FSA}, respectively. Prior work presented in Section 2 indicates that this is due to the DCT PoS, which is more capable of conserving the speech quality than the DFT PhS for the same level of global distortion. It is worth noting that, the proposed estimator, G_N (or G_L), and the DFT-based approach, G_{EM} (or G_{L-FSA}), are directly comparable techniques, with DCT and DFT the only difference and all other factors being equivalent. Consequently, the premise that the DCT can provide superior performance is justified. Undoubtedly, there are more advanced techniques for estimating the DFT noise PSD or the *a priori* SNR but this comparison shows that the DCT has potential for better performance overall if those techniques are adapted to it.

Regarding the priors we note that Laplacian and Gamma priors achieve much higher SegSNR scores, mainly due to better preservation of speech spectral components. As could be judged by subjective listening, G_L or G_G clearly gives better enhanced speech quality than G_N and G_{EM}. Listening results also reveal that the performance of G_L and G_G are comparable. Among the algorithms based on the Laplacian prior, G_L gives similar or better noise attenuation but less speech distortions than G_{L-CSC} and G_{L-FSA}. Note that the perceivable differences between the various estimators are generally small,

since the maximum suppression was limited to 0.1 for all methods. Comparing the outcome of the blind experiments to the results of the oracle experiments, we observe that the proposed estimators are more robust to errors in practical scenarios (i.e., the noise PSD must be estimated from the given noisy signal). Nevertheless, the oracle information about the noise PSD results in considerable improvements relative to the blind case. Thus, the algorithms can still benefit from more precise noise estimates.

Finally, when compared to phase-aware DFT-based STSA estimators, it was found that the proposed estimators offer better perceived speech quality and were widely preferred by the listeners. Future work will be towards on examining other optimization criteria, developing a more precise *a prior* SNR estimator, and incorporating DCT polarity information to exploit the full potential of DCT-based STSA estimation.

CRediT authorship contribution statement

Sisi Shi: Investigation, Methodology, Software, Visualization, Data curation, Writing - original draft, Writing - review & editing. **Kuldip Paliwal:** Conceptualization, Supervision, Project administration, Writing - review & editing. **Andrew Busch:** Supervision, Writing - review & editing.

Data availability

Data will be made available on request.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Subjective testing procedure

In this appendix, we describe the procedure used to obtain the subjective quality scores in Fig. 2. These tests were done in the form of AB listening tests [36], 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 play list 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.

Since use of the entire corpus was not feasible for human listening tests, two utterances (one from a male speaker and one from a female speaker) from the test set described in Section 6 were used. Each utterance was modified as described in [Sec. III, B-(b)] [33] for each value of SegSNR. Thus, a total of 18 modified utterances were generated for the subjective test, and since each stimuli pair was also played in reverse order, each participant scored 36 stimuli pairs. Each listening test is conducted in a separated session, in a quiet room using closed circumaural headphones at a comfortable listening level.

Appendix B. Derivation of (15)

Assuming Gaussian distributions for speech and noise spectral components, it follows that the PDF of Y is also a Gaussian and $\sigma_Y^2 = \sigma_X^2 + \sigma_D^2$. Upon substituting (8) and (14) into (13b) we obtain

$$|\widehat{X}| = \frac{\sigma_Y}{\sqrt{2\pi}\sigma_X\sigma_D}\left\{\int_0^\infty x\exp\left[-\frac{\left(x\sigma_Y^2-\sigma_X^2 Y\right)^2}{2\sigma_X^2\sigma_D^2\sigma_Y^2}\right]\mathrm{d}x + \int_0^\infty x\exp\left[-\frac{\left(x\sigma_Y^2+\sigma_X^2 Y\right)^2}{2\sigma_X^2\sigma_D^2\sigma_Y^2}\right]\mathrm{d}x\right\}. \tag{B.1}$$

By using [eq.2.33.1,2.33.6][38] we get from (B.1)

$$|\widehat{X}| = \frac{\sqrt{2}\sigma_X\sigma_D}{\sqrt{\pi}\sigma_Y}\exp\left(-\frac{\sigma_X^2 Y^2}{2\sigma_D^2\sigma_Y^2}\right) + \frac{\sigma_X^2 Y}{\sigma_Y^2}\mathrm{erf}\left(\frac{\sigma_X}{\sqrt{2}\sigma_D\sigma_Y}Y\right). \tag{B.2}$$

The equivalent form of $|\widehat{X}|$ as given in (15) is obtained from (B.2) by using (16), (17) and (18).

Appendix C. Derivation of (27)

Using the expression for the parabolic cylinder function from (26) and the relations [Th.9.212.2,9.212.3][38] yields

$$\begin{aligned}\mathscr{D}_{-\frac{3}{2}}(z) =& 2^{-\frac{3}{4}}e^{-\frac{z^2}{4}}\left\{\frac{\sqrt{\pi}}{\Gamma\left(\frac{5}{4}\right)}\Phi\left(\frac{3}{4},\frac{1}{2};\frac{z^2}{2}\right) - \frac{\sqrt{2\pi}z}{\Gamma\left(\frac{3}{4}\right)}\Phi\left(\frac{5}{4},\frac{3}{2};\frac{z^2}{2}\right)\right\}\\ =& 2^{-\frac{3}{4}}e^{-\frac{z^2}{4}}\left\{\frac{\sqrt{\pi}}{\Gamma\left(\frac{5}{4}\right)}\left[\Phi\left(-\frac{1}{4},-\frac{1}{2};\frac{z^2}{2}\right)+\frac{z^2}{2}\Phi\left(\frac{3}{4},\frac{3}{2};\frac{z^2}{2}\right)\right]\right.\\ &\left. - \frac{\sqrt{2\pi}z}{\Gamma\left(\frac{3}{4}\right)}\left[\Phi\left(\frac{1}{4},\frac{1}{2};\frac{z^2}{2}\right)+\frac{z^2}{6}\Phi\left(\frac{5}{4},\frac{5}{2};\frac{z^2}{2}\right)\right]\right\},\end{aligned} \tag{C.1}$$

and

$$\mathscr{D}_{-\frac{1}{2}}(z) = 2^{-\frac{1}{4}}e^{-\frac{z^2}{4}}\left\{\frac{\sqrt{\pi}}{\Gamma\left(\frac{3}{4}\right)}\Phi\left(\frac{1}{4},\frac{1}{2};\frac{z^2}{2}\right) - \frac{\sqrt{2\pi}z}{\Gamma\left(\frac{3}{4}\right)}\Phi\left(\frac{3}{4},\frac{3}{2};\frac{z^2}{2}\right)\right\}. \tag{C.2}$$

After substituting (C.1) and (C.2) into (25), and using [eq.9.6.2,13.6.3] [42], we obtain (27).

Appendix D. MMSE-based Noise Power Estimation

The DFT-based MMSE noise estimator proposed in [35] is commonly used as a baseline method for estimating the DFT noise power spectral density (PSD). It was derived under complex Gaussian distributions. It's important to use the same noise estimation approach when comparing the performance among different speech enhancement algorithms. As DCT is a real-valued transform, we cannot simply carry it over to the DCT domain, so an equivalent estimator must be developed. In this section, we derive the MMSE estimator of DCT noise PSD, to be comparable to the method proposed in [35]. The differences between the two noise estimators are summarised in D.3, Table 5.

The noise and speech DCT spectral coefficients are assumed to have a Gaussian distribution. Under SPU, the optimal MMSE estimation of the noise PSD is given by

$$E\left\{|D|^2|Y\right\} = E\left\{|D|^2|Y,H_0\right\}p(H_0|Y) + E\left\{|D|^2|Y,H_1\right\}p(H_1|Y) \tag{D.1}$$

where $p(H_1|Y)$ denotes the *a posteriori* SPP as in (31) and here $p(H_0|Y) = 1 - p(H_1|Y)$. The conditional probability terms can be interpreted as the smoothing factors between two estimators. In Section D.1, we will estimate the SPP with a fixed *a priori* SNR, ξ_{H1}, which is optimal in terms of the total probability error. Under hypothesis H_0, $|Y| = |D|$, we can approximate the preceding estima-

Table 5
Noise estimator comparisons.

	DCT Domain (proposed)	**DFT** Domain (as given in [35])
Prior Density	Gaussian	Complex Gaussian
$p(H_1\|Y)$	$\left\{1+\frac{p(H_0)}{p(H_1)}\sqrt{\xi_{H_1}+1}\exp\left[-\frac{\|Y\|^2\,\xi_{H_1}}{2\sigma_D^2(1+\xi_{H_1})}\right]\right\}^{-1}$	$\left\{1+\frac{p(H_0)}{p(H_1)}(\xi_{H_1}+1)\exp\left[-\frac{\|Y\|^2\,\xi_{H_1}}{\sigma_D^2(1+\xi_{H_1})}\right]\right\}^{-1}$
$\|y*\|$	$\sqrt{\sigma_D^2\left(\frac{\xi_{H_1}+1}{\xi_{H_1}}\right)\ln\left\{(\xi_{H_1}+1)\left[\frac{P(H_0)}{P(H_1)}\right]^2\right\}}$	$\sqrt{\sigma_D^2\left(\frac{\xi_{H_1}+1}{\xi_{H_1}}\right)\ln\left((\xi_{H_1}+1)\frac{P(H_0)}{P(H_1)}\right)}$
$P_f(\xi_{H_1})$	$\mathrm{erfc}\left(\sqrt{\mu(\xi_{H_1})}\right),\ \mu(\xi_{H_1})=\frac{1}{2}\left(\frac{\xi_{H_1}+1}{\xi_{H_1}}\right)\ln(\xi_{H_1}+1)$	$\left(\frac{p(H_0)}{p(H_1)}[\xi_{H_1}+1]\right)^{-\frac{\xi_{H_1}+1}{\xi_{H_1}}}$
$P_m(\xi_{H_1},\xi)$	$\mathrm{erf}\left(\sqrt{\frac{\mu(\xi_{H_1})}{1+\xi}}\right)$	$1-P_f(\xi_{H_1})^{\frac{\xi}{\xi+1}}$
$10\log(\xi_{H_1})$	15.6 dB	15 dB

tor, $E\{|D|^2|Y,H_0\}$, with the periodogram of the noisy speech $|Y|^2$. Furthermore, $P(|Y||H_0)$ has the folded-normal PDF [37]

$$P(|Y||H_0)=\frac{2}{\sqrt{2\pi\sigma_D^2}}\exp\left[-\frac{|Y|^2}{2\sigma_D^2}\right] \tag{D.2}$$

Under hypothesis H_1, $|Y|=|X+D|$, and since Y is a Gaussian variate with variance, $\sigma_Y^2=\sigma_X^2+\sigma_D^2$, the PDF for $P(|Y||H_1)$ is

$$P(|Y||H_1)=\frac{2}{\sqrt{2\pi\sigma_D^2(1+\xi_{H1})}}\exp\left[-\frac{|Y|^2}{2\sigma_D^2(\xi_{H1}+1)}\right] \tag{D.3}$$

where ξ is the *a priori* SNR. We can compute the optimal estimator for speech presence as

$$E\{|D|^2|Y,H_1\}=\frac{\int_{-\infty}^{\infty}|d|^2p(Y|d)p(d)\mathrm{d}d}{\int_{-\infty}^{\infty}p(Y|d)p(d)\mathrm{d}d} \tag{D.4}$$

From (8) and the additive and independence assumption of the speech and noise, it follows that the conditional PDF $p(Y|D)$ is given by

$$p(Y|D)=\frac{1}{\sqrt{2\pi}\sigma_X}\exp\left[\frac{-(Y-D)^2}{2\sigma_X^2}\right] \tag{D.5}$$

Substituting (D.5) and (7) into (D.4) and using [eq.3.46.8] [38], we find

$$E\{|D|^2|Y,\widehat{\xi},H_1\}=\frac{\widehat{\xi}}{\widehat{\xi}+1}\widehat{\sigma}_D^2+\frac{|Y|^2}{\left(\widehat{\xi}+1\right)^2} \tag{D.6}$$

Note that evaluation of (D.6) requires the estimates $\widehat{\xi}$ and $\widehat{\sigma}_D$ of the *a priori* SNR and noise variance, respectively. As suggested in [35], we use the noise power estimate of the previous frame, i.e., $\widehat{\sigma}_D^2=\widehat{\sigma}_D^2(l-1)$, and the maximum-likelihood (ML) estimate of the *a priori* SNR, i.e.,

$$\widehat{\xi}_{ml}=\widehat{\gamma}-1=\frac{|Y|^2}{\widehat{\sigma}_D^2(l-1)}-1 \tag{D.7}$$

for (D.6) and attain

$$E\{|D|^2|Y,\widehat{\xi},H_1\}=\widehat{\sigma}_D^2(l-1) \tag{D.8}$$

After making these approximations, the noise estimate in (D.1) takes the form

$$E\{|D|^2|Y\}=p(H_0|Y)|Y|^2+p(H_1|Y)\widehat{\sigma}_D^2(l-1) \tag{D.9}$$

Therefore, when the probability of speech being present in the noisy input is extremely low, i.e., $p(H_1|Y)\approx 0$, the noise estimate will follow the noisy speech periodogram, $|Y|^2$. Conversely, when $p(H_1|Y)\approx 1$, the noise update will cease and the noise estimate will remain the same as the previous frame's estimate. Following the preceding computation of $E\{|D|^2|Y\}$, the long-term noise PSD was then obtained via recursive smoothing with $\beta=0.95$

$$\widehat{\sigma}_D^2(l)=\beta\,\widehat{\sigma}_D^2(l-1)+(1-\beta)E\{|D|^2|Y\} \tag{D.10}$$

D.1. a posteriori SPP estimation

The smoothing factor $p(H_1|Y)$is a function of the likelihood ratio, Λ, as in (31). Under the Gaussian assumption, we get an expression for the conditional SPP

$$p(H_1|Y)=\left\{1+\frac{p(H_0)}{p(H_1)}\sqrt{\xi_{H_1}+1}\exp\left[-\frac{|Y|^2\,\xi_{H_1}}{2\sigma_D^2(1+\xi_{H_1})}\right]\right\}^{-1} \tag{D.11}$$

which follows from substitution of (35) into (31). Where ξ_{H_1} is a fixed model parameter for speech presence and is used to guarantee a desired performance in terms of false alarms and missed detections [10]. The idea of optimizing ξ_{H_1} for SPP estimation has been proposed the first time in [58]. In the next section we determine the optimal value of ξ_{H_1}.

D.2. Optimal ξ_{H_1} estimation

The optimal ξ_{H_1} is found by minimising the total probability of error [59]

$$P_\epsilon\triangleq p(H_0)P_f+p(H_1)P_m \tag{D.12}$$

where P_f and P_m denote the probabilities of false-alarm and missed-hit, respectively. Assuming that the two hypotheses are equally likely (a worst case assumption [10]), $P(H_1)=P(H_0)=\frac{1}{2}$. We specify P_f as the probability that $p(H_1|y)>\frac{1}{2}$ when speech is absent, and P_m as the probability that $p(H_1|y)<\frac{1}{2}$ when speech is present. To evaluate (D.12) numerically, we can express P_f and P_m in terms of $p(|Y||H_0)$ and $p(|Y||H_1)$, respectively

$$P_f(\xi_{H_1})=\int_{|y*|}^{\infty}p(|y||H_0)\mathrm{d}|y|=\mathrm{erfc}\left(\sqrt{\mu(\xi_{H_1})}\right) \tag{D.13}$$

$$P_m(\xi_{H_1},\xi)=\int_0^{|y*|}p(|y||H_1)\mathrm{d}|y|=\mathrm{erf}\left(\sqrt{\frac{\mu(\xi_{H_1})}{1+\xi}}\right) \tag{D.14}$$

where $|y*|$ is the point corresponding to $p(H_1|y)=0.5$ and results from (D.11)

$$|y*|=\sqrt{\sigma_D^2\left(\frac{\xi_{H_1}+1}{\xi_{H_1}}\right)\ln\left\{(\xi_{H_1}+1)\left[\frac{P(H_0)}{P(H_1)}\right]^2\right\}} \tag{D.15}$$

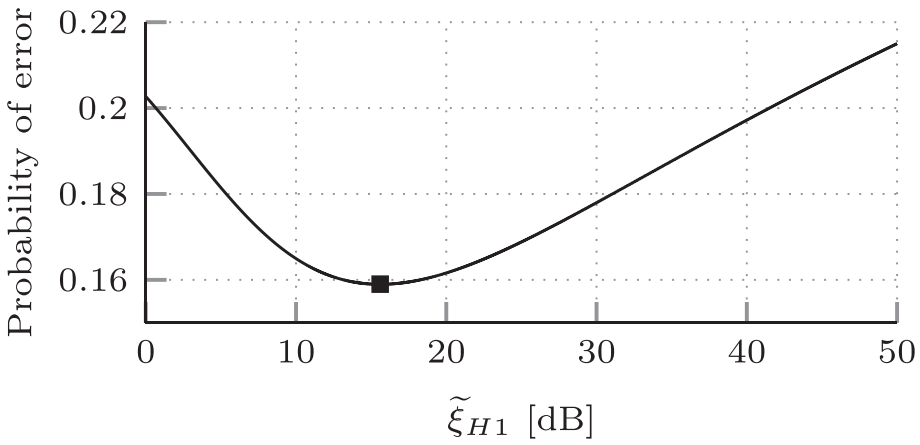

Fig. 18. The total probability of error as a function of $\widetilde{\xi}_{H1}$.

and $\mu(\xi_{H_1})$ is defined by

$$\mu(\xi_{H_1}) \triangleq \frac{|y*|^2}{2\sigma_D^2} = \frac{1}{2}\left(\frac{\xi_{H_1}+1}{\xi_{H_1}}\right)\ln(\xi_{H_1}+1) \tag{D.16}$$

The optimal ξ_{H_1} is found by minimizing the total probability of error when ξ is uniformly distributed between $\xi_{low}=0$ and $\xi_{up}=100$,

$$\xi_{H_1} = \min_{\widetilde{\xi}_{H_1}} \int_{\xi_{low}}^{\xi_{up}} p(H_0)P_f\left(\widetilde{\xi}_{H_1}\right) + p(H_1)P_m\left(\widetilde{\xi}_{H_1},\xi\right)\mathrm{d}\,\xi \tag{D.17}$$

Using [eq.4.1.1,4.1.2] [60], (D.17) can be expressed as shown in (D.18). The lower bound $10\log(\xi_{low}) = -\infty$ dB and upper bound $10\log(\xi_{up}) = 20$ dB are considered appropriate for a noise reduction application [35]. The expression given in (D.18) is plotted against $\widetilde{\xi}_{H_1}$ in Fig. 18. The minimum is at $10\log(\xi_{H_1}) = 15.6$ dB. This is the point of interest at which minimum P_ϵ is the criterion. Therefore, we use $10\log(\xi_{H_1}) = 15.6$ dB for (D.11), and throughout the algorithm.

$$\begin{aligned}\xi_{H_1} = \min_{\widetilde{\xi}_{H_1}} \frac{1}{2}\{&\xi_{up}P_f\left(\widetilde{\xi}_{H_1}\right) + [\left(2\mu\left(\widetilde{\xi}_{H_1}\right)+\xi_{up}+1\right)P_m\left(\widetilde{\xi}_{H_1},\xi_{up}\right)\\ &-\left(2\mu\left(\widetilde{\xi}_{H_1}\right)+1\right)\left(1-P_f\left(\widetilde{\xi}_{H_1}\right)\right)\\ &+\frac{2}{\sqrt{\pi}}\left(\sqrt{\mu\left(\widetilde{\xi}_{H_1}\right)(\xi_{up}+1)}\exp\left[-\frac{\mu\left(\widetilde{\xi}_{H_1}\right)}{\xi_{up}+1}\right]\right.\\ &\left.-\sqrt{\mu\left(\widetilde{\xi}_{H_1}\right)}\exp\left[-\mu\left(\widetilde{\xi}_{H_1}\right)\right]\right)\Big]\Big\}\end{aligned} \tag{D.18}$$

D.3. Noise estimator comparisons

The differences between the complex DFT noise estimator given in [35] and the noise estimator proposed in D are summarized in Table 5.

References

[1] Loizou PC, Introduction, in: Speech enhancement: theory and practice, 2nd Edition, CRC Press, Boca Raton, NW, USA, 2013, Ch. 1, pp. 1–2.
[2] Loizou PC, Statistical-model-based methods, in: Speech enhancement: theory and practice, 2nd Edition, CRC Press, Boca Raton, NW, USA, 2013, Ch. 7, pp. 209–227.
[3] Ahmed N, Natarajan T, Rao K. Discrete cosine transform. IEEE Trans Computers 1974;100(1):90–3.
[4] Quatieri TF. Discrete-time speech signal processing: principles and practice. Pearson Education India; 2006.
[5] Wang D, Lim J. The unimportance of phase in speech enhancement. IEEE Trans Acoust Speech Signal Process 1982;30(4):679–81.
[6] Ephraim Y, Malah D. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans Acoust Speech Signal Process 1984;32(6):1109–21.
[7] Martin R. Speech enhancement based on minimum mean-square error estimation and supergaussian priors. IEEE Trans Speech Audio Process 2005;13(5):845–56.
[8] Chen B, Loizou PC. A laplacian-based mmse estimator for speech enhancement. Speech Commun 2007;49(2):134–43.
[9] Erkelens JS, Hendriks RC, Heusdens R, Jensen J. Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors. IEEE Trans Audio, Speech, Language Processing 2007;15(6):1741–52.
[10] McAulay R, Malpass M. Speech enhancement using a soft-decision noise suppression filter. Acoust, Speech Signal Process, IEEE Trans 1980;28 (2):137–45.
[11] Soon Y, Koh SN, Yeo CK. Noisy speech enhancement using discrete cosine transform. Speech Commun 1998;24(3):249–57.
[12] Ding H, Soon Y, Yeo CK. A dct-based speech enhancement system with pitch synchronous analysis. IEEE Trans Audio, Speech, Language Process 2011;19 (8):2614–23.
[13] Soon I, Koh S. Low distortion speech enhancement. IEE Proceedings-Vision, Image Signal Process 2000;147(3):247–53.
[14] Erkelens JS, Hendriks RC, Heusdens R. On the estimation of complex speech dft coefficients without assuming independent real and imaginary parts. IEEE Signal Process Lett 2008;15:213–6. https://doi.org/10.1109/LSP.2007.911730.
[15] Andrianakis I, White PR. Speech spectral amplitude estimators using optimally shaped gamma and chi priors. Speech Commun 2009;51(1):1–14.
[16] Hasan T, Hasan MK. Mmse estimator for speech enhancement considering the constructive and destructive interference of noise. IET Signal Process 2010;4 (1):1–11.
[17] Mahmmod BM, Ramli AR, Abdulhussian SH, Al-Haddad SAR, Jassim WA. Low-distortion mmse speech enhancement estimator based on laplacian prior. IEEE Access 2017;5:9866–81.
[18] Zou X, Zhang X. Speech enhancement using an mmse short time dct coefficients estimator with supergaussian speech modeling. J Electron 2007;24(3):332–7.
[19] Aroudi A, Veisi H, Sameti H. Speech enhancement based on hidden markov model with discrete cosine transform coefficients using laplace and gaussian distributions. In: 2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA). IEEE; 2012. p. 304–9.
[20] Aroudi A, Veisi H, Sameti H. Hidden markov model-based speech enhancement using multivariate laplace and gaussian distributions. IET Signal Proc 2015;9 (2):177–85.
[21] Hamidi M, Pearl J. Comparison of the cosine and fourier transforms of markov-1 signals. IEEE Trans Acoust Speech Signal Process 1976;24(5):428–9.
[22] Gazor S, Zhang W. Speech probability distribution. IEEE Signal Process Lett 2003;10(7):204–7.
[23] Aroudi A, Veisi H, Sameti H, Mafakheri Z. Speech signal modeling using multivariate distributions. EURASIP J Audio, Speech, Music Processing 2015;2015(1):1–14.
[24] Hendriks RC, Gerkmann T, Jensen J. Dft-domain based single-microphone noise reduction for speech enhancement: A survey of the state of the art. Synthesis Lectures Speech Audio Process 2013;9(1):1–80.
[25] Vary P, Eurasip M. Noise suppression by spectral magnitude estimation-mechanism and theoretical limits-. Signal Processing 1985;8(4):387–400.
[26] Krawczyk M, Gerkmann T. Stft phase reconstruction in voiced speech for an improved single-channel speech enhancement. IEEE/ACM Trans Audio, Speech, Language Process 2014;22(12):1931–40.

[27] Krawczyk-Becker M, Gerkmann T. On mmse-based estimation of amplitude and complex speech spectral coefficients under phase-uncertainty. IEEE/ACM Trans Audio, Speech, Language Process 2016;24(12):2251–62.
[28] Samui S, Chakrabarti I, Ghosh SK, Auditory model based phase-aware bayesian spectral amplitude estimator for single-channel speech enhancement, in: arXiv preprint arXiv:2202.04882, 2022.
[29] Mowlaee P, Kulmer J, Stahl J, Mayer F. Single channel phase-aware signal processing in speech communication: theory and practice. John Wiley & Sons; 2016.
[30] Mowlaee P, Kulmer J. Phase estimation in single-channel speech enhancement: Limits-potential. IEEE/ACM Trans Audio, Speech, Language Process 2015;23(8):1283–94.
[31] Mowlaee P, Kulmer J. Harmonic phase estimation in single-channel speech enhancement using phase decomposition and snr information. IEEE/ACM Trans Audio, Speech, Language Process 2015;23:1521–32.
[32] Mowlaee P, Saiedi R, Martin R, Phase estimation for signal reconstruction in single-channel speech separation, in: Proceedings of the International Conference on Spoken Language Processing, 2012, pp. 1–4.
[33] Shi S, Busch A, Paliwal K, Fickenscher T. On the use of discrete cosine transform polarity spectrum in speech enhancement. In: 2020 28th European Signal Processing Conference (EUSIPCO). IEEE; 2021. p. 421–5.
[34] Rix AW, Beerends JG, Hollier MP, Hekstra AP. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat No. 01CH37221), Vol. 2. IEEE; 2001. p. 749–52.
[35] Gerkmann T, Hendriks RC. Unbiased mmse-based noise power estimation with low complexity and low tracking delay. IEEE Trans Audio, Speech, Language Process 2011;20(4):1383–93.
[36] So S, Paliwal KK. Modulation-domain kalman filtering for single-channel speech enhancement. Speech Commun 2011;53(6):818–29.
[37] Johnson NL, Kotz S, Balakrishnan N. Continuous univariate distributions. New York: Wiley; 1994.
[38] Gradshteyn IS, Ryzhik IM. Table of integrals, series, and products. Academic press; 2007.
[39] Wiener N, Extrapolation, interpolation and smoothing of stationary, Time Series, with Engineering Appli cations.
[40] Cohen I. Relaxed statistical model for speech enhancement and a priori snr estimation. IEEE Trans Speech Audio Process 2005;13(5):870–81.
[41] Hendriks R, Toolbox for mmse estimators of dft coefficients under the generalized gamma density [cited 18.03.2022]. https://www.mathworks.com/matlabcentral/fileexchange/25408-toolbox-for-mmse-estimators-of-dft-coefficients-under-the-generalized-gamma-density.
[42] Abramowitz M, Stegun IA, Handbook of mathematical functions with formulas, graphs, and mathematical table, in: US Department of Commerce, National Bureau of Standards Applied Mathematics series 55, 1964.
[43] Cappé O. Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor. IEEE Trans Speech Audio Process 1994;2(2):345–9.
[44] Middleton D, Esposito R. Simultaneous optimum detection and estimation of signals in noise. IEEE Trans Inform Theory 1968;14(3):434–44.
[45] Loizou PC, Statistical-model-based methods, in: Speech enhancement: theory and practice, 2nd Edition, CRC Press, Boca Raton, NW, USA, 2013, Ch. 7, pp. 257–258.
[46] Loizou PC, Statistical-model-based methods, in: Speech enhancement: theory and practice, 2nd Edition, CRC Press, Boca Raton, NW, USA, 2013, Ch. 7, p. 263.
[47] Kabal P, Tsp speech database, McGill University, Database Version 1 (0) (2002) 09–02.
[48] Varga A, Steeneken HJ. 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 1993;12(3):247–51.
[49] Loizou PC, Objective quality and intelligibility measures, in: Speech enhancement: theory and practice, 2nd Edition, CRC Press, Boca Raton, NW, USA, 2013, Ch. 11, pp. 502–503.
[50] Taal CH, Hendriks RC, Heusdens R, Jensen J. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. IEEE Trans Audio, Speech, Language Process 2011;19(7):2125–36.
[51] Gerkmann T, Krawczyk M. Mmse-optimal spectral amplitude estimation given the stft-phase. IEEE Signal Process Lett 2013;20(2):129–32.
[52] Gaich A, Mowlaee P. On speech quality estimation of phase-aware single-channel speech enhancement. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2015. p. 216–20.
[53] Gaich A, Mowlaee P, On speech intelligibility estimation of phase-aware single-channel speech enhancement, in: Sixteenth Annual Conference of the International Speech Communication Association, 2015.
[54] Mowlaee P, Phaselab toolbox [cited 30.09.2022]. https://www2.spsc.tugraz.at/people/pmowlaee/PhaseLab.
[55] Ephraim Y, Malah D. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans Acoust Speech Signal Process 1985;33(2):443–5.
[56] Loizou PC. Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum. IEEE Trans Speech Audio Process 2005;13(5):857–69.
[57] You CH, Koh SN, Rahardja S. /spl beta/-order mmse spectral amplitude estimation for speech enhancement. IEEE Trans Speech Audio Process 2005;13(4):475–86.
[58] Gerkmann T, Breithaupt C, Martin R. Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors. IEEE Trans Audio, Speech, Language Process 2008;16(5):910–9.
[59] Van Trees HL. Detection, estimation, and modulation theory, part I: detection, estimation, and linear modulation theory. John Wiley & Sons; 2004. Ch. 2.
[60] Ng EW, Geller M. A table of integrals of the error functions. J Res National Bureau Standards B 1969;73(1):1–20.