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

Sisi Shi, Kuldip Paliwal, Andrew Busch

Signal Processing Laboratory, Griffith School of Engineering, Griffith University, Nathan QLD 4111, Australia

School of Engineering, Griffith University, Nathan QLD 4111, Australia

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
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].

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 to account the Speech Presence Uncertainty (SPU) in the noisy observers implies the amplitudes and phases of DFT coefficients are statistically 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].

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.

### Table 1

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Spectral Amplitude</th>
<th>Coefficient</th>
<th>Spectral Amplitude (proposed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>super-Gaussian prior [7]</td>
<td>complex Laplacian prior [8]*</td>
<td>( G_\lambda ), Laplacian, refer to [22]</td>
<td></td>
</tr>
<tr>
<td>generalized Gamma [14]</td>
<td>( G_{\lambda, \text{SEM}} ), generalized Gamma [9]<em>[15]</em></td>
<td>( g_c ), Gamma, refer to [27]</td>
<td></td>
</tr>
<tr>
<td>( G_\text{Wiener} ), Wiener filter [11,12]</td>
<td>( G_{\text{DCW}}, \text{dual-gain Wiener} [13] )</td>
<td>( G_{\text{NBLG}}, \text{dual-gain MMSE, Laplacian prior} [17] )</td>
<td></td>
</tr>
<tr>
<td>( G_{\text{DCW}}, \text{dual-gain Wiener} [13] )</td>
<td>( G_{\text{DGW}}, \text{dual-gain MMSE, Laplacian prior} [16] )</td>
<td>( G_{\text{GL}}, \text{super-Gaussian} [18] )</td>
<td></td>
</tr>
<tr>
<td>Generalized Gamma [14]</td>
<td>( G_{\text{GL}}, \text{super-Gaussian} [18] )</td>
<td>( G_{\text{GL}}, \text{super-Gaussian} [18] )</td>
<td></td>
</tr>
</tbody>
</table>

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. These 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 is 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
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:

\[
\begin{align*}
\text{S. Shi, K. Paliwal and A. Busch Applied Acoustics 202 (2023) 109134}
\end{align*}
\]
y(n) = x(n) + d(n), \quad 0 \leq n \leq N - 1 \quad (1)

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

\[ Y(i, k) = \sum_{n=0}^{N-1} y(n + iN_i)w(n) \cos \left( \frac{(2n + 1)k\pi}{2L} \right) \quad (2) \]

where \( 0 \leq k \leq L - 1 \) and:

\[ c_{mk} = \begin{cases} \frac{1}{\sqrt{L}} & \text{for } k = 0 \\ \frac{1}{\sqrt{2}} & \text{for } k \neq 0 \end{cases} \quad (3) \]

\( n, k \) and \( i \) are the discrete time, frequency and frame index, respectively, \( w(n) \) is the analysis window function of length \( N_w \), \( N_i \) 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:

\[ \phi_Y|Y| = \phi_X|X| + \phi_D|D| \quad (4) \]

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 \( |Y| \) from the degraded signal \( Y \). The optimal solution can be computed as a gain function multiplied by the noisy DCT STSA:

\[ |\hat{X}| \triangleq G(\cdot, \cdot)|Y| \quad (5) \]

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., \( \hat{X} \). 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, \( |\hat{X}| \), with the noisy PoS, \( \phi_Y \), to get the final estimate of the spectral component

\[ \hat{X} \triangleq \phi_Y|\hat{X}| \triangleq G(\cdot, \cdot)Y \quad (6) \]

The Gaussian assumption holds for the distribution of the noise coefficients:

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

where \( p(\cdot) \) denotes the probability density function (PDF) and \( \sigma_D^2 \) 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

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

- Laplacian speech prior

\[ p(X) = \frac{1}{\sqrt{2\sigma_X}} \exp \left( -\frac{\sqrt{2}|X|}{\sigma_X} \right) \quad (9) \]

- Double-sided Gamma speech prior

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

where \( \sigma_X \) and \( \sigma_X^2 \) 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]:

\[ |\hat{X}| = E(|X| | Y) = \int_0^\infty |X|p(|X| | Y)d|X| \quad (11) \]

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:

\[ p(|X| = x) = p(X = x) + p(X = -x) \quad (12) \]

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

\[ p(|X| | Y) = \frac{1}{\sqrt{2\pi \sigma_X}} \exp \left( -\frac{|X|^2}{2\sigma_X^2} \right) \quad (13) \]
\[
|\hat{X}| = \int_{-\infty}^{\infty} |x|f(x)dx
\]  
(13a)

\[
|\hat{X}| = \frac{1}{2\pi\sigma_x} \exp\left[-\frac{(Y - X)^2}{2\sigma_x^2}\right] 
\]  
(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

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

\[
|\hat{X}| = \frac{\xi}{1 + \xi} \left\{ \frac{2}{\pi\nu} \exp\left(-\frac{\nu^2}{2}\right) + \text{erf}\left(\frac{\nu}{\sqrt{2}}\right) \right\} |Y|
\]

\[\Delta = G_n(\xi, \gamma)|Y| \]  
(15)

where \( \text{erf}() \) denotes the error function [eq.8.250.1][38] and \( \nu \) is defined by

\[
\nu = \frac{\xi}{1 + \xi} \gamma
\]

(16)

where \( \gamma \) and \( \xi \) are the a priori and a posteriori signal-to-noise (SNR), respectively and are defined by [6]

\[
\xi = \frac{E\{X^2\}}{E\{D^2\}} - \frac{\sigma_x^2}{\sigma_D^2}
\]

\[
\gamma = \frac{\sqrt{\sigma_x^2}}{\sqrt{\sigma_D^2}}
\]

(17)

(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]

\[
G_W = \frac{\xi}{1 + \xi}
\]

(19)

and it is independent of the a posteriori SNR, \( \gamma \). Therefore, \( G_W \) 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 priory, 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:

\[
\hat{\xi}(i) = \max\left\{ a_i \sqrt{\frac{|\hat{X}(i-1)|^2}{\sigma_D^2(i-1)}} + (1 - a_i) \max[\gamma(i-1) - 0], \hat{\xi}_\text{min} \right\}
\]

(20)

where \( |\hat{X}(i-1)| \) and \( \sigma_D^2(i-1) \) are the estimates of the spectral amplitude and the noise variance in the past frame, respectively. The max( ) operator denotes the maximum function to ensure the positiveness of the estimator, while \( a_i = 0.98 \) (was determined by simulations and informal listening tests in [6]) is the smoothing factor and \( \hat{\xi}_\text{min} = -25 \text{dB} \) is the SNR floor value for eliminating low-level musical noise [40].

The relation between \( G_n \) and \( G_{\text{MMSE}} \) [6] can be observed from their respective gain curves. The closeness of the gain curves, which correspond to the same value of \( \hat{\xi} \), implies that \( G_n \) and \( G_{\text{MMSE}} \) are nearly equivalent (Fig. 3). However, despite their similarity in behavior, \( G_n \) may yield better perceived speech quality than \( G_{\text{MMSE}} \). Because \( G_n \) or \( G_{\text{MMSE}} \) 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 \( G_{\text{MMSE}} \) 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:

\[
L_+ = \frac{\sigma_D^2}{\sigma_x^2} + \frac{|Y|^2}{\sqrt{2\sigma_D^2}} - \frac{1}{\xi} + \sqrt{\frac{2}{\gamma}}
\]

(21a)

\[
L_- = \frac{\sigma_D^2}{\sigma_x^2} + \frac{|Y|^2}{\sqrt{2\sigma_D^2}} - \frac{1}{\xi} - \sqrt{\frac{2}{\gamma}}
\]

(21b)

\[
M_+ = \exp(\frac{L_+}{2}) \cdot \text{erfc}(L_+)
\]

(21c)

\[
M_- = \exp(\frac{L_-}{2}) \cdot \text{erfc}(L_-)
\]

(21d)

where \( \text{erfc}(\cdot) \) 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:

\[
|\hat{X}| = \frac{2}{\sqrt{\pi}} \left\{ \frac{L_+ M_+ + L_- M_-}{M_+ + M_-} \right\} |Y|
\]

(22)

and the DD approach is used to estimate \( \hat{\xi}' \)

\[
\hat{\xi}'(i) = \max\left\{ a_i \sqrt{\frac{|\hat{X}(i-1)|^2}{\sigma_D^2(i-1)}} + (1 - a_i) \sqrt{\gamma(i-1) - 0}, \hat{\xi}'_\text{min} \right\}
\]

(23)

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

Note that the equivalent of \( G_n \) 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] \( G_{\text{L-FSA}} \) matches closely to the DCT estimator \( G_i \) for low a priori SNRs (-20 dB \( \leq \xi \leq -10 \text{ dB} \)). As \( \xi \) increases, the value of \( G_{\text{L-FSA}} \) is up to 5 \text{ dB} greater than \( G_i \) (-10 dB \( \leq \xi \leq -5 \text{ dB} \)). However, \( G_{\text{L-FSA}} \) no longer offers any noise attenuation when \( \xi \) 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] \( G_{\text{L-CSC}} \) is almost identical to its DFT counterpart [14] \( G_{\text{L-FSA}} \), for all SNR conditions. It can be seen that, the
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

\[ A_+ = \frac{\sqrt{3} \sigma_0}{2 \sigma_x} \frac{|y|}{\|y\|} = \frac{\sqrt{3}}{2 \varepsilon} + \sqrt{\gamma} \]

(24a)

\[ A_- = \frac{\sqrt{3} \sigma_0}{2 \sigma_x} \left( \frac{|y|}{\|y\|} - 1 \right) = \frac{\sqrt{3}}{2 \varepsilon} - \sqrt{\gamma} \]

(24b)

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

\[ \hat{X} = \frac{\sigma_0}{2} \left\{ \exp \left( \frac{\sqrt{3}}{2} \varepsilon \right) \hat{p}_A(A_+) + \exp \left( \frac{\sqrt{3}}{2} \varepsilon \right) \hat{p}_A(A_-) \right\} \]

(25)

where \( \hat{p}_A(z) \) denotes the parabolic cylinder function defined as [Th.9.240] [38]

\[ \hat{p}_A(z) = 2^{1/4} \left\{ \frac{\sqrt{\pi}}{\Gamma(1/2)} \Phi \left( -\frac{1}{2}, 1, \frac{3}{2} \right) - \frac{2 \pi z}{\sqrt{2}} \Phi \left( 1, \frac{3}{2}, \frac{3}{2} \right) \right\} \]

(26)

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)

\[ |\hat{X}| = (2\sqrt{\gamma})^{-1} \left\{ \exp \left( \frac{\sqrt{3}}{2} \varepsilon \right) S + \exp \left( \frac{\sqrt{3}}{2} \varepsilon \right) U \right\} \]

\[ \hat{p}_A \left( \gamma, \frac{\varepsilon}{\gamma} \right) |Y| \]

(27)

where

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

(28a)

\[ T = \frac{A_+^2}{A_-^2} \left[ K_\frac{1}{2} \left( \frac{A_-^2}{4} \right) \right] \]

(28b)

\[ U = \frac{\pi}{\sqrt{2}} |A_-|^2 \left\{ I_{\frac{1}{2}} \left( \frac{A_-^2}{4} \right) + I_{\frac{1}{2}} \left( \frac{A_+^2}{4} \right) - \phi_\varepsilon \left[ I_{\frac{1}{2}} \left( \frac{A_-^2}{4} \right) + I_{\frac{1}{2}} \left( \frac{A_+^2}{4} \right) \right] \right\} \]

(28c)

\[ V = \frac{\pi}{\sqrt{2}} |A_-|^2 \left\{ I_{\frac{1}{2}} \left( \frac{A_-^2}{4} \right) - \phi_\varepsilon \left[ I_{\frac{1}{2}} \left( \frac{A_-^2}{4} \right) \right] \right\} \]

(28d)

where \( \phi_\varepsilon \equiv \text{sgn}(A_-), I_{\frac{1}{2}}(z) \) and \( K_{\frac{1}{2}}(z) \) denote the modified Bessel functions of the first and second kind, respectively [eq.9.6.1] [42].

Fig. 5 illustrates the gain
characteristics for the Gamma speech model. We find a behaviour similar to the case of Laplacian speech prior. Due to inaccurate approximations, $G_{L/C0}$ 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 $G_{L/C0}$ and the approximated DFT spectral amplitude estimator $[9]$ is generally small (within 1 dB) for low a priori SNRs. This suggests that $G_L$ (or $G_{GL}$) can be used as a computationally simpler alternative to the DFT MMSE estimator $G_{L/C0}$ 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, $\gamma$, and the instantaneous SNR (ISNR), $\gamma - 1$, to describe the whole variations of the gain characteristics. Note that the relationship between $\xi$ and $\xi'$ can be easily obtained as

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

(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 $G_{EM}$ (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, $G_C$ (or $G_L$) provides significantly less attenuation than the Wiener filter due to the heavy-tailed (e.g., leptokurtic) speech prior. 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
statistical independence assumption of the spectral components in Section 3, the MMSE estimator that incorporates speech presence uncertainty (SPU) is given by
\[ j_b X j = E j X j j_{Yj} H_1 \text{fg} p_{H_1}(Y) \] \( \text{(30)} \)

The a posteriori speech presence probability (SPP) can be obtained using Bayes’ rule
\[ p(H_1|Y) = \frac{\Lambda}{\Lambda + 1} \] \( \text{(31)} \)

where \( \Lambda \) denotes the generalized likelihood ratio
\[ \Lambda = \frac{p(Y|H_1)}{p(Y|H_0)} \mu \] \( \text{(32)} \)

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 \( \sigma_D^2 \), it follows that
\[ p(Y|H_0) = \frac{1}{\sqrt{2\pi}\sigma_D} \exp \left( -\frac{Y^2}{2\sigma_D^2} \right) \] \( \text{(33)} \)

Under hypothesis \( H_1 \), \( Y = X + D \), we have
\[ p(Y|H_1) = \int_{-\infty}^{\infty} p(Y|X)p(X)dX \] \( \text{(34)} \)

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
\[ \Lambda_n(\xi, \gamma) = \exp \left( \frac{\nu}{2} \right) \mu \] \( \text{(35)} \)

Therefore, The amplitude estimator (30) can be written as
\[ |\tilde{X}| = \frac{\Lambda_n(\xi, \gamma)}{\Lambda_n(\xi, \gamma) + 1} G_n(\xi, \gamma)|Y| \] \( \text{(36)} \)

where \( G_n(\cdot, \cdot) \) is defined by (15). Similarly, for the Laplacian speech model
\[ |\tilde{X}| = \frac{\Lambda_l(\xi, \gamma)}{\Lambda_l(\xi, \gamma) + 1} G_l(\xi, \gamma)|Y| \] \( \text{(37)} \)

where \( G_l(\cdot, \cdot) \) is defined by (22) and
\[ c(\xi) = \sqrt{\frac{\pi}{\xi}} \left\{ \exp\left( \frac{\xi^2}{4} \right) + \exp\left( \frac{\xi^2}{4} \right) \right\} \mu \]  

where \( C(\xi, \gamma) \) is defined by (27) and

\[ \Lambda_c(\xi', \gamma) = \frac{\sqrt{\pi}}{2} (M_+ + M_-) \mu \]  

For the Gamma speech model

\[ \hat{X} = \frac{A_c(\xi', \gamma)}{A_c(\xi', \gamma) + 1} G_c(\xi', \gamma) \]  

\[ \triangleq G_{\text{SPU}}(\xi', \gamma) \]  

where \( G(\xi', \gamma) \) is defined by (27) and

\[ \Lambda_c(\xi', \gamma) = \frac{\sqrt{\frac{3}{4}}}{4\sqrt{\pi}} \left\{ \exp\left( \frac{\xi'^2}{4} \right) + \exp\left( \frac{\xi'^2}{4} \right) \right\} \mu \]  

Note that the original definition of \( \xi \) or \( \xi' \) was unconditional, whereas \( \xi \) or \( \xi' \) 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 \( C_{\text{SPU}} \) and \( C_{\text{EM}} \) are illustrated in Fig. 8, along with the corresponding curves of the Ephraim and Malah MMSE STSA estimator with SPU weighting \([30][6] \) as a comparison. The gain functions for \( q = 0.2 \) are plotted against the \( a \) priori SNR and the ISNR. It shows \( C_{\text{EM}} \) of (36), follows the Ephraim and Malah 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 \( C_{\text{EM}} \) are included for reference. It can be seen that \( C_{\text{EM}} \) yields around 5 dB higher attenuation than \( C_{\text{SPU}} \) 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 \( \xi \) 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 converge to the Wiener filter as ISNR increase. As ISNR decreases, \( C_{\text{SPU}} \) gives increased attenuation for which case \( q = 0.2 \), whereas \( C_{\text{EM}} \) has decreased attenuation for which case \( q = 0 \). For unfavorable conditions, \( C_{\text{SPU}} \) gives increased attenuation for which case \( q = 0.2 \), whereas \( C_{\text{EM}} \) has decreased attenuation for which case \( q = 0 \). For unfavorable...
acoustic conditions (i.e., $\xi = -5, -15$ dB), the gain curves of $G_{\text{L}}^{\text{SPT}}$ shows increased attenuation compared with those of $G_{N}$. 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. 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]) [4] $G_{\text{EM}}$ (complex Gaussian speech prior), are respectively plotted with dash-dotted and loosely dotted lines for comparison.](image-url)
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

$$\text{SegSNR} = \frac{10}{M} \sum_{i=0}^{M-1} \log_{10} \frac{\sum_{n=N_w}^{N_w-1} x^2(n)}{\sum_{n=-N_w}^{N_w-1} [x(n) - \hat{x}(n)]^2}$$  \hspace{1cm} (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. 8. Gain curves for the MMSE STSA estimators incorporating speech presence uncertainty (SPU) with $q = 0.2$. (a) $G_{SPU}^{GN}$ (Gaussian speech prior) defined by (36). (b) $G_{SPU}^{GL}$ (Laplacian speech prior) defined by (37), (c) $G_{SPU}^{GG}$ (Gamma speech prior) defined by (39), and (d) the respective Ephraim and Malah solution $G_{SPU}^{EM}$ (complex Gaussian speech prior), as seen in [30] [6].
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 (%), 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 GW and GSPW, respectively) [11]; the Laplacian-based MMSE DCT spectral coefficient estimator [18] implemented without and with SPU (cases GL and GSPL, 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 GEM and GSPM, respectively); MMSE DFT spectral amplitude estimator [9,41], with $\nu = 1$ (case $G_{L-ESA}$).

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

$$\hat{\sigma}_b^2 = |D|^2$$

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.
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), $G_{L-CSC}$, $G_t$ and $G_c$ 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 $G_t$, closely followed by $G_c$, $G_n$, and $G_w$. As compared to the oracle case, $G_t$ 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, $G_t$, $G_c$, 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 increases with increase in input SNR. $G_t$ and $G_c$ give about 1% and 2% higher STOI gain over $G_{L-CSC}$ for SNRs less than 10 dB, respectively. Finally, $G_t$ and $G_c$ yield the highest SegSNR gain for all SNRs. $G_t$ 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 $G_{SPU,L}$ and $G_{SPU,G}$ over $G_t$ and $G_c$ for input SNR in the range 0 dB to 5 dB. Whereas an intelligibility drop is suggested for $G_{SPU,EM}$ 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 $G_t$ (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 $G_c$ (Gamma PDF) performs best]. For the voice babble noise, the proposed $G_n$ (Gaussian PDF) and $G_t$ obtained the highest PESQ and SegSNR scores, respectively. The objective intelligibility scores attained by each method are given in Table 4. The proposed $G_t$ 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.

<table>
<thead>
<tr>
<th>Noise</th>
<th>Method</th>
<th>0 dB</th>
<th>5 dB</th>
<th>10 dB</th>
<th>15 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-16 two seat</td>
<td>GEM</td>
<td>0.071</td>
<td>0.212</td>
<td>0.567</td>
<td>0.770</td>
</tr>
<tr>
<td></td>
<td>GL</td>
<td>0.084</td>
<td>0.228</td>
<td>0.575</td>
<td>0.709</td>
</tr>
<tr>
<td></td>
<td>GW</td>
<td>0.121</td>
<td>0.375</td>
<td>0.689</td>
<td>0.990</td>
</tr>
<tr>
<td></td>
<td>GL-CSC</td>
<td>0.069</td>
<td>0.278</td>
<td>0.537</td>
<td>0.896</td>
</tr>
<tr>
<td></td>
<td>DGW</td>
<td>0.071</td>
<td>0.282</td>
<td>0.545</td>
<td>0.923</td>
</tr>
<tr>
<td></td>
<td>NBLG</td>
<td>0.098</td>
<td>0.315</td>
<td>0.628</td>
<td>0.989</td>
</tr>
<tr>
<td></td>
<td>G_C</td>
<td>0.123</td>
<td>0.377</td>
<td>0.694</td>
<td>1.076</td>
</tr>
<tr>
<td></td>
<td>G_C</td>
<td>0.140</td>
<td>0.385</td>
<td>0.696</td>
<td>1.086</td>
</tr>
<tr>
<td>Voice babble</td>
<td>GEM</td>
<td>0.039</td>
<td>0.150</td>
<td>0.383</td>
<td>0.377</td>
</tr>
<tr>
<td></td>
<td>GL</td>
<td>0.026</td>
<td>0.110</td>
<td>0.285</td>
<td>0.188</td>
</tr>
<tr>
<td></td>
<td>GW</td>
<td>0.044</td>
<td>0.161</td>
<td>0.271</td>
<td>0.239</td>
</tr>
<tr>
<td></td>
<td>GL-CSC</td>
<td>0.024</td>
<td>0.117</td>
<td>0.211</td>
<td>0.295</td>
</tr>
<tr>
<td></td>
<td>DGW</td>
<td>0.054</td>
<td>0.162</td>
<td>0.336</td>
<td>0.438</td>
</tr>
<tr>
<td></td>
<td>NBLG</td>
<td>0.067</td>
<td>0.208</td>
<td>0.452</td>
<td>0.528</td>
</tr>
<tr>
<td></td>
<td>G_C</td>
<td>0.075</td>
<td>0.244</td>
<td>0.465</td>
<td>0.627</td>
</tr>
<tr>
<td></td>
<td>G_C</td>
<td>0.052</td>
<td>0.191</td>
<td>0.306</td>
<td>0.498</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Noise</th>
<th>Method</th>
<th>0 dB</th>
<th>5 dB</th>
<th>10 dB</th>
<th>15 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-16 two seat</td>
<td>GEM</td>
<td>3.974</td>
<td>4.333</td>
<td>2.184</td>
<td>0.917</td>
</tr>
<tr>
<td></td>
<td>GL</td>
<td>4.012</td>
<td>4.385</td>
<td>2.229</td>
<td>0.861</td>
</tr>
<tr>
<td></td>
<td>GW</td>
<td>5.258</td>
<td>4.997</td>
<td>3.558</td>
<td>1.901</td>
</tr>
<tr>
<td></td>
<td>GL-CSC</td>
<td>5.278</td>
<td>5.408</td>
<td>4.327</td>
<td>2.687</td>
</tr>
<tr>
<td></td>
<td>DGW</td>
<td>4.843</td>
<td>4.609</td>
<td>3.172</td>
<td>1.282</td>
</tr>
<tr>
<td></td>
<td>NBLG</td>
<td>4.945</td>
<td>4.608</td>
<td>3.044</td>
<td>0.725</td>
</tr>
<tr>
<td></td>
<td>G_C</td>
<td>5.584</td>
<td>5.473</td>
<td>4.516</td>
<td>3.052</td>
</tr>
<tr>
<td></td>
<td>G_C</td>
<td>6.009</td>
<td>5.743</td>
<td>4.751</td>
<td>3.132</td>
</tr>
<tr>
<td>Voice babble</td>
<td>GEM</td>
<td>3.433</td>
<td>2.342</td>
<td>1.595</td>
<td>0.640</td>
</tr>
<tr>
<td></td>
<td>GL</td>
<td>3.508</td>
<td>2.395</td>
<td>1.642</td>
<td>0.716</td>
</tr>
<tr>
<td></td>
<td>GW</td>
<td>4.179</td>
<td>2.931</td>
<td>2.860</td>
<td>1.465</td>
</tr>
<tr>
<td></td>
<td>GL-CSC</td>
<td>4.086</td>
<td>3.050</td>
<td>2.394</td>
<td>2.034</td>
</tr>
<tr>
<td></td>
<td>DGW</td>
<td>3.942</td>
<td>2.801</td>
<td>1.868</td>
<td>1.194</td>
</tr>
<tr>
<td></td>
<td>NBLG</td>
<td>4.066</td>
<td>2.892</td>
<td>1.962</td>
<td>1.506</td>
</tr>
<tr>
<td></td>
<td>G_C</td>
<td>4.330</td>
<td>3.292</td>
<td>2.787</td>
<td>2.090</td>
</tr>
<tr>
<td></td>
<td>G_C</td>
<td>4.558</td>
<td>3.399</td>
<td>2.878</td>
<td>2.236</td>
</tr>
<tr>
<td></td>
<td>G_C</td>
<td>4.516</td>
<td>3.360</td>
<td>2.863</td>
<td>2.229</td>
</tr>
</tbody>
</table>
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 \( G_N \) and \( G_C \) 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.

### Table 4

<table>
<thead>
<tr>
<th>Noise Method</th>
<th>0 dB</th>
<th>5 dB</th>
<th>10 dB</th>
<th>15 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-16 two seat</td>
<td>( G_{EM} )</td>
<td>0.096</td>
<td>0.086</td>
<td>0.063</td>
</tr>
<tr>
<td>( G_{L-FSA} )</td>
<td>0.083</td>
<td>0.091</td>
<td>0.067</td>
<td>0.049</td>
</tr>
<tr>
<td>( G_Y )</td>
<td>0.129</td>
<td>0.129</td>
<td>0.092</td>
<td>0.078</td>
</tr>
<tr>
<td>( G_{L-CSC} )</td>
<td>0.118</td>
<td>0.128</td>
<td>0.096</td>
<td>0.085</td>
</tr>
<tr>
<td>DGW</td>
<td>0.114</td>
<td>0.124</td>
<td>0.092</td>
<td>0.080</td>
</tr>
<tr>
<td>NBLG</td>
<td>0.121</td>
<td>0.126</td>
<td>0.093</td>
<td>0.079</td>
</tr>
<tr>
<td>( G_N )</td>
<td>0.139</td>
<td>0.131</td>
<td>0.101</td>
<td>0.086</td>
</tr>
<tr>
<td>( G_C )</td>
<td><strong>0.146</strong></td>
<td><strong>0.143</strong></td>
<td><strong>0.107</strong></td>
<td><strong>0.089</strong></td>
</tr>
<tr>
<td>( G_{L-FSA} )</td>
<td>0.083</td>
<td>0.091</td>
<td>0.067</td>
<td>0.049</td>
</tr>
<tr>
<td>( G_Y )</td>
<td>0.129</td>
<td>0.129</td>
<td>0.092</td>
<td>0.078</td>
</tr>
<tr>
<td>( G_{L-CSC} )</td>
<td>0.118</td>
<td>0.128</td>
<td>0.096</td>
<td>0.085</td>
</tr>
<tr>
<td>DGW</td>
<td>0.114</td>
<td>0.124</td>
<td>0.092</td>
<td>0.080</td>
</tr>
<tr>
<td>NBLG</td>
<td>0.121</td>
<td>0.126</td>
<td>0.093</td>
<td>0.079</td>
</tr>
<tr>
<td>( G_N )</td>
<td>0.139</td>
<td>0.131</td>
<td>0.101</td>
<td>0.086</td>
</tr>
<tr>
<td>( G_C )</td>
<td><strong>0.146</strong></td>
<td><strong>0.143</strong></td>
<td><strong>0.107</strong></td>
<td><strong>0.089</strong></td>
</tr>
<tr>
<td>( G_{L-FSA} )</td>
<td>0.083</td>
<td>0.091</td>
<td>0.067</td>
<td>0.049</td>
</tr>
<tr>
<td>( G_Y )</td>
<td>0.129</td>
<td>0.129</td>
<td>0.092</td>
<td>0.078</td>
</tr>
<tr>
<td>( G_{L-CSC} )</td>
<td>0.118</td>
<td>0.128</td>
<td>0.096</td>
<td>0.085</td>
</tr>
<tr>
<td>DGW</td>
<td>0.114</td>
<td>0.124</td>
<td>0.092</td>
<td>0.080</td>
</tr>
<tr>
<td>NBLG</td>
<td>0.121</td>
<td>0.126</td>
<td>0.093</td>
<td>0.079</td>
</tr>
<tr>
<td>( G_N )</td>
<td>0.139</td>
<td>0.131</td>
<td>0.101</td>
<td>0.086</td>
</tr>
<tr>
<td>( G_C )</td>
<td><strong>0.146</strong></td>
<td><strong>0.143</strong></td>
<td><strong>0.107</strong></td>
<td><strong>0.089</strong></td>
</tr>
</tbody>
</table>

(continued on next page)
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, $G_C$ 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-
igate how the proposed methods perform 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 detection 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. 14.](image1)

**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.](image2)

**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.
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 priori 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].

**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., $G_a$ [15], $G_c$ [22], and $G_G$ [27]) using blind a priori 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.
The results obtained from the objective (Fig. 16) 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 (\( G_{\text{GMM}} \%21.4, \ G_{\text{ADP}} \%21.4, \) and \( G_{\text{AUP}} \%44.6 \)) were much less preferred by the listeners than those enhanced by the proposed methods (\( G_c \%75, \ G_i \%73.2 \) and \( G_a \%64.3 \)). 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), \( \phi_{\text{dev, max}} \approx 0.679 \) 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.

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 \( G_M \) [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 \( G_c \) (Laplacian PDF) achieves a higher perceptual quality than \( G_{c, FSA} \) [19], for \( \gamma = 1 \), complex Laplacian PDF. Subjective listening tests reveal that the residual noise level in the \( G_N \) and \( G_c \) estimation is lower than that in the \( G_M \) and \( G_{c, FSA} \), 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 \( G_c \)), and the DFT-based approach, \( G_{M, D} \) (or \( G_{c, FSA} \)), 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 DCT 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, \( G_c \) or \( G_N \) clearly gives better enhanced speech quality than \( G_m \) and \( G_{M, D} \). Listening results also reveal that the performance of \( G_c \) and \( G_N \) are comparable. Among the algorithms based on the Laplacian prior, \( G_c \) gives similar or better noise attenuation but less speech distortions than \( G_{c, CSC} \) and \( G_{c, FSA} \). 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 priori 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 $\sigma_Y^2 = \sigma_S^2 + \sigma_N^2$. Upon substituting (8) and (14) into (13b) we obtain

$$\hat{X} = \frac{\sigma_Y}{2\pi \sigma_N \sigma_S} \int_0^\infty dx \exp \left[ -\frac{1}{2 \sigma_Y^2} \left( \frac{\sigma_N^2}{\sigma_N^2 + \sigma_S^2} x - \frac{\sigma_S^2}{\sigma_N^2 + \sigma_S^2} y \right)^2 \right] \exp \left[ -\frac{1}{2 \sigma_N^2} x^2 \right] \exp \left[ -\frac{1}{2 \sigma_S^2} y^2 \right]$$

By using [eq.2.33.1.23.36][38] we get from (B.1)

$$\hat{X} = \frac{\sqrt{2 \sigma_Y \sigma_N}}{\sqrt{2 \pi \sigma_N \sigma_Y}} \exp \left[ -\frac{\sigma_S^2}{2 \sigma_Y^2} x^2 \right] \exp \left[ -\frac{\sigma_N^2}{2 \sigma_Y^2} y^2 \right] \exp \left[ -\frac{1}{2 \sigma_N^2} \left( \frac{\sigma_N^2}{\sigma_N^2 + \sigma_S^2} x - \frac{\sigma_S^2}{\sigma_N^2 + \sigma_S^2} y \right)^2 \right]$$

The equivalent form of $\hat{X}$ 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.9.212.29.212.3][38] yields

$$\varphi_{\frac{1}{2}}(z) = 2^\frac{1}{4} \pi^{-\frac{1}{4}} \phi \left( \frac{1}{4} \frac{1}{2} \frac{1}{2} \right)$$

and

$$\varphi_{\frac{1}{2}}(z) = 2^\frac{1}{4} \pi^{-\frac{1}{4}} \phi \left( \frac{1}{4} \frac{1}{2} \frac{1}{2} \right)$$

After substituting (C.1) and (C.2) into (25), and using [eq.9.6.4.13.6.3][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

$$E\left[D_n^2|Y\right] = E\left[D_n^2|Y, H_0\right] p(H_0|Y) + E\left[D_n^2|Y, H_1\right] p(H_1|Y)$$

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, $\zeta_m$, which is optimal in terms of the total probability error. Under hypothesis $H_0$, $|Y| = |D|$, we can approximate the preceding estima-
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 \([37]\)

\[
P(|Y| | H_0) = \frac{2}{\sqrt{2\pi} \sigma^2_D} \exp \left[ -\frac{|Y|^2}{2\sigma^2_D} \right]
\]

Under hypothesis \(H_1, |Y| = |X + D|\), and since \(Y\) is a Gaussian variate with variance, \(\sigma^2_Y = \sigma^2_X + \sigma^2_D\), the PDF for \(P(|Y| | H_1)\) is

\[
P(|Y| | H_1) = \frac{2}{\sqrt{2\pi} \sigma^2_Y} \exp \left[ -\frac{|Y|^2}{2\sigma^2_Y} \right]
\]

where \(\xi\) is the a priori SNR. We can compute the optimal estimator for speech presence as

\[
E \{ |D|^2 | Y, H_1 \} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{|Y|^2}{\sqrt{2\pi} \sigma^2} \exp \left[ -\frac{(Y-D)^2}{2\sigma^2} \right] dy \, dx
\]

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

\[
p(Y | D) = \frac{1}{\sqrt{2\pi} \sigma_Y} \exp \left[ -\frac{(Y-D)^2}{2\sigma_Y^2} \right]
\]

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

\[
E \{ |D|^2 | Y, \hat{\xi}, H_1 \} = \frac{\hat{\xi}}{\xi + 1} \sigma^2_D + \frac{|Y|^2}{\xi + 1}
\]

Note that evaluation of \((D.6)\) requires the estimates \(\hat{\xi}\) and \(\hat{\sigma}_D\) 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., \(\hat{\sigma}^2_D = \hat{\sigma}^2_D(l-1)\), and the maximum-likelihood (ML) estimate of the a priori SNR, i.e.,

\[
\hat{\xi}_{ml} = \hat{\xi} - 1 = \frac{|Y|^2}{\sigma^2_Y(l-1)} - 1
\]

for \((D.6)\) and attain

\[
E \{ |D|^2 | Y, \hat{\xi}, H_1 \} = \hat{\sigma}^2_D(l-1)
\]

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

\[
E \{ |D|^2 | Y \} = p(H_0 | Y) |Y|^2 + p(H_1 | Y) \hat{\sigma}^2_D(l-1)
\]

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 \(\beta = 0.95\)

\[
\hat{\sigma}^2_D(l) = \beta \hat{\sigma}^2_D(l-1) + (1-\beta)p(H_1 | Y) |Y|^2
\]

\[\text{(D.10)}\]

\[\text{D.1. a posteriori SPP estimation}\]

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

\[
p(H_1 | Y) = \left\{ 1 + \frac{p(H_0)}{p(H_1)} \sqrt{\bar{\zeta}_h} + 1 \exp \left[ -\frac{|Y|^2}{2\sigma^2_Y(l-1)} \right] \right\}^{-1}
\]

\[\text{(D.11)}\]

which follows from substitution of \((35)\) into \((31)\). Where \(\bar{\zeta}_h\) 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 \(\bar{\zeta}_h\) for SPP estimation has been proposed the first time in \([58]\). In the next section we determine the optimal value of \(\bar{\zeta}_h\).

\[\text{D.2. Optimal } \bar{\zeta}_h \text{ estimation}\]

The optimal \(\bar{\zeta}_h\) is found by minimising the total probability of error \([59]\)

\[
P_d = p(H_0)P_f + p(H_1)P_m
\]

\[\text{(D.12)}\]

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

\[
P_f(\bar{\zeta}_h) = \int_{|y|}^{\infty} p(|y| | H_0) \, dy = \text{erfc} \left( \sqrt{\frac{\bar{\zeta}_h}{\xi + 1}} \right)
\]

\[\text{(D.13)}\]

\[
P_m(\bar{\zeta}_h, \xi) = \int_{0}^{\infty} p(|y| | H_1) \, dy = \text{erf} \left( \sqrt{\frac{\bar{\zeta}_h}{\xi + 1}} \right)
\]

\[\text{(D.14)}\]

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

\[
|y| = \left( \frac{\bar{\zeta}_h + 1}{\xi + 1} \right) \ln \left( \frac{\sigma^2_D}{\bar{\zeta}_h + 1} \right) \left\{ \frac{[p(H_0)]^2}{p(H_1) \bar{\zeta}_h} \right\}
\]

\[\text{(D.15)}\]
and \( \mu(\tilde{z}_{H1}) \) is defined by
\[
\mu(\tilde{z}_{H1}) = \frac{|y|^2}{2\sigma_y^2} = \frac{1}{2} \left( \frac{H_{1}^2 + 1}{\tilde{z}_{H1}} \right) \ln(\tilde{z}_{H1} + 1) \quad (D.16)
\]
The optimal \( \tilde{z}_{H1} \) is found by minimizing the total probability of error when \( \xi \) is uniformly distributed between \( \tilde{z}_{\text{low}} = 0 \) and \( \tilde{z}_{\text{up}} = 100 \).
\[
\tilde{z}_{H1} = \min_{\tilde{z}_{H1}} \int_{\tilde{z}_{\text{low}}}^{\tilde{z}_{\text{up}}} p(H_0)P_f(\tilde{z}_{H1}) + p(H_1)P_m(\tilde{z}_{H1}, \xi) d\xi
\]

Using [eq.4.1.4.1.2] [60], (D.17) can be expressed as shown in (D.18). The lower bound 10 log(\( \tilde{z}_{\text{low}} \)) = \(-\infty\) dB and upper bound 10 log(\( \tilde{z}_{\text{up}} \)) = 20 dB are considered appropriate for a noise reduction application [35]. The expression given in (D.18) is plotted against \( \tilde{z}_{H1} \) in Fig. 18. The minimum is at 10 log(\( \tilde{z}_{H1} \)) = 15.6 dB. This is the point of interest at which minimum \( \tilde{z}_{H1} \) is the criterion. Therefore, we use 10 log(\( \tilde{z}_{H1} \)) = 15.6 dB for (D.11), and throughout the algorithm.

\[
\tilde{z}_{H1} = \min_{\tilde{z}_{H1}} \left\{ \int_{\tilde{z}_{\text{low}}}^{\tilde{z}_{\text{up}}} p(H_0)P_f(\tilde{z}_{H1}) + p(H_1)P_m(\tilde{z}_{H1}, \xi) d\xi \right\}
\]

\[
\tilde{z}_{H1} = \min_{\tilde{z}_{H1}} \left\{ \int_{\tilde{z}_{\text{low}}}^{\tilde{z}_{\text{up}}} \left[ (2\mu(\tilde{z}_{H1}) + \frac{1}{2})P_m(\tilde{z}_{H1}, \xi) + \frac{1}{2} \mu(\tilde{z}_{H1}) \right] d\xi \right\}
\]

\[
\tilde{z}_{H1} = \min_{\tilde{z}_{H1}} \left\{ \int_{\tilde{z}_{\text{low}}}^{\tilde{z}_{\text{up}}} \left[ \mu(\tilde{z}_{H1}) + \frac{1}{2} \right] \right\}
\]

**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**


