

Received August 1, 2020, accepted September 6, 2020, date of publication September 10, 2020, date of current version September 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3023348

Performance Improvement of Adaptive Wavelet Thresholding for Speech Enhancement Using Generalized Gaussian Priors and Frame-Wise Context Modeling

PARNA CHAKRABORTY BHATTACHARYA¹,
NISACHON TANGSANGIUMVISAI¹ , (Senior Member, IEEE),
AND APISAK WORAPISHET², (Senior Member, IEEE)

¹Department of Electrical Engineering, Faculty of Engineering, Multimedia Data Analytics and Processing Research Unit, Chulalongkorn University, Bangkok 10330, Thailand

²Mahanakorn Institute of Innovation (MII), Mahanakorn University of Technology, Bangkok 10530, Thailand

Corresponding author: Nisachon Tangsangiumvisai (nisachon.t@chula.ac.th)


This work was supported in part by the 100th Anniversary Chulalongkorn University for Doctoral Scholarship, and in part by the 90th Anniversary of the Chulalongkorn University Fund (Ratchadaphiseksomphot Endowment Fund).

ABSTRACT This work aims at developing an adaptive wavelet thresholding algorithm for speech enhancement with significant performance improvement over other wavelet-based counterparts. This is accomplished through the formulation of the optimum threshold for noise reduction, based on the generalized Gaussian priors to fully characterize the statistics of speech and noise wavelet coefficients. In addition, through the frame-wise context modeling which enables tracking of the statistical characteristics of each individual coefficient on the frame-wise basis, the optimum threshold is accurate and adaptive at both the coefficient level and frame level. The frame-wise context model is formulated by virtue of the context subspace projection of the wavelet coefficients, with the context index employed as the invariant correspondence between successive frame parameters, thereby enabling the frame-wise tracking at the coefficient level. Simulation results show significant improvement over the wavelet-based speech enhancement algorithms in terms of the segmental signal-to-noise ratio improvement by as much as 226%, the perceptual evaluation of speech quality by 36%, the short-time objective intelligibility by 17.8% and the cepstral distance by 33.3%. When benchmarked with the well-established short-time-Fourier-transform-based counterparts, the proposed wavelet thresholding algorithm offers favorable and more robust performances, particularly under non-stationary noise conditions, with no adverse musical noise effect.

INDEX TERMS Context modeling, speech enhancement, wavelet thresholding.

I. INTRODUCTION

Besides its typical applications in teleconferencing, hands-free communications, hearing devices, etc., signal processing for speech enhancement (SE) has recently witnessed increasing demand in emerging applications. Much of this growth is attributed to the increased adoption of voice-control applications, such as voice-activated robots and in-vehicle voice navigation [1]–[3]. Moreover, Internet-of-Things (IoT) enabled applications, such as smart home appliances and connected machines, generally make use of voice commands [4]–[6].

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

Since these applications require automatic speech recognition, it is inevitable that environmental background noise is picked up and the quality of speech signals to be processed can be adversely affected. To this end, the SE is of the utmost importance in maintaining the integrity of these evolving applications.

Most of the SE algorithms developed so far are based on the short-time Fourier transform (STFT). Among these, the spectral subtraction (SS) has become one of the most popular algorithms due to its simplicity [7]. However, its major drawback is the associated high level of musical noise, which usually has a negative impact on speech quality. Various modified versions of the SS algorithms have been

proposed to reduce the musical noise, such as multi-band SS and iterative SS based on high-order statistics [8]–[12]. Another well-established STFT-based SE algorithms relies upon the minimum mean square error (MMSE) estimation of the speech spectrum [13]. Various MMSE estimators have been developed to obtain enhanced speech from noisy speech signals using different approximations of the complex-valued spectrum [14]–[19]. As compared to their SS counterparts, the MMSE estimators have been demonstrated to provide a better performance trade-off between the noise attenuation and musical noise effect, thereby yielding better enhanced speech quality. Further improvement on the trade-off can be obtained by incorporating uncertainty in speech presence [17], [25] and a perceptual speech model [26].

Following its tremendous success for image denoising [27]–[30], the wavelet thresholding has made inroads into speech enhancement [31]–[41], which is hereafter also referred to as speech denoising. This was mainly motivated by advantages offered by the wavelet transform (WT) over the STFT. Since the WT offers time-frequency *multi*-resolution processing, it is more suitable to handling non-stationary signals inherent in the characteristics of speech and its environmental noise. In addition, the WT makes no use of windowing, which inevitably entails bias-variance trade-off in spectral estimation, causing possible generation of musical noise as experienced in the STFT domain [39]. Another benefit of the WT is its simplicity in processing real values typically associated with speech wavelet coefficients instead of complex values in the STFT domain.

Thus far, however, the level of speech enhancing performances in the WT domain has been inferior to its STFT counterparts. The main bottleneck lies in the difference in the underlying characteristics between image and speech signals. Since typical image wavelet coefficients exhibit a near-sparse condition in each subband, image denoising by thresholding the noise coefficients using the conventional hard/soft threshold function is effective [27], [28]. On the contrary, speech wavelet coefficients exhibit significant deviation from the near-sparse condition, making them inseparable from the noise coefficients in a noisy speech environment. As a consequence, not only the wavelet thresholding algorithms are less effective for speech denoising, but the enhanced speech signals are also subject to high distortion.

To enable more distinctive separation between speech and noise wavelet coefficients in each subband, various techniques have been proposed. Instead of using the octave-band WT, the wavelet packet transform (WPT) was employed for fine-resolution uniform subband decomposition in [31]–[34], and the perceptual wavelet packet (PWP) for non-uniform subbands based on models of human auditory speech perception in [35]–[38]. To provide improved extraction of speech from noise, the Teager energy (TE) operator was also utilized in [31], [37], and [38]. With the use of a two microphone system, the blind source separation (BSS) technique was employed to separate the speech and noise wavelet

coefficients in [40], [41]. To reduce speech distortion, various custom threshold functions with continuous derivative characteristic were proposed in [36]–[41] to replace the hard/soft threshold function.

In addition, various improved optimum thresholding methods that incorporate characteristics of speech and/or noise coefficients at both frame and subband levels have been developed. In [32], the segmental signal-to-noise ratios (SNR) of the wavelet coefficients were included for the threshold calculation. In [33], the band recursive threshold, based on a weighted sum of noise wavelet coefficients from other subbands, was presented. In [34], the iterative Kalman filter was applied to the thresholding method, assuming Gaussian noise. It is not until recently that the prior probability distributions of *both* speech and noise have been incorporated, and significant performance improvement was achieved. In [36], the threshold formulation was based on the symmetric Kullback Leibler divergence and a Gaussian distribution for both speech and noise wavelet coefficients. In [37] and [38], the student's *t*-distribution and the Rayleigh distribution were employed, respectively, to model the TE operated speech and noise coefficients.

In this work, we propose an adaptive wavelet thresholding algorithm with the generalized Gaussian (GG) priors and frame-wise context modeling for general purpose speech enhancement, with emphasis on emerging voice-control applications. The optimum threshold is made adaptive at both the coefficient and frame levels by virtue of modeling each individual wavelet coefficient as a GG random variable with its standard deviation estimated and updated on a frame-wise basis. Summarized below are the technical contributions of the proposed algorithm and its main advantages over other wavelet thresholding counterparts.

- Instead of employing the Gaussian priors in [36] and [42], the student's *t*-distribution prior in [37] and the Rayleigh prior in [38] for the derivation of the optimum threshold, the use of the GG priors in the proposed algorithm fully represents the statistical characteristics of speech and noise wavelet coefficients, and hence results in more accurate optimum threshold, particularly over various non-stationary noise conditions.
- The direct statistical modeling of the wavelet coefficients in the proposed algorithm is more accurate than the indirect modeling of the TE operated PWP coefficients in [37], [38], where only instantaneous energy characteristics of wavelet coefficients are retained after the TE operation.
- Through the frame-wise context modeling, the proposed algorithm offers a different optimum threshold value for each individual wavelet coefficient in each subband, instead of using the same threshold value for all the wavelet coefficients in each subband, as in the WPT algorithms in [31]–[34], and the PWP algorithms in [35]–[38].
- The frame-wise context modeling also offers statistical estimation and update of each individual noise

coefficient on a frame-to-frame basis. This is in contrast to the WT-based algorithms with statistical priors in [36]–[38], where the noise statistics were estimated and updated at the subband level using the methods developed elsewhere, i.e. the noise energy method in [32], and the improved minima controlled recursive averaging (IMCRA) method in [43]. Note that, for the conventional context modeling in [42], there was no noise update, and the estimation was directly made at the highest subband, and used across all the subbands.

- The coefficient-level thresholding results in a small number of subbands as typically required in image denoising, e.g. four octave bands (see Fig.3). This is unlike the subband-level thresholding in other WT-based algorithms [31]–[38], which invariably requires more than twice the number of subbands as compared to the proposed algorithm, so as to achieve acceptable speech enhancement performance.
- The coefficient-level thresholding enables the use of a simple wavelet type (the Daubechies family in this work), and the simple soft-threshold function with no adverse effect on speech quality. This is in contrast to the need for special psychoacoustic or perceptual wavelets in [35]–[38] and more sophisticated custom threshold functions by the subband-level thresholding algorithms in [32], [33], [35], [37], and [38] to avoid excessive distortion in the enhanced speech.
- Unlike its BSS counterparts in [40], [41], the proposed adaptive wavelet thresholding algorithm requires only one single microphone.

This paper is organized as follows. In Section II, the operational overview of the proposed wavelet thresholding SE algorithm is given. In Section III, the GG distribution model for typical speech and noise wavelet coefficients is characterized, followed by the formulation under the GG priors to find the optimum threshold. In Section IV, the frame-wise context modeling using a subspace formulation for the frame-to-frame estimation and update of the statistical parameters is detailed. The implementation, simulation, and comparative performance evaluation with the recent state-of-the-art WT-based SE algorithms are provided in Section V. In addition, a benchmark with some well-established STFT-based SE algorithms is also included. It is demonstrated that the proposed adaptive wavelet thresholding offers superior performances than the WT-based counterparts, and exhibits more favorable and robust performances as compared to the STFT-based algorithms, without the musical noise effect. Finally, the conclusion and prospects are given in Section VI.

II. OVERVIEW OF PROPOSED ADAPTIVE WAVELET THRESHOLDING

To describe the operational overview, the block diagram of the proposed adaptive wavelet thresholding algorithm for SE is shown in Fig. 1. From the diagram, the noisy speech input signal is divided into frames, and each frame is

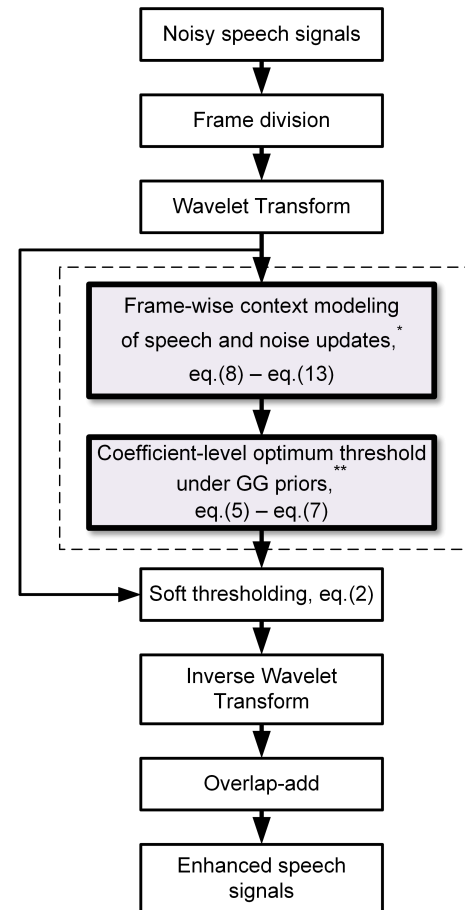


FIGURE 1. Block Diagram of the proposed wavelet thresholding algorithm with frame-wise context modeling for speech enhancement. (*See Section IV, **See Section III).

transformed to the wavelet domain using a regular WT with octave bands. This is followed by the frame-wise context modeling where the statistical parameters of each individual speech and noise wavelet coefficients, including the shape parameter and standard deviation under the GG model, are estimated and updated on a frame-wise basis. Subsequently, the optimum coefficient-level threshold value, formulated based on the Bayesian MMSE under the GG prior, is determined. Each noisy speech coefficient is thresholded accordingly using the soft-thresholding function to obtain its clean speech estimate. The inverse WT is then applied, and the denoised speech signal is reconstructed through the overlap-add method.

In contrast to image denoising in [30], where an approximation using the Gaussian distribution for image and noise coefficients was shown to be sufficient, the incorporation of the GG distribution to the optimum threshold formulation in this algorithm makes it highly compatible with speech and its noise environments. As a result, consistent speech enhancement performance under various noisy conditions can be achieved. Through numerical computation, an empirical near-optimal threshold as a function of the shape

parameter and standard deviation of the GG model is developed, as will be described in Section III.

The frame-wise context modeling developed in this work is the extension of the context modeling for image denoising in [42] to speech enhancement which requires a frame-based processing. Unlike image denoising, the frame-wise basis of speech processing necessitates recursive updating and transferring operation of parameters between frames. While the operation can be executed at the subband level in the WT domain, it is ill-defined at the individual wavelet coefficient level, because there is no unique correspondence of the wavelet coefficients between successive frames, due to their inherent temporal characteristic. Based upon the operational insight of the context modeling as low-dimensional subspace projection and clustering, it is thus proposed in this work to project the current-frame temporal coefficients into a subspace formed by the previous-frame counterparts, before subsequent pairing between the context parameters of the projected coefficients with the same context index. This is henceforth denoted as the *frame-wise context modeling*. Its operational details will be described in Section IV.

III. OPTIMUM THRESHOLD FORMULATION FOR SPEECH ENHANCEMENT

Let the noisy speech signal, $y(t)$, be modeled in the time domain by $y(t) = s(t) + n(t)$, where $s(t)$ denotes the clean speech and $n(t)$ denotes the additive noise. Consider a frame-wise processing, where the signals are divided into a sequence of windowed time-domain frames, with each frame represented by their corresponding $\tilde{M} \times 1$ vectors $\mathbf{y}(l)$, $\mathbf{s}(l)$ and $\mathbf{n}(l)$, where \tilde{M} is the number of time-domain samples per frame or frame length, and l denotes the frame index.

The WT is typically implemented as a critically-sampled octave-band filter bank. This essentially groups the wavelet coefficients into low/high subbands of different scales, and each subband is related to the frequency band of the signals. The wavelet coefficients in the high-frequency subband, H_i , with $i = 1, 2, \dots, I$ is called the details and those in the low-frequency subband, L_i , is called the approximations, where i is the scale, and I is the largest number of scales. A subband at scale i has size $M = \tilde{M}/2^i$.

For each frame, the wavelet transform of the noisy speech model is expressed by

$$\mathbf{Y}^i(l) = \mathbf{S}^i(l) + \mathbf{N}^i(l) \quad (1)$$

where $\mathbf{Y}^i(l) = W^i \mathbf{y}(l)$, $\mathbf{S}^i(l) = W^i \mathbf{s}(l)$, and $\mathbf{N}^i(l) = W^i \mathbf{n}(l)$ denote the wavelet coefficients vectors of $\mathbf{y}(l)$, $\mathbf{s}(l)$, and $\mathbf{n}(l)$, respectively, and W^i is the one-dimensional orthogonal WT operator at scale i . For notational convenience, the scale, i , and the frame index, l , will not be explicitly included, unless otherwise necessary for clarity.

Speech denoising using the wavelet thresholding method is accomplished by applying the noisy speech wavelet coefficients, $Y(j)$, in each subband with a threshold function, where $j = 1, 2, \dots, M$ is the coefficient index. In speech denoising,

the soft-thresholding function

$$\mathcal{T}(Y, T) = \text{sgn}(Y) \cdot \max(|Y| - T, 0) \quad (2)$$

is typically employed, where T denotes the threshold value, and $\text{sgn}(\cdot)$ denotes the sign of the wavelet coefficient, Y . The function in (2) essentially shrinks its argument Y towards zero by the threshold value, T , and hence is also called the shrinkage function. The soft-thresholding offers small discontinuity and less abrupt artifacts in the thresholded coefficients as compared to the hard-thresholding, as well as less computational complexity as compared to other non-linear thresholding functions [32], [38]. Note that, the thresholded wavelet coefficients are subsequently transformed by $W^{-1} \mathcal{T}(\mathbf{Y}, T)$, where W^{-1} is the inverse WT operator, to reconstruct the time-domain denoised signal.

A. PROBABILITY DISTRIBUTIONS OF SPEECH/NOISE COEFFICIENTS

In order to achieve effective noise reduction based on a statistical estimation framework, the probability distribution function (pdf) associated with clean speech and noise signals must be accurately modeled. In the STFT domain, a Gaussian or normal distribution is typically assumed for the real and imaginary discrete Fourier transform (DFT) coefficients of the signals. However, it has been shown that, for a short-frame period, a non-Gaussian distribution, such as a Laplacian or Gamma distribution, provides a better fit [20]. These non-Gaussian models are in fact special cases of a GG distribution model, where the GG pdf of a random variable X with a zero mean is given by

$$p(X) = GG(\sigma, \beta) = \frac{1}{2\Gamma(1 + \frac{1}{\beta})\mathcal{A}(\sigma, \beta)} \exp\left(-\left|\frac{X}{\mathcal{A}(\sigma, \beta)}\right|^\beta\right) \quad (3)$$

where Γ denotes the gamma function, the shape parameter, $\beta > 0$, is the measure of a peakness of the distribution, $\mathcal{A} = [\sigma^2\Gamma(1/\beta)/\Gamma(3/\beta)]^{1/2}$ is a scaling factor with $\sigma > 0$ being the standard deviation. Note that, (3) becomes a Laplacian distribution at $\beta = 1$, and a Gaussian or normal distributions at $\beta = 2$.

Since the noise attenuation by thresholding of noisy speech wavelet coefficients relies upon statistical estimation, a suitable probability distribution model for the wavelet coefficients is essential and must therefore be characterized. Fig. 2 shows histogram plots of the wavelet coefficients at subbands L_3, H_3, H_2, H_1 for full-length clean speech and raw noise data selected from the NOIZEUS database [44]. Note that, all the histograms have the standard deviation normalized to $\sigma = 1$ for ease of comparison. Also shown in each of the plots is the GG distribution curve (solid lines) with its shape parameter, β , estimated to fit the histogram by using the maximum likelihood estimation (MLE) algorithm available in MATLAB [45]. To show that the GG model can better represent the statistical distribution of the wavelet coefficients than the models employed by other algorithms, the fitted

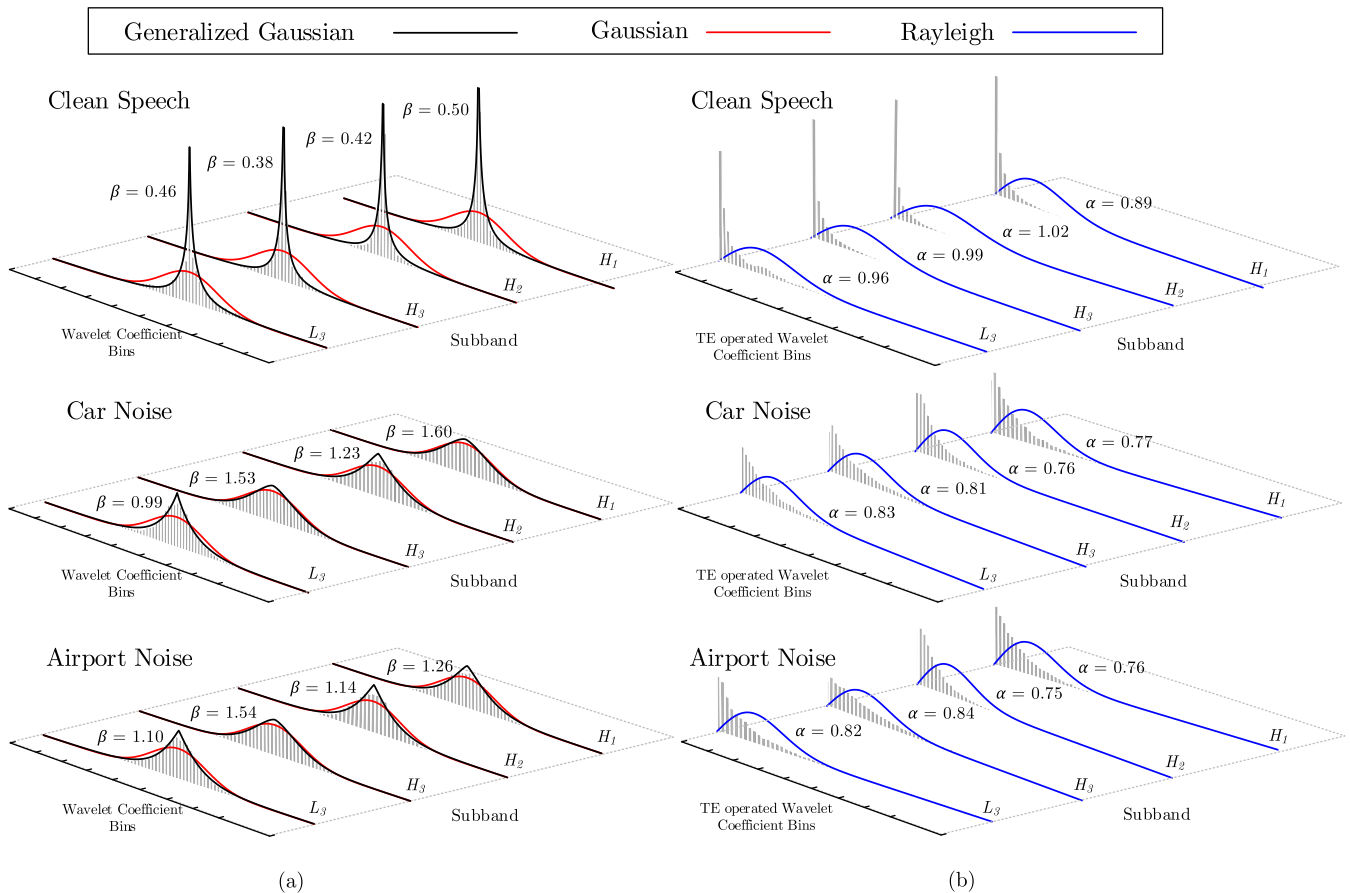


FIGURE 2. Histograms and fitted distribution curves at subbands L_3, H_3, H_2, H_1 for speech and noise signals selected from the NOIZEUS database. (a) Histograms and fitted curves for wavelet coefficients of clean speech, car noise, and airport noise using GG model (black lines) and Gaussian model (red lines), (b) histograms and fitted curves for TE operated wavelet coefficients of the same signals using Rayleigh model (blue lines).

distribution curve using the Gaussian model (red lines), as employed in [36] and [42]), are included in Fig. 2(a) for comparison. In Fig. 2(b), the distribution curves using the Rayleigh model in [38] fitted to the histograms of the TE operated wavelet coefficients of the same speech and noise signals of Fig. 2(a) are given. As clearly observed in the figures, unlike the other distribution models, the GG model can accurately capture the statistical distribution of the wavelet coefficients across various conditions. Following this, the GG distribution will be employed as the prior for speech denoising using wavelet thresholding in this work.

B. OPTIMUM THRESHOLD UNDER GG PRIORS

Fig. 3 shows the variation of the shape parameters of the wavelet coefficients at all subbands with the largest number of scales, $I = 3$, for a wide range of signals from the NOIZEUS database [44], including the shape parameters of clean speech signals, β_S , in Fig. 3(a), the shape parameters of various noise signals, β_N , i.e. a white Gaussian noise (WGN) in Fig. 3(b), babble noise in Fig. 3(c), car noise in Fig. 3(d), and airport noise in Fig. 3(e). Note that, for the noisy speech corpus NOIZEUS, there are 30 IEEE speech sentences produced and

corrupted by different types of real-world noise signals from the AURORA database [46] at different SNRs.

By inspecting Fig. 3, it is noticed that whereas the shape parameters of clean speech signals, β_S , at each subband only vary slightly and stay less than 0.8, the shape parameters of the noise signals, β_N , exhibit more significant variations with β_N ranging from 1.0 to about 2.0 depending on the type of noise. It is interesting to note that although the GG model was also employed as the prior for the image denoising in [30], it was only applied to the clean image, while a Gaussian distribution is assumed for the noise. Evidently, this is not suitable in speech denoising where the distribution of both the speech and noise wavelet coefficients tend to be of a super-Gaussian characteristic.

Under the GG distribution as the prior, the wavelet thresholding algorithm using the shrinkage function for speech enhancement can be formulated based on the Bayesian statistical framework. It is assumed that the wavelet coefficients are modeled as independent samples of the GG distribution at each subband, where $S \sim GG(\sigma_S, \beta_S)$ for the clean speech wavelet coefficients with σ_S being their standard deviation, and $N \sim GG(\sigma_N, \beta_N)$ for the noise wavelet coefficients with σ_N being their standard deviation. Note that, the fol-

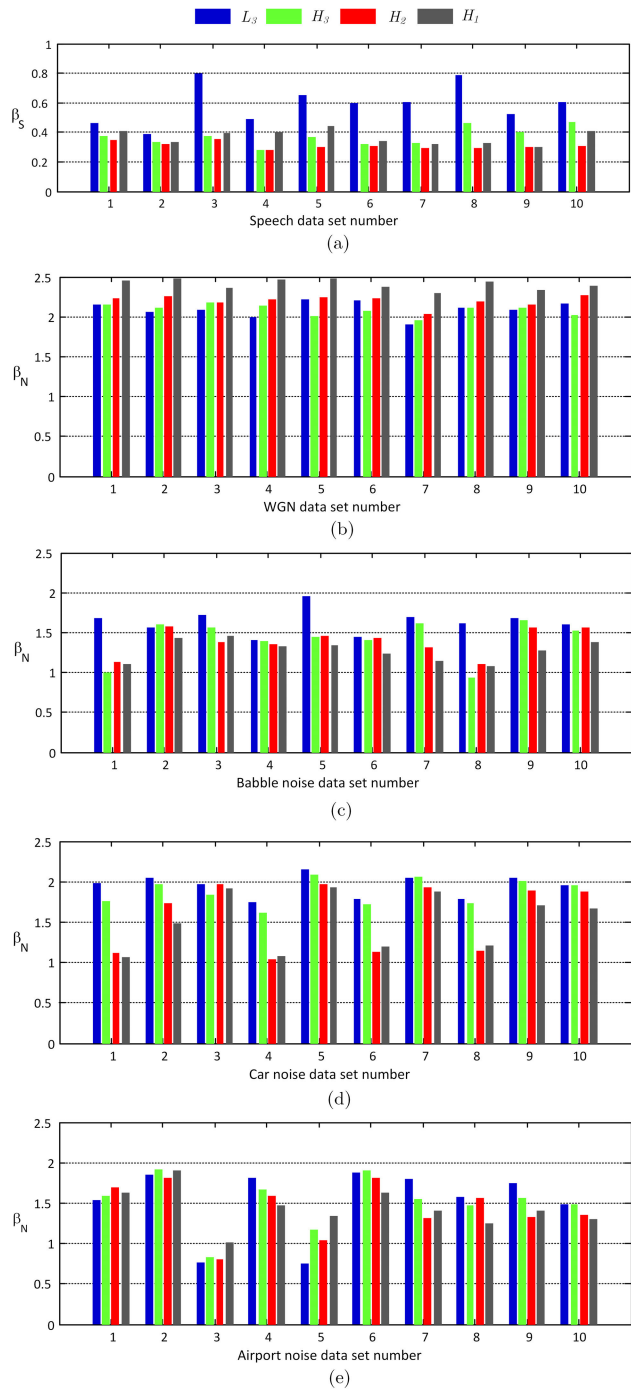


FIGURE 3. Variation of the shape parameters, β , of the GG models for the wavelet coefficients of speech and noise signals at all subbands (L_3, H_3, H_2, H_1) with the largest number of scales $l = 3$. (a) β_S of clean speech data set, (b) β_N of WGN, (c) β_N of babble noise, (d) β_N of car noise, and (e) β_N of airport noise. (10 data sets for speech and noises were chosen from the NOIZEUS database.)

lowing derivation is also applicable to the case where each wavelet coefficient at a subband is individually modeled as a random variable in Section IV. Let the estimator of the clean speech wavelet coefficients, \hat{S} , be the shrinkage function with the argument being the noisy wavelet coefficients, Y ,

as given by (2), i.e., $\hat{S} = \mathcal{T}(Y, T)$. Based upon the Bayesian MMSE approach, the optimum threshold, T_B , is defined as the value that minimizes the expected square error between the coefficients of the estimated and actual clean speech with respect to the joint pdf, $p(Y|S)$. Thus, we have

$$\begin{aligned} \mathbb{E}\{\hat{S} - S\}^2 &= \mathbb{E}_S \mathbb{E}_{Y|S} \{\mathcal{T}(Y, T) - S\}^2 \\ &= \iint_{-\infty}^{\infty} (\mathcal{T}(Y, T) - S)^2 p(Y|S) p(S) dY dS \end{aligned} \quad (4a)$$

with

$$T_B = \arg \min_T \mathbb{E}_S \mathbb{E}_{Y|S} \{\mathcal{T}(Y, T) - S\}^2 \quad (4b)$$

where the conditional coefficients are $Y|S \sim GG(\sigma_N, \beta_N)$, and $\mathbb{E}\{\cdot\}$ denotes the expectation operator. It is known that, under the GG priors, the minimization given in (4a) and (4b) has no closed-form solution for T_B and one must resort to numerical computation to find the optimum threshold.

Given the values of the shape parameters as summarized in Fig. 3, it is assumed that β_S is constant at 0.5 and the range of β_N is limited from 1.0 to 2.0, for simplification of the empirical optimum threshold expression. The empirical expression was assumed to be a product between σ_N and the fractional polynomial expansion of the ratio σ_S/σ_N , with a minimum possible number of terms that yield acceptable fitting. After a few iterations to determine the forms of the exponents and coefficients of the fractional polynomial, we arrive at the following closed-form approximation of T_B with a good fit to the numerical computation using (4a) and (4b),

$$T_B = \sigma_N \left[\frac{\frac{K_0}{\beta_S}}{\left(\frac{\sigma_S}{\sigma_N}\right)^{K_1 \left(\frac{\beta_S}{\beta_N} - \frac{1}{6}\right)}} + \frac{\left(1 - \frac{K_0}{3\beta_S}\right)}{\left(\frac{\beta_S}{\beta_N} + \frac{\sigma_S}{\sigma_N}\right)^{K_2 \left(\frac{\beta_S}{\beta_N}\right)}} + K_3 \right]. \quad (5)$$

The approximated optimum threshold for SE in (5) is dependent on both the clean speech and noise standard deviations, σ_S and σ_N , as well as the ratios σ_S/σ_N and β_S/β_N . Note that, the squared standard deviation ratio, σ_S^2/σ_N^2 , is equivalent to the *a priori* SNR of the DFT coefficients defined in the frequency-domain speech processing. As compared to the simple approximated threshold based on the Gaussian priors for image and noise signals in [30], the empirical closed-form threshold in (5) is more involved because both clean speech and noise wavelet coefficients are modeled by the GG distribution.

For validation, Fig. 4 shows comparative plots of the normalized optimum threshold, T_B/σ_N , against σ_S/σ_N , using the numerical computation in (4a), (4b), and the approximation in (5), at a fixed $\beta_S = 0.5$, and different β_N values from 1.0 up to 2.0. The constants in (5) at $K_0 = 0.5, K_1 = 5, K_2 = 6$ and $K_3 = 0.3$ show good agreement to fit the numerical curves. The closed-form equation offers the approximated threshold values within a small deviation from the MMSE obtained via numerical calculation over the given ranges of the shape parameters. From the plots in Fig. 4, it can be deduced that

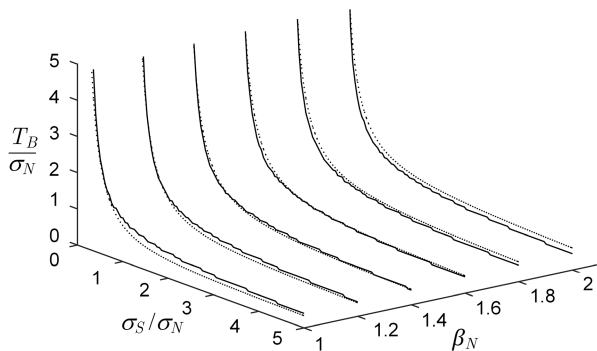


FIGURE 4. Numerical (dotted \dots) and approximated (solid $-$) normalized optimum threshold T_B/σ_N against σ_S/σ_N at $\beta_S = 0.5$ and different values of β_N .

T_B/σ_N generally increases as σ_S/σ_N decreases. In addition, the rate of increase is higher at a smaller σ_S/σ_N .

Further insight can be gained by plotting T_B/σ_N against β_N at different σ_S/σ_N ratios, as shown in Fig. 5. When $\sigma_S/\sigma_N \geq 1$, T_B/σ_N stays practically independent on β_N . On the contrary, as σ_S/σ_N gets smaller, T_B/σ_N starts to increase sharply, particularly at a small $\beta_N < 1$, where the noise distribution becomes more super-Gaussian. This is attributed to more occurrence of small noise amplitudes which has more impact on speech degradation when the σ_S/σ_N ratio is small, thereby resulting in a higher T_B . Such a high adaptivity of T_B at a low σ_S/σ_N ratio indicates the significance of incorporating the non-Gaussian distribution model in the noise statistics for speech denoising.

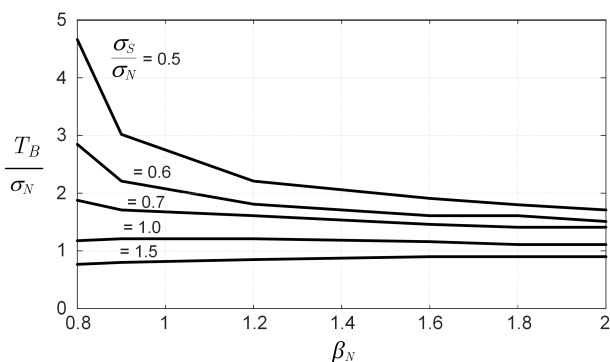


FIGURE 5. Plot of normalized optimum threshold T_B/σ_N against β_N at different values of σ_S/σ_N to investigate dependence of T_B/σ_N on β_N and σ_S/σ_N .

The calculation of the optimum threshold, T_B , under the GG priors requires the estimation of the shape parameter, β_N , and the standard deviations, σ_S and σ_N . The shape parameter estimation is carried out at a subband level since it normally exhibits small variation over each subband. The estimate of the shape parameter, $\hat{\beta}_N$, can be determined at each subband level by using the following equations derived by the method of moment in [47] to avoid using the MLE algorithm in

MATLAB which entails high computational complexity, i.e.,

$$\kappa(Y) = \frac{\left[\frac{1}{M} \sum_{j=1}^M |Y(j)| \right]^2}{\frac{1}{M} \sum_{j=1}^M |Y(j)|^2} \quad (6)$$

$$\hat{\beta}_N = \frac{1}{2a_1} \left(-a_1 + \sqrt{a_2^2 - 4a_1a_3 + 4a_1\kappa(Y)} \right), \quad \kappa(Y) \in [0.131, 0.449] \quad (7)$$

$a_1 = -0.536$, $a_2 = 1.169$, $a_3 = -0.152$. Note that, with reference to [47], the parameter values are truncated to three significant digits, and only the equations for the shape parameter, $\beta_N \in [0.277, 2.632]$, are given here to cover the typical range of speech wavelet coefficients summarised in Fig. 3. If the estimate, $\hat{\beta}_N$, is outside this range, it will be limited to the corresponding min/max range values.

In [30], the estimates of σ_S and σ_N were also accomplished at a subband level, where all the wavelet coefficients in the subband were employed to determine both σ_S and σ_N , and thus they all share the same T_B . In particular, σ_N was only obtained at H_1 and its value was used at all subbands. However, such a simple estimate will invariably result in poor quality in the denoised speech because of inherent non-sparsity in speech wavelet coefficients. A means to estimate and update σ_S and σ_N at an individual coefficient level in each subband on a frame-wise basis is the subject of the next section.

IV. FRAME-WISE CONTEXT MODELING

In the frame-wise context model, the noisy speech wavelet coefficients, \mathbf{Y} , in (1) at each subband are modeled as a temporal mixture of GG random variables with different standard deviations. As a consequence, a different optimum threshold, $T_B(j)$, for each individual $Y(j)$ can be determined for $j = 1, 2, \dots, M$, thereby yielding significant improvement in speech denoising performance without the need for fine subband resolutions as employed in most of the reported WT-based algorithms for SE [38], [39].

The frame-wise context model for speech denoising computes and updates the threshold value on a frame-by-frame basis. In order to enable recursive updates of the model over successive frames, it is proposed in this work an interpretation of the context modeling in view of subspace projection and clustering. Specifically, the context model is essentially a low-dimensional subspace representation of the wavelet coefficients. The main assumption is that high-dimensional data can be better clustered by exploiting a certain similarity measure in the projected lower-dimensional subspace, which adequately describes the data. Such an interpretation is useful for both in-frame and successive-frame correspondences between the wavelet coefficients and their contextual parameters, as will be later described later in Section IV-C.

The block diagram showing the operation of the frame-wise context modeling is given in Fig. 6. The wavelet coefficients are first projected onto the context subspace. This is followed by a formation of clusters, or context clustering,

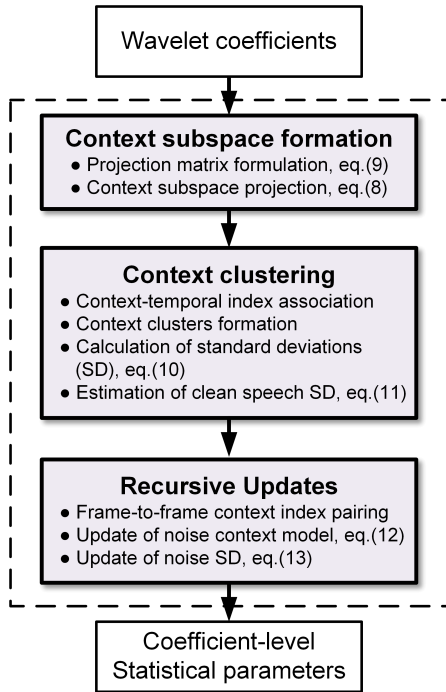


FIGURE 6. Block diagram of the proposed frame-wise context modeling.

in the context subspace by grouping the coefficients whose associated context values are close within a specified range. Note that, these clustered coefficients are not necessarily adjacent in the temporal WT domain. With respect to the cluster in the context subspace, the shape parameter, β , and standard deviation, σ , are estimated under the GG model assumption. By pairing between the same index in the projected context subspace, the recursive update of the statistical parameters within the same frame and between successive frames can be accomplished.

A. CONTEXT SUBSPACE PROJECTION

The subspace projection of the frame-wise context model is formulated by expressing the context of each wavelet coefficient, $Y(j)$, as a weighted average of the absolute values of its neighbouring coefficients, with the least-square (LS) error minimization being the objective function. Note that, the absolute value is employed because more correlation, and hence more statistical information from the otherwise uncorrelated nearby coefficients of the orthogonal wavelets, can be acquired [42]. As theoretically shown in [48], unlike other minimization criteria such as sparseness and low rank, the LS method takes the correlation structure of data into account. As a result, highly correlated data tend to be grouped together in the low-dimensional LS projected subspace, yielding better clustering, and thus more accurate estimation of the associated GG model parameters of each wavelet coefficient.

Under the LS subspace formulation, the context modeling can be described as follows. At a given subband with M wavelet coefficients, the context value, $Z(j)$, of the

coefficient $Y(j)$, which is essentially the projection of $|Y(j)|$ onto the LS subspace, is given by

$$\mathbf{Z} = \mathbf{P}|\mathbf{Y}| \tag{8}$$

with

$$\mathbf{P} = \mathbf{U}(\mathbf{U}^T \mathbf{U})^{-1} \mathbf{U}^T + \lambda \mathbf{I} \tag{9}$$

where \mathbf{Y} is the $M \times 1$ vector containing all the $Y(j)$ coefficients, \mathbf{Z} is an $M \times 1$ vector containing all the associated context values $Z(j)$ of $Y(j)$, \mathbf{P} is the $M \times M$ projection matrix analytically derived from the $M \times p$ matrix \mathbf{U} , \mathbf{I} is an $M \times M$ identity matrix. The operator $|\cdot|$ denotes the absolute value of each element in a vector. The regularization parameter, λ , helps control over-fitting in the LS method [48]. The matrix \mathbf{U} has each of its rows formed by the absolute values of p neighbouring coefficients of $Y(j)$, including its parent coefficient at lower subband, with the choice of p appropriately selected to capture to local contextual standard deviation of $Y(j)$ [42]. Note that, in (8) and (9), the matrix \mathbf{P} essentially projects $|\mathbf{Y}|$ onto the column space of the matrix \mathbf{U} , yielding the context vector \mathbf{Z} .

B. CONTEXT CLUSTERING

The clusters in the projected context subspace, \mathbb{Q} , are formed by grouping the coefficients, $Y(j)$, that have close context values, $Z(j)$. Following this, the context values $Z(j)$ of $Y(j)$ are sorted in ascending orders, and subsequently given indices $k = 1, 2, \dots, M$, to yield $Z(k)$ of $Y(k)$, where k denotes the context index. This temporal-context index association, or $j - k$ index association, between $Z(j)$ and $Z(k)$ leads to the rearrangement of $Y(j)$ to $Y(k)$, which enables classification of the wavelet coefficients into the same local cluster by virtue of some nearest context indices, k , as the similarity measure.

The noisy speech projection matrix, \mathbf{P}_Y , projects $|\mathbf{Y}|$ onto a noisy speech subspace, \mathbb{Q}_Y , and the noise projection matrix, \mathbf{P}_N , projects $|\mathbf{N}|$ onto a noise subspace, \mathbb{Q}_N . Although, by using (9), the projection matrix, \mathbf{P}_Y , can be determined during a speech-activity (SA) frame when $\mathbf{Y} = \mathbf{S} + \mathbf{N}$, and the projection matrix, \mathbf{P}_N , during a non-speech-activity (NSA) frame (or noise-only frame) when $\mathbf{Y} = \mathbf{N}$, the definitions of \mathbf{P}_Y and \mathbf{P}_N are general irrespective of the frame type.

Fig. 7 shows the diagram illustrating the projection and association among the wavelet coefficient, $Y(j)$, in the temporal domain, $Y(k_Y)$ in the noisy speech context subspace, \mathbb{Q}_Y , and $Y(k_N)$ in the noise context subspace, \mathbb{Q}_N . Through the projection matrix \mathbf{P}_Y , $Y(j)$ is rearranged as $Y(k_Y)$, forming the $j - k_Y$ index association. Similarly, through the projection matrix \mathbf{P}_N , $Y(j)$ is rearranged as $Y(k_N)$, forming the $j - k_N$ index association. In the context subspace, the standard deviations, σ_Y associated with $Y(k_Y)$, and σ_N associated with $Y(k_N)$, can then be calculated via the context clustering.

Fig. 8 helps illustrate the rearrangement, clustering, and $j - k_{Y,N}$ index association through the subspace projection. The plots in the figure show the rearrangement of $Y(j)$ from

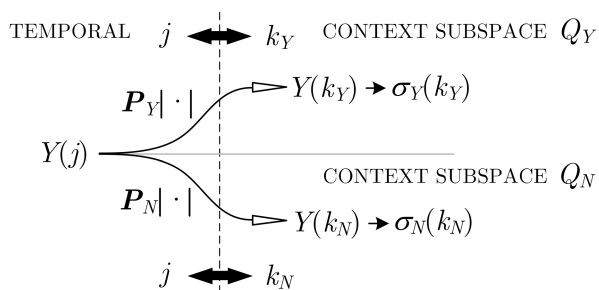


FIGURE 7. Diagram illustrates the rearrangement and $j - k$ association between $Y(j)$ in the temporal domain and $Y(k_{Y,N})$ in the context subspace through the projection matrices, P_Y (during an SA frame) and P_N (during an NSA frame). By using the context clustering, $\sigma_{Y,N}$ associated with $Y(k_{Y,N})$ can then be calculated.

an SA frame as $Y(k_Y)$ in a noisy speech context subspace in Fig. 8(a), and of $Y(j)$ from an NSA frame as $Y(k_N)$ in a noise context subspace in Fig. 8(b). Note that, the speech samples of these plots are corrupted by additive WGN at 5-dB SNR, with $M = 256$ and the number of neighbouring coefficients is $p = 3$ at subband H_1 . By examining the coefficient samples, a, b, c, d from an SA frame in Fig. 8(a), and e, f, g, h from an NSA frame in Fig. 8(b), it is seen that although they are far apart under the temporal index, j , they can become adjacent and hence are within the same local cluster under the context index, k_Y and k_N . This indicates the correlation between the coefficient samples in the corresponding context subspace, as determined by the LS subspace formulation. The plots also illustrate the index association between the coefficient samples. With the samples, a, b, c and d from an SA frame in Fig. 8(a), we have $Y(j)$ at $j = 29, 70, 129$, and 207 , associated with $Y(k_Y)$ at $k_Y = 218, 208, 217$, and 211 , respectively. With the samples, e, f, g and h from an NSA frame in Fig. 8(b), we have $Y(j)$ at $j = 25, 77, 126$, and 211 , associated with $Y(k_N)$ at $k_N = 96, 92, 87$, and 101 , respectively.

An approach introduced here to find the cluster members of $Y(k_0)$ under the context index makes use of two opposite-sliding windows along the context index, k , each containing L coefficients to make a total of $2L + 1$ cluster points. The choice of the parameter L is a trade-off between the locality of each cluster and the accuracy of its estimated standard deviation.

Illustrated in Fig. 9 are two possible conditions encountered by the opposite-sliding windows. Under the condition $L < k_0 < (M - L)$ (e.g., the window pair A-B for $k_0 = 128$ in Fig. 9), those $Y(k)$ whose context indices fall within the two opposite windows adjacent to the index k_0 are recruited as the members of $Y(k_0)$. On the other hand, under the conditions $k_0 \leq L$ (e.g., the window pair C-D for $k_0 = 10$ in Fig. 9) and $k_0 \geq (M - L)$, one of the L -point windows can no longer slide, and thus the members within the window become fixed. With such a condition, the cluster of $Y(k_0)$ is instead formed by the members of the non-sliding window that contains k_0 , and

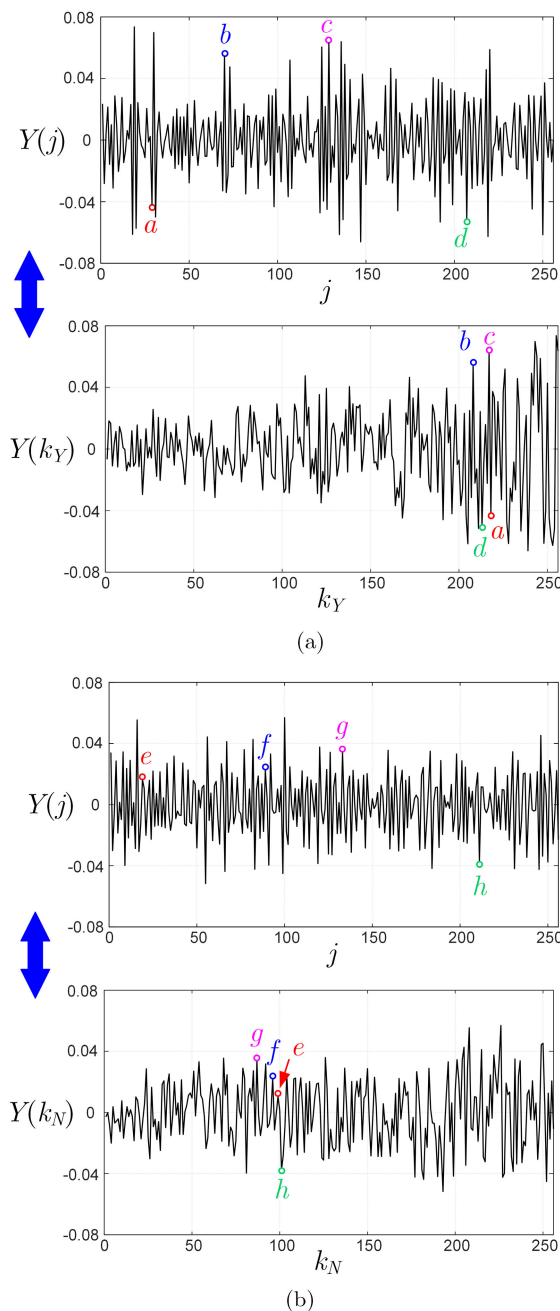


FIGURE 8. Plots of sample wavelet coefficients $Y(j)$ versus temporal index, and their corresponding $Y(k)$ versus context index after subspace projection onto (a) a noisy speech context subspace, and (b) a noise context subspace. Two groups of coefficient points (a, b, c, d) from an SA frame and (e, f, g, h) from an NSA frame are included to help examine the rearrangement, context clustering, and $j - k$ index association.

the sliding window adjacent to k_0 . Unlike the use of a single sliding window in [42], the two-window approach guarantees non-fixed members for all the clusters. This provides a consequent benefit to better estimation of the standard deviation associated with the random variable, $Y(k_0)$, particularly when k_0 is near both ends of the context index.

Having finished forming the clusters, the contextual standard deviation of the wavelet coefficient, $Y(k_0)$, which

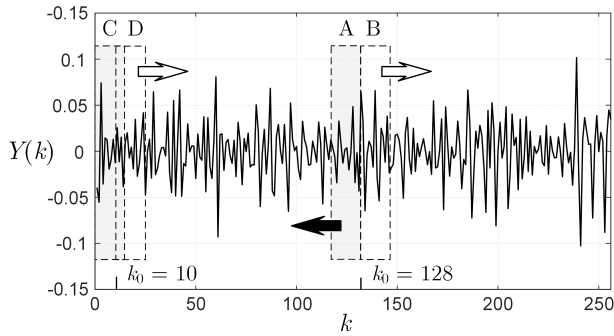


FIGURE 9. Opposite-sliding window pair A-B and C-D, with $L = 16$ at $M = 256$, for determining the clusters of $Y(k_0)$ at $k_0 = 128$ ($k_0 = M/2$) and $k_0 = 10$ ($k_0 \leq L$), respectively.

is $\sigma_Y(k_0)$, is estimated from the root mean square value of all the coefficients within its cluster, which is

$$\sigma_Y^2(k_0) = \frac{1}{2L + 1} \sum_{k \in B_{k_0}} Y^2(k) \quad (10)$$

where B_{k_0} denotes the cluster member set.

Fig. 10 shows sample plots of the standard deviations versus the context index of wavelet coefficients, calculated by the opposite-sliding window approach to form the cluster members, and by using (10) with $L = 16$ and $M = 256$.

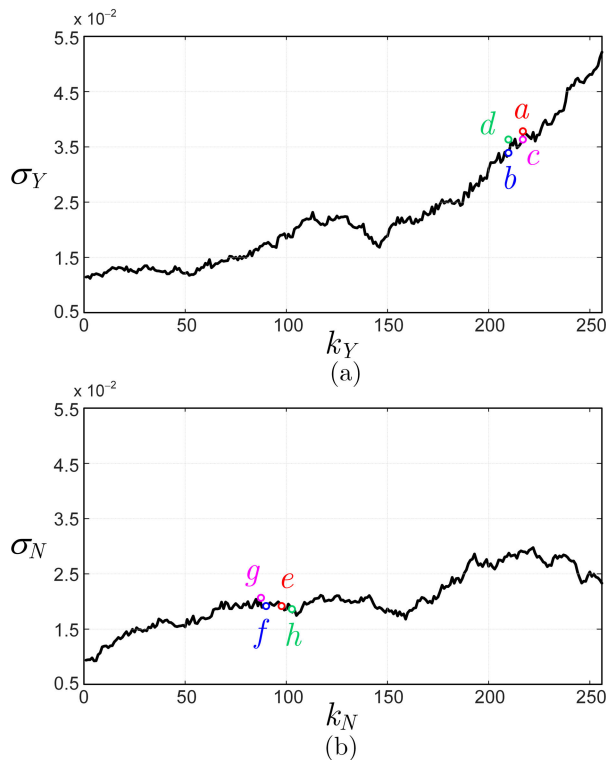


FIGURE 10. Plots of standard deviations versus context index, determined from the clustering of the sample wavelet coefficients using (10), for (a) $Y(k_Y)$ of Fig. 8(a) and (b) $Y(k_N)$ of Fig. 8(b), with $L = 16$.

Fig. 10(a) shows $\sigma_Y(k_Y)$ associated with each individual wavelet coefficient, $Y(k_Y)$, in Fig. 8(a). Fig. 10(b) shows $\sigma_N(k_N)$ associated with each $Y(k_N)$ in Fig. 8(b). Note that, each coefficient is modelled as a random variable under the GG pdf as explained in Section 3. By examining the coefficient samples a, b, c, d in Fig. 10(a), which correspond to the sample coefficients in Fig. 8(a), the standard deviation, $\sigma_Y(k_Y)$, of $Y(k_Y)$ are $\sigma_Y(218) = 0.0375$, $\sigma_Y(208) = 0.034$, $\sigma_Y(217) = 0.0373$, and $\sigma_Y(211) = 0.0359$. Similarly, for the samples e, f, g, h in Fig. 10(b), which correspond to the sample coefficients in Fig. 8(b), the standard deviation, $\sigma_Y(k_N)$, of $Y(k_N)$ are $\sigma_N(96) = 0.0199$, $\sigma_N(92) = 0.0198$, $\sigma_N(87) = 0.0202$, and $\sigma_N(101) = 0.0195$.

It should be noted that the standard deviation of a random variable is proportional to the spreading of its value. As noticed in Fig. 10(a) and Fig. 10(b), the value of $\sigma_Y(k_Y)$ falls at $k_Y = 146$, and that of $\sigma_Y(k_N)$ falls at $k_N = 155$ and $k_N = 230$. Therefore, as observed in Fig. 8, the corresponding wavelet coefficients $Y(k_Y)$ in Fig. 8(a) and $Y(k_N)$ in Fig. 8(b) tend to have lower values than their neighbouring coefficients. As also noticed from Fig. 10, σ_Y and σ_N tend to increase at a larger context index, where more coefficients with higher values are observed. In addition, σ_Y increases more sharply at high context indices. This reflects the typical characteristic of speech wavelet coefficients in $Y(k_Y)$, which exhibits more probability at higher amplitudes than its noise, with the shape parameter condition $\beta_S < \beta_N$ in the GG models.

Under the assumption of independent speech and noise signals, the estimated contextual standard deviation, $\hat{\sigma}_S$, of the clean speech coefficient, $S(j)$, can be determined by

$$\hat{\sigma}_S^2(j) = \max \left[\left(\sigma_Y^2(j) - \sigma_N^2(j) \right), \sigma_\epsilon^2 \right] \quad (11)$$

where σ_ϵ is a lower-bound value for σ_S , introduced to enable a control of possible low-level speech artifacts after denoising. It is important to point out that the calculation of $\hat{\sigma}_S^2$ in (11) is carried out in the temporal WT domain with the association between the temporal index, j , and the context index, k_Y, k_N , determined through P_Y, P_N , respectively, as already described and summarized in the diagram of Fig. 7.

C. FRAME-WISE CONTEXT UPDATE OF NOISE

Since a speech signal is processed on a frame-wise basis, it is evident from (11) that in order to compute $\hat{\sigma}_S, \sigma_Y$, and σ_N must be simultaneously available at the same time-frame index. One simple method is to set $\sigma_N = \sigma_Y$ during an NSA frame, and according to (11), $\hat{\sigma}_S$ is equal to σ_ϵ . However, this method can result in a significant attenuation of speech when an SA frame is incorrectly identified as an NSA frame by the voice activity detector (VAD). Moreover, during an SA frame, σ_N is not available and thus needs to be estimated from previous NSA frames. Whereas σ_Y can be calculated independently at a current frame, the estimation of σ_N , or $\hat{\sigma}_N$, during an SA frame requires the data between frames in order to obtain more accuracy [49]. This invariably

requires an approach to accomplish frame-to-frame correspondence between the contextual standard deviations. A practical means to the estimation of σ_N and the correspondence will be developed later in this section.

1) CONTEXT PAIRING

One practical approach widely employed in the STFT-based SE algorithms is to estimate and update noise signal during NSA frames. By assuming that noise is more stationary than speech, and that there exists certain noise correlation between successive frames, the estimated noise parameters can be maintained and employed during subsequent SA frames. To this end, the NSA and SA frames are detected by means of a VAD.

In the STFT, signals are transformed from the time domain onto the frequency domain, and the frame-invariant frequency index is employed as the correspondence for recursive updating and transferring operation of parameters between successive frames. By contrast, the wavelet coefficients of the WT is of temporal nature, making the coefficient index j varying between frames. To resolve this issue, it is proposed that the context index, k_N , obtained from the noise subspace projection in the context model, serves as the frame-invariant index to perform the contextual noise-parameter correspondence, both within the same frame and between successive frames. This is possible under the quasi-stationary noise assumption where the noise subspace and statistics between successive frames are considered to be correlated. In particular, the noise projection matrix, \mathbf{P}_N , is employed to project $|\mathbf{Y}|$ onto \mathbb{Q}_N in order to obtain the $j - k_N$ index association at each frame. The frame correspondence is then accomplished by means of pairing the noise parameters with the same noise context index k_N in \mathbb{Q}_N .

2) RECURSIVE UPDATE

Having established the noise context index as the frame-invariant correspondence, the estimation and update of the standard deviation, $\hat{\sigma}_N$, and the noise projection matrix, \mathbf{P}_N , can now be described. During the first few frames, it is assumed that $\mathbf{Y} = \mathbf{N}$. Thus, both the parameters $\hat{\sigma}_N$ and \mathbf{P}_N of the noise wavelet coefficients can be initialized by the context modeling. For subsequent frame index, l , if an NSA frame is detected, $\mathbf{Y}(l) = \mathbf{N}(l)$ will be used to partially modify the parameters in a recursive manner. This recursive update using partial estimates of the current frame serves not only to track the non-stationarity associated with the noise signal, but also to make the denoising algorithm more robust to inaccuracy of the VAD. On the other hand, if an SA frame is detected, $\mathbf{Y}(l) = \mathbf{S}(l) + \mathbf{N}(l)$, $\hat{\sigma}_N$ and \mathbf{P}_N obtained in the previous frame index, $l - 1$, are maintained.

Based upon the described operation, the update equations can be given as follows. Since the noise projection matrix, \mathbf{P}_N , is a function of \mathbf{U} , which is in turn derived from $\mathbf{Y}(l) = \mathbf{N}(l)$ during an NSA frame, it can be recursively estimated

and updated under the noise context index by

$$N(k_N, l) = \alpha_n N(k_N, l - 1) + (1 - \alpha_n) \tilde{N}(k_N, l) \quad (12a)$$

with

$$\tilde{N}(k_N, l) = \begin{cases} N(k_N, l - 1) & \text{SA} \\ Y(k_N, l) & \text{NSA} \end{cases} \quad (12b)$$

where α_n is a tracking factor between successive frames. Following this, $\mathbf{P}_N(l)$ is updated, and $\sigma_N(l)$ is obtained by the context modeling using $\mathbf{P}_N(l)$. Subsequently, the estimated noise standard deviation, $\hat{\sigma}_N(l)$, is recursively updated and estimated by

$$\hat{\sigma}_N^2(k_N, l) = \alpha_\sigma \cdot \hat{\sigma}_N^2(k_N, l - 1) + (1 - \alpha_\sigma) \cdot \tilde{\sigma}_N^2(k_N, l) \quad (13a)$$

with

$$\tilde{\sigma}_N^2(k_N, l) = \begin{cases} \hat{\sigma}_N^2(k_N, l - 1) & \text{SA} \\ \sigma_N^2(k_N, l) & \text{NSA} \end{cases} \quad (13b)$$

where α_σ is a smoothing factor between successive frames.

It can be summarized from the above equations that, when an NSA frame is detected, $\tilde{N}(l)$, and thus $\mathbf{P}_N(l)$, $\sigma_N(l)$ and $\tilde{\sigma}_N(l)$ are successively updated, followed by the update of the estimated noise standard deviation, $\hat{\sigma}_N(l)$, using (13a) and (13b). On the other hand, when an SA frame is detected, $\tilde{N}(l)$ is maintained at $\tilde{N}(l - 1)$, and $\tilde{\sigma}_N(l)$ is maintained at $\tilde{\sigma}_N(l - 1)$. Thus, it follows from (13a) and (13b) that $\mathbf{P}_N(l)$ and hence $\hat{\sigma}_N(l)$ are maintained as the previous values at the frame index, $l - 1$, for an SA frame.

V. SIMULATION AND PERFORMANCE EVALUATION

The proposed adaptive wavelet thresholding algorithm for SE based on the GG priors and frame-wise context modeling is hereafter denoted as GGFC. The performance of the GGFC algorithm was evaluated by making a comparison to most recently reported WT-based SE algorithms. With extensive evaluation and overall state-of-the-art performances, the wavelet thresholding using symmetric Kullback–Leibler divergence [36], denoted as the SKL algorithm, and the Rayleigh modeled TE operated wavelet thresholding [38], denoted here as the RTE algorithm, were included. Note that the SKL and RTE algorithms made use of the PWP transform, with the threshold formulation at the subband level derived by incorporating the Gaussian priors for the PWP coefficients in the SKL, and the Rayleigh priors for TE operated PWP coefficients in the RTE. Whereas the GGFC algorithm is fully equipped with the frame-wise noise estimation and update at the coefficient level, both the SKL and RTE algorithms relied on the existing noise estimation methods at the subband level, developed in [32] for the SKL, and in [43] for the RTE.

To appreciate the level of performance gained by the GGFC, a comparison with the wavelet thresholding using the Gaussian prior and the conventional context modeling in [42], denoted as the GC algorithm, was also included. Unlike the GGFC which makes use of the GG priors, the use

of the Gaussian prior in the GC algorithm is ineffective in capturing the varying statistical distributions of speech and noise wavelet coefficients (see also Fig. 2(a)), yielding a less accurate optimum threshold equation. In addition, the conventional context modeling only provides a simple noise estimation using the median value of the coefficients obtained in the highest subband, with no frame-wise updating and tracking of the noise statistics [42].

To benchmark the performance with the well-established STFT-based counterparts, a comparison to the MMSE log-spectral amplitude (LSA) algorithm [15] and the soft-mask with posteriori SNR uncertainty (SMPO) algorithm [17] was also given. Whereas the proposed algorithm operates in the wavelet domain and employs the GG priors, both the LSA and SMPO operated in the frequency domain and made use of the Gaussian priors in their formulation of the speech enhancing gain function. Unlike the proposed algorithm, the LSA and SMPO also incorporated speech presence probability and SNR uncertainty, respectively, for further improvement. Also note that, the LSA algorithm included the estimation and update of noise during speech pauses, and the SMPO relied on the continuous noise estimation method developed elsewhere in [43].

For the performance evaluation, the NOIZEUS database in [44], which is a standard noisy speech corpus designed specifically for evaluating speech enhancement algorithms, was employed. Note that, the GGFC algorithm is developed with emphasis on emerging voice-control applications, such as smart home appliances and in-vehicle voice navigation. With this application perspective, the selected practical noise conditions for performance evaluation, in addition to white Gaussian noise (WGN), were babble noise, car noise, and airport noise, which are of non-stationary and colored noise types. Whereas the WGN is stationary and employed mainly for primary testing of noise attenuation performance, the other selected noise types are non-stationary and typically present in practical noise environment of the targeted voice-control applications. Note that, the babble noise represents conversations among groups of people, the car noise represents in-vehicle sounds from engine, wind, etc., and the airport noise represents mixed types of indoor noise in a busy public area, including multi-tone noise sounds.

Whereas the comparison to the LSA and SMPO algorithms could be made extensively using the noise types and speech measures as summarized above, the comparison to the SKL and RTE algorithms were made based on the available data in [36] and [38], due to the lack of public access to their implementation codes. Note that, the speech data was based on the NOIZEUS database for the RTE algorithm, and the compatible TIMIT database for the SKL algorithm. Both the NOIZEUS and TIMIT databases have been widely utilized for SE performance evaluation, and yielded similar average results under the tested WGN and babble noise conditions. Also note that, for the SKL algorithm in [36], only the average performance across the noise types was reported.

A. SIMULATION RESULTS

1) PARAMETER SETUP OF THE GGFC ALGORITHM

The proposed GGFC algorithm was implemented using the Daubechies wavelet family, which has been extensively employed in other applications including image denoising. Note that, since the use of the Daubechies type with the order more than seven yielded similar performance results, the Daubechies-8 type was employed for efficient computation with some performance margins. The octave-band decomposition was chosen at three levels, i.e., the largest scale at $I = 3$, so as to cover three major frequency bands of typical speech, including low- and high-frequency voiced spectral bands, and high frequency unvoiced spectral band. An overcomplete wavelet expansion implemented using non-subsampled filter bank in [42] was adopted to obtain more attenuated artifacts in denoised signals. The VAD was based on a combination of the simple energy-ratio test using the average variance ratio, σ_Y^2/σ_N^2 , over the subband, and the zero-crossing rate test of the coefficients $Y^i(j)$ over the frame.

The details and values of the constant parameters employed in the GGFC algorithm were summarized below. The shape parameter of the clean speech in the GG model was set at $\beta_S = 0.5$, based on the plot of β_S variation against speech data set in Fig. 3(a). For the following parameters, they were optimized based on the perceptual evaluation of speech quality (PESQ) scores in 5-dB babble noise, and consistent results were obtained for other noise types. In the frame-wise context model, the number of the neighboring points, p , of the matrix \mathbf{U} , that forms the projection matrix \mathbf{P} in (9), from $p = 3$ up to $p = 9$ was verified through simulations to capture the locality of the contextual standard deviation of each coefficient $Y(j)$, where similar speech enhancement performances were obtained. Therefore, $p = 3$ (two neighboring and one parent coefficient points) was selected for efficient computation. The regularization parameter λ in (9), which helps control overfitting, was chosen to be 0.01. The number of coefficients L in (10) for one sliding window of the context cluster was set at $L = 16$. In the noise update equations (12) and (13), the tracking and smoothing factors were optimized at $\alpha_n = 0.22$ and $\alpha_\sigma = 0.98$, which help control the effect of abrupt noise fluctuation between frames, particularly under non-stationary noise conditions. Note that, for initialisation, we set $\alpha_n = \alpha_\sigma = 0$ at frame $l = 1$.

The details of the computed parameters in the GGFC algorithm are summarized as follows. The projection matrix, \mathbf{P} , which projects the wavelet coefficients into the context subspace is calculated using (9), with the context parameters of the coefficients computed using (8). The estimated standard deviations of noisy speech, σ_Y , and clean speech, σ_S , are given by (10) and (11), respectively. The estimated standard deviation of noise, σ_N , is given by (12) and (13). The estimated shape parameter of the noise in the GG model, β_N , is determined by using (6) and (7). The optimum threshold value, T_B , for removing noise from noisy

speech at the coefficient-level, through the soft threshold function in (2), is determined using (5).

Below is the outline in a step-by-step manner of the GGFC algorithm.

Step 1: Divide the input noisy speech, $y(t)$, into a sequence of overlapped frames to obtain $y(l)$ with frame length, M , at frame l .

Step 2: Apply the octave-band wavelet transform on $y(l)$ to obtain the noisy wavelet coefficients, $Y^i(j, l)$, at the coefficient index, j , of scale, i ($i = 1, 2, \dots, I$), where I is the largest number of scale.

Step 3: Use (8) and (9) to calculate the noisy speech projection matrix, $\mathbf{P}_Y^i(l)$, and project $Y^i(j, l)$ onto the noisy speech context subspace to obtain $Y^i(k_Y, l)$. Subsequently, apply the context clustering and calculate the standard deviation, $\sigma_Y^i(k_Y, l)$, using (10). Use the $j-k_Y$ association obtained from the subspace projection to map from $\sigma_Y^i(k_Y, l)$ to $\sigma_Y^i(j, l)$.

Step 4: Use (12a) and (12a) along with (6) and (7) to update the noise projection matrix, $\mathbf{P}_N^i(l)$, in the noise context subspace, k_N . Then, use $\mathbf{P}_N^i(l)$ to project $Y^i(j, l)$ onto the noise context subspace, apply the context clustering and calculate the noise standard deviation, $\sigma_N^i(k_N, l)$, at the context index k_N .

Step 5: Use (13a) and (13b) to update the estimated noise standard deviation, $\hat{\sigma}_N^i(k_N, l)$. Then apply the $j-k_N$ association obtained from the noise subspace projection to map from $\hat{\sigma}_N^i(k_N, l)$ to $\hat{\sigma}_N^i(j, l)$.

Step 6: Use (11) to obtain the standard deviation of the estimated clean speech, $\hat{\sigma}_S^i(j, l)$, by putting $\sigma_N(j) = \hat{\sigma}_N^i(j, l)$.

Step 7: If an NSA frame, update the estimated noise shape parameter, $\hat{\beta}_N^i(j, l)$, by using (7). Then, calculate the adaptive threshold, $T_B^i(j, l)$ by using (5), with the shape parameter $\beta_N = \hat{\beta}_N^i(j, l)$ and the standard deviations, $\sigma_S = \hat{\sigma}_S^i(j, l)$ and $\sigma_N = \hat{\sigma}_N^i(j, l)$.

Step 8: Apply the soft-threshold function in (2) to $Y^i(j, l)$ at $T = T_B^i(j, l)$ to obtain $\hat{S}^i(j, l)$.

Step 9: Apply the inverse wavelet transform to the estimated clean speech wavelet coefficients, $\hat{\mathbf{S}}^i(l)$, to obtain the time-domain clean speech, $\hat{s}(l)$. Reconstruct the enhanced speech signal, $\hat{s}(t)$, by the overlap-add method.

In the following simulations, the sampling frequency of the speech data was at 8 kHz, and the frame period at 32 ms, yielding the number of samples, $M = 256$, samples per frame. The frame overlap was chosen to be 50%. To obtain each SNR level, the active speech level of the clean speech was first determined. Subsequently, a noise segment of the same length as the speech was selected, appropriately scaled to reach the desired SNR level, and added to the clean speech signal.

The measures employed to evaluate performance of these investigated algorithms were the segmental SNR improvement (ΔSegSNR) for noise attenuation; the perceptual evaluation of speech quality (PESQ) score [50], the short-time objective intelligibility (STOI) score [51], and the cepstral distance (CD) measure [44], for speech quality. Different

levels of input SNR were employed ranging from -5 dB to $+15$ dB, with a 5-dB step. Each data point in the plots of the performance measures was based on average simulation results using a total of 20 sentences with ten male and ten female noisy speech signals selected from the NOIZEUS database [44]. These objective results shown in Fig. 11 to Fig. 14 were averaged from 20 enhanced speech signals. In addition, an investigation using the spectrograms of enhanced speech signals was also included.

2) NOISE ATTENUATION PERFORMANCE

The evaluation of noise attenuation performance was performed using the SegSNR improvement, ΔSegSNR , defined as the difference between the SegSNR of the output enhanced speech and that of the input noisy speech, during all SA frames. Since the SegSNR of the enhanced speech should be larger than that of noisy speech, higher ΔSegSNR values indicate better noise attenuation performance.

Under the WGN in Fig. 11(a), and the car noise in Fig. 11(c), the proposed GGFC algorithm offered significantly higher ΔSegSNR values than the WT-based GC and RTE algorithms. As compared to the LSA and SMPO algorithms, the GGFC yielded comparable ΔSegSNR performance, with slightly lower values at low to medium SNRs, and slightly higher at medium to high SNR levels.

Under the babble noise in Fig. 11(b), the GGFC algorithm still significantly outperformed the GC and RTE, and offered higher ΔSegSNR values than the SMPO at low to medium SNR levels, and the LSA at medium to high SNR levels.

Under the airport noise in Fig. 11(d), the GGFC outperformed the GC, LSA, and SMPO algorithms, with similar trends to the babble noise condition. Note that, results of the RTE algorithm under the airport noise were not available.

As compared to the SKL algorithm, the GGFC offered higher ΔSegSNR values than the average ΔSegSNR across the noise types, as illustrated in Fig. 11(a) to Fig. 11(d), except at 15-dB SNR level in the babble noise condition. However, at such a high SNR level, speech is less corrupted and hence the improvement is less critical. Moreover, it is evident that the noise attenuation offered by the SKL algorithm at other SNR levels was considerably lower than other algorithms.

3) SPEECH QUALITY PERFORMANCE VIA PESQ

The PESQ measure relies upon the psychoacoustics about how human listeners process tones and bands of noise. Specifically, it takes into account the frequency resolution non-uniformity, frequency-dependent sensitivity, and loudness perception non-linearity of human auditory processing [44]. The PESQ score normally ranging from 0.5 to 4.5 was demonstrated to have high correlation (> 0.92) with speech quality measured by the subjective listening tests evaluated using the Mean Opinion Score (MOS) [52]. A higher noise attenuation, better clean speech preservation, and more distributed residual noise in the enhanced speech spectrum yield a higher PESQ score.

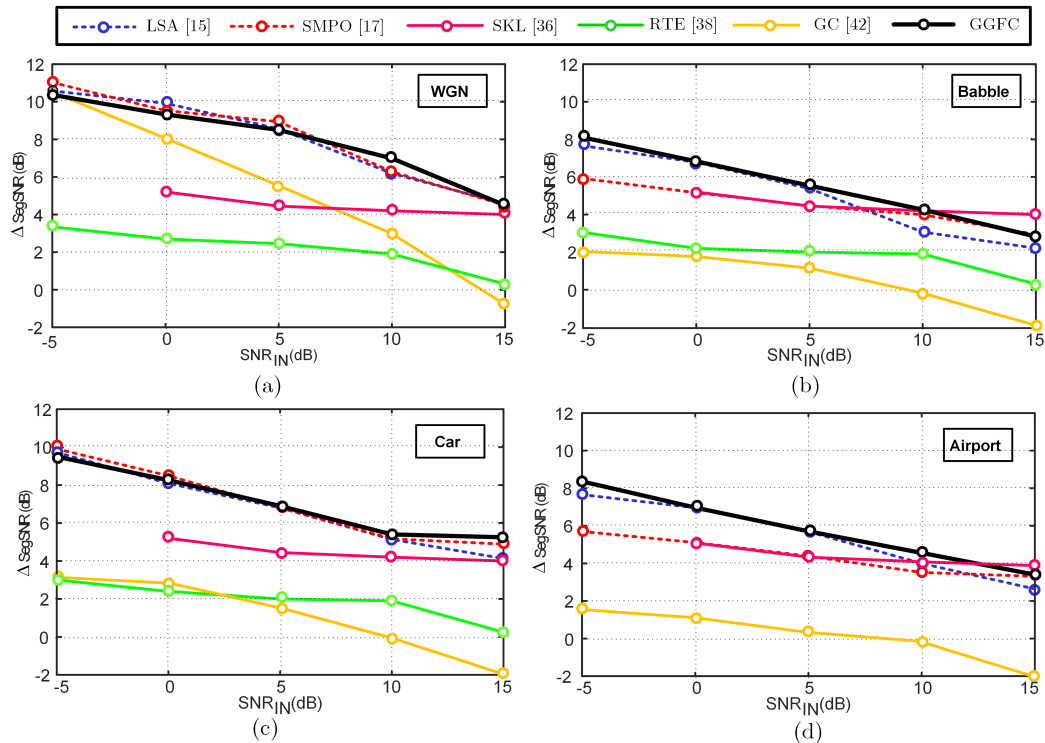


FIGURE 11. Δ SegSNR values of the enhanced speech signals using different SE algorithms in (a) WGN, (b) babble, (c) car, and (d) airport noise conditions. A higher Δ SegSNR value indicates better performance. The results of the SKL algorithm at -5 dB were not available in [36]. The RTE results under the airport noise condition were not available in [38].

The plots in Fig. 12 show the PESQ scores obtained by the GGFC algorithm in comparison to its WT-based and STFT-based counterparts under different noise conditions. Note that, the results of the RTE algorithms under the airport noise were not available.

Under the WGN condition in Fig. 12(a), the GGFC algorithm clearly exhibited higher PESQ scores than the GC, and comparable PESQ scores to the RTE, with slightly less at low and high SNRs, and slightly higher at medium SNR levels. As compared to its STFT-based counterparts, the GGFC yielded slightly less PESQ scores than the SMPO, and comparable PESQ scores to the LSA, with slightly less at $+15$ -dB SNR and slightly higher values at low to medium SNR levels.

Under the babble noise in Fig. 12(b) and the car noise in Fig. 12(c), the GGFC algorithm offered higher PESQs than the GC, RTE, and LSA algorithms. The GGFC algorithm also yielded higher PESQs than the SMPO, except at 0-dB SNR level under the car noise condition.

Under the airport noise in Fig. 12(d), whereas the GGFC algorithm yielded comparable PESQ scores at high SNR levels, it clearly offered higher PESQ scores than the LSA and SMPO algorithms at low to medium SNR levels. The GGFC also clearly offered higher PESQ scores than the GC algorithm.

In comparison to the WT-based SKL algorithm, the GGFC offers higher PESQ scores than the average PESQ values

obtained across the noise types, as shown in Fig. 12(a) to Fig. 12(d).

4) SPEECH INTELLIGIBILITY PERFORMANCE VIA STOI

The STOI measure is based on the correlation between temporal envelopes of the clean and degraded speeches in short-time segments [51]. The STOI scores provide high correlation with speech intelligibility and listening tests, with a higher STOI value indicating better performance.

Shown in the plots of Fig. 13 are the STOI scores obtained by the GC, GGFC, LSA, and SMPO algorithms under different noise conditions. Note that for the SKL and RTE algorithms in [36] and [38], the STOI measures were not reported.

Under the WGN condition in Fig. 13(a), the GGFC algorithm exhibited slightly higher scores than the LSA, and slightly lower scores than the SMPO algorithm. Under the babble noise in Fig. 13(b), it yielded comparable scores to the LSA and SMPO, with slightly higher scores at low SNRs, and slightly less at high SNR levels. Under the car noise in Fig. 13(c), it also yielded comparable STOI scores to the LSA and SMPO, with slightly lower scores at -5 -dB and $+15$ -dB SNR levels. Under the airport noise in Fig. 13(d), it exhibited slightly less STOI scores than the SMPO at high SNR levels, and comparable scores to the LSA. Under all the noise conditions, the GGFC algorithm offered higher STOI scores as compared to the GC algorithm.

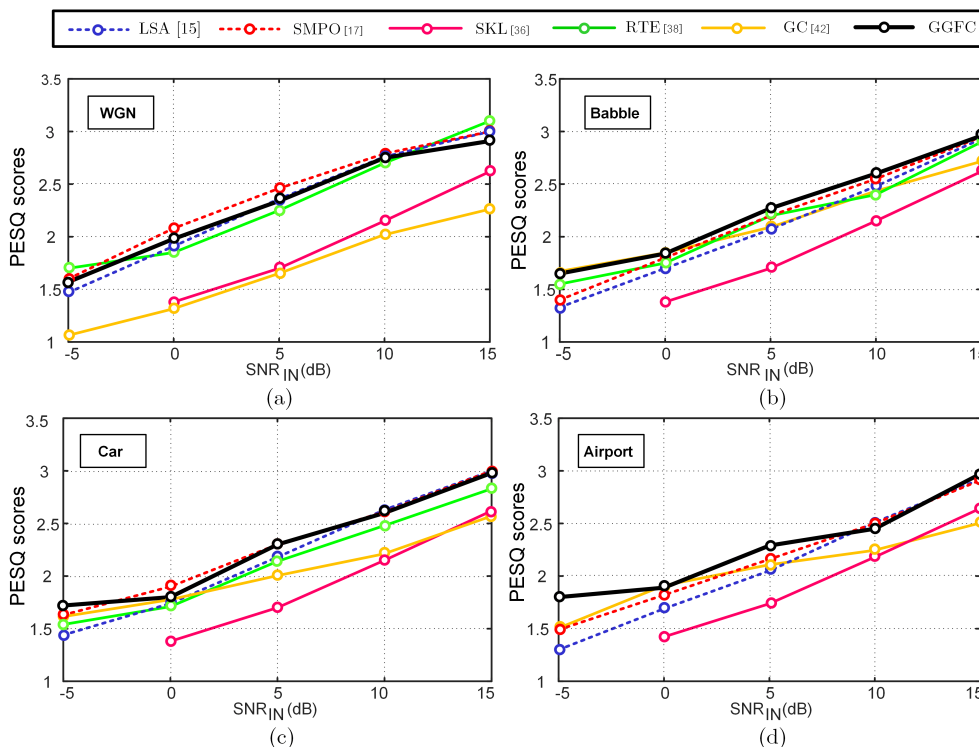


FIGURE 12. PESQ scores of the enhanced speech signals using different SE algorithms in in (a) WGN, (b) babble, (c) car, and (d) airport noise conditions. A higher PESQ score indicates better performance. The results of the SKL algorithm at -5dB were not available in [36]. The RTE results under the airport noise condition were not available in [38].

5) SPEECH QUALITY PERFORMANCE VIA CD

The CD measure is based on the log-spectral distance between the enhanced and clean speech spectra to determine their dissimilarity [44]. A higher noise attenuation and better preservation of enhanced speech spectrum yields lower CD values.

Fig. 14 shows the plots of the CD measures under different noise conditions obtained by the GGFC, GC and the STFT-based algorithms, where a lower CD value indicates a better performance. Note that, CD comparisons were not available for the SKL and RTE algorithms in [36] and [38].

Consider the CD results under the WGN in Fig. 14(a), the GGFC exhibited lower CD values than the GC and LSA, and comparable CD values to the SMPO, with slightly lower values at low to medium SNRs, and slightly higher at high SNR levels. Under the babble noise in Fig. 14(b), the car noise in Fig. 14(c), and the airport noise in Fig. 14(d), the GGFC algorithm yielded comparable CD values to the SMPO, and clearly offered lower CD values than the GC and LSA algorithms.

6) SPECTROGRAMS

A comparison among the WT-based GC and GGFC, and the STFT-based algorithms is also investigated via the spectrogram plots of their enhanced speech signals under the babble noise condition with the input SNR of 10 dB. The spectrogram plots of the clean speech and noisy speech signals with 10-dB input SNR of babble noise, and

their corresponding enhanced speech signals employing the SE algorithms are shown in Fig. 15.

Note that, by comparing the spectrograms of the enhanced, noisy, and clean speech signals, the objective and subjective measures can be implied. A higher noise attenuation in the enhanced speech indicates a higher segmental SNR improvement. Better preservation of the enhanced speech spectrum implies lower CD measures and higher PESQ scores.

By inspecting Fig. 15, it is evident that the GC algorithm yielded the highest residual noise, and the least preserved spectrum in the enhanced speech signal. Also, whereas the GGFC, LSA, and SMPO algorithms can significantly reduce the background noise, the spectral components of the enhanced speech obtained by the GGFC is better preserved. Furthermore, it can be noticed that the spectral components of the residual noise associated with the LSA and SMPO appears to be isolated, and those of the GGFC algorithm distribute more uniformly over frequencies, thereby entailing no musical noise effect.

7) SUBJECTIVE LISTENING TEST

A subjective listening test was carried out based on the 5-scale Mean Opinion Score (MOS) with 15 listeners. Listeners were asked to rate the sound quality of each speech signal from 1-poor, 2-bad, 3-fair, 4-good, and 5-excellent. Table 1 shows the MOS values for the tested clean and noisy speech signals with additive babble noise at a 5-dB input SNR. Also given in the table are the MOS values of the enhanced

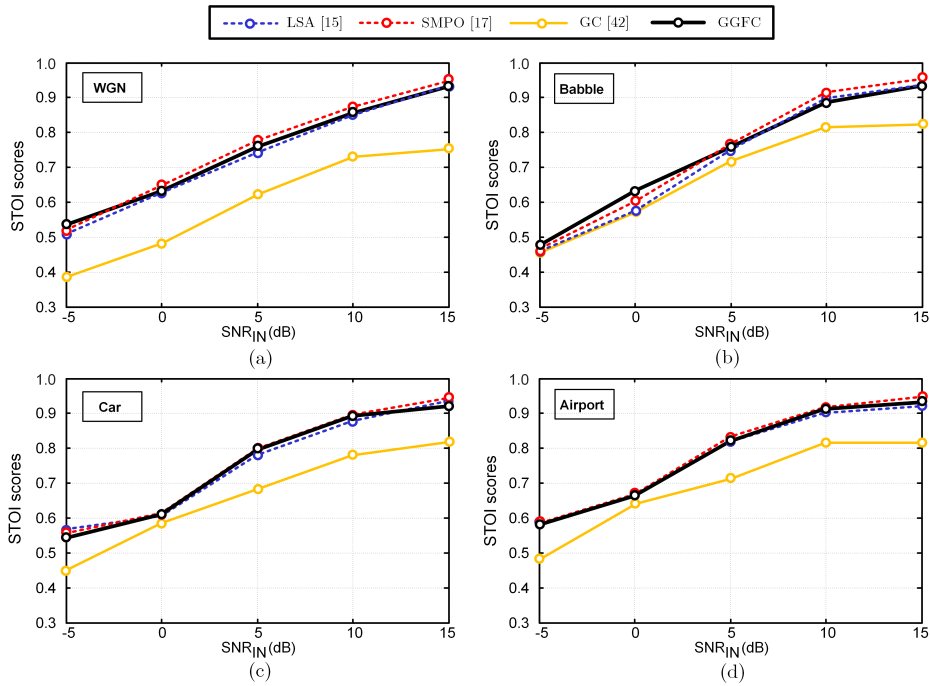


FIGURE 13. STOI scores of the enhanced speech signals using different SE algorithms in (a) WGN, (b) babble, (c) car, and (d) airport noise conditions. A higher STOI value indicates better performance. The SKL and RTE results in [36] and [38] were not available.

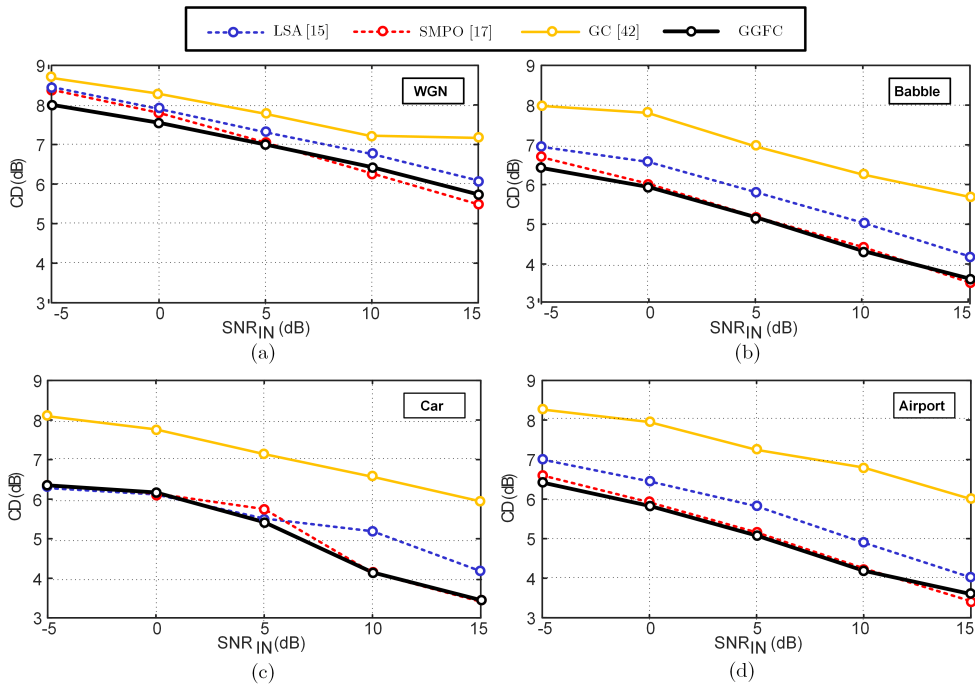


FIGURE 14. CD values of the enhanced speech signals using different SE algorithms in (a) WGN, (b) babble, (c) car, and (d) airport noise conditions. A lower CD value indicates better performance. The SKL and RTE results in [36] and [38] were not available.

speech signals using the LSA, SMPO, GC, and the proposed GGFC algorithms. It is evident from Table 1 that the listeners, on average, were more satisfied with the enhanced speech quality of the GGFC algorithm. It should be noted that, by listening to these enhanced speech signals, the residual noise of

the GC and GGFC was found to be similar to a white noise type, while those of the LSA and SMPO resembled a musical noise type, in line with the spectrogram plots in Fig. 15. Also, the subjective results were consistent with the highest PESQ score offered by the GGFC under the babble noise and 5-dB

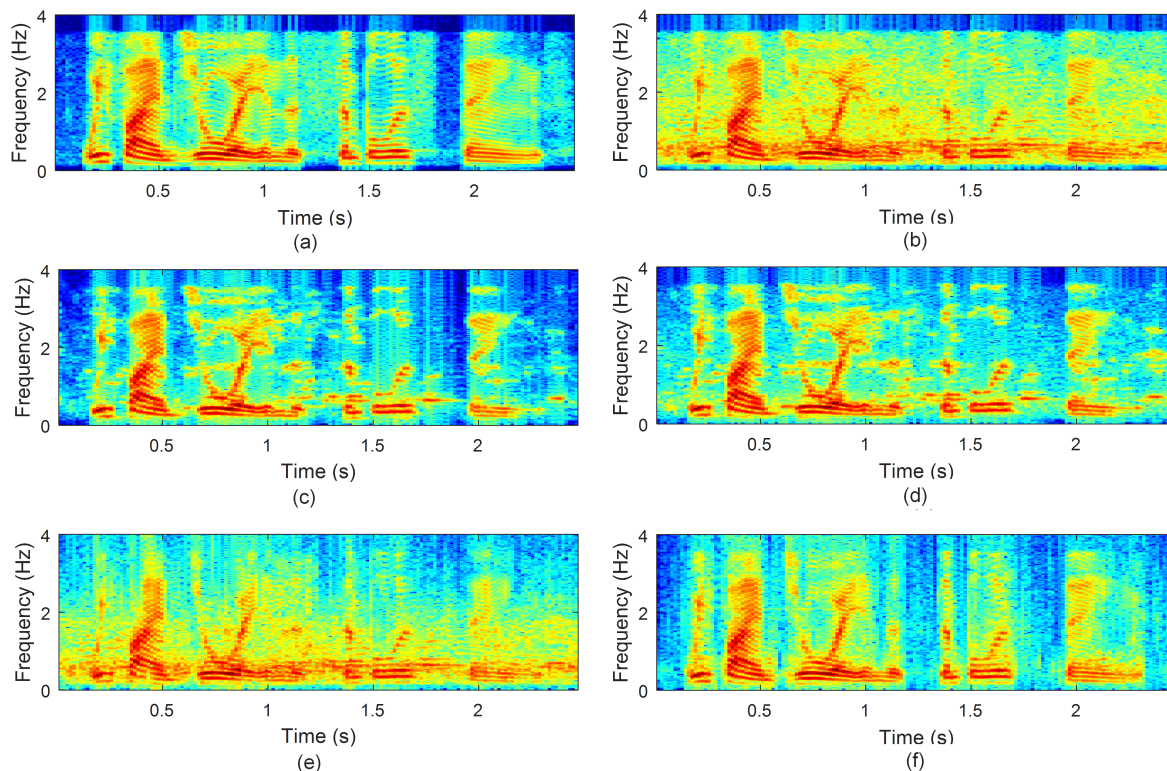


FIGURE 15. Spectrogram plots of (a) clean speech, (b) noisy speech signal with 10-dB SNR babble noise, and the enhanced speech signals using the (c) LSA, (d) SMPO, (e) GC, and (f) proposed GGFC algorithms. The sentence (“The birch canoe slid on the smooth planks”) was taken from the NOIZEUS database.

TABLE 1. MOS values of the investigated SE algorithms.

Signals	MOS values
clean speech	4.80
noisy speech (babble, input SNR = 5dB)	1.87
enhanced speech (LSA) [15]	2.87
enhanced speech (SMPO) [17]	2.93
enhanced speech (GC) [42]	2.45
enhanced speech (proposed GGFC)	3.20

SNR conditions in Fig. 12. Note that, the PESQ measure is highly correlated with the subjective MOS test [52].

B. OVERALL PERFORMANCE EVALUATION ACROSS NOISE TYPES

From the description based on Fig. 11 to Fig. 14 in the previous sections, it is clear that the proposed GGFC algorithms significantly outperformed the WT-based GC, RTE, and SKL algorithms. When benchmarked with the well-established STFT-based LSA and SMPO counterparts, the GGFC offered competitive performances in the non-stationary noise conditions, and slight underperformance in the WGN condition. This is mainly attributed to the direct use of the Gaussian priors in the derivation of the LSA and SMPO algorithms. Furthermore, unlike the GGFC algorithm which is developed in its intrinsic form in this work, additional enhancement techniques, also based on the Gaussian distribution, were incorporated, including the use of speech presence probability in the LSA, and SNR uncertainty

in the SMPO. Note however that, because their operations mainly rely on the Gaussian assumption, both STFT-based algorithms are less effective under various non-stationary noise types. As will be evident later in this subsection, more consistent and robust performance, particularly across non-stationary noise types in practical scenarios, can be achieved by the proposed GGFC algorithm.

To provide a comparative overall performance evaluation in a quantitative manner, the Δ_{SegSNR} , PESQ, STOI, and CD measures in Fig. 11 to Fig. 14 were averaged across all noise types at each of the input SNR level. The results are given in Table 2 to Table 5, where the boldface and underlined numbers indicate the best and second-best performances, respectively. Also, each table provides the average measures *with* and *without* the stationary WGN. Note that, the performance under the WGN is normally included for a preliminary test. Because typical speech noise is non-stationary by nature, the average measures across the non-stationary noise types should better provide a true comparative assessment among the investigated SE algorithms.

In terms of Δ_{SegSNR} in Table 2, it is evident that the proposed GGFC algorithm significantly outperformed the WT-based GC, RTE, and SKL algorithms at all the SNR levels. In comparison to the STFT-based LSA and SMPO, the GGFC offered the best performance except at 0-dB SNR level when including the WGN. Under practical non-stationary noise conditions, however, the

TABLE 2. The average values of ΔSegSNR of the enhanced speech signals using different SE algorithms, across all non-stationary noise types (babble, car, and airport noises) and all noise types (WGN, babble, car, and airport noises), for various input SNR levels. A higher ΔSegSNR value indicates better performance. The boldface and underlined numbers indicate the best and second-best performances, respectively.

input SNR (dB)	Average ΔSegSNR (non-stationary noises)						Average ΔSegSNR (all noises)					
	LSA	SMPO	RTE	SKL	GC	GGFC	LSA	SMPO	RTE	SKL	GC	GGFC
-5	<u>8.46</u>	7.64	3.03	-	2.04	8.68	<u>9.09</u>	8.76	3.14	-	5.94	9.17
0	<u>7.32</u>	6.56	2.30	5.20	1.71	7.39	8.13	7.51	2.44	5.20	4.21	<u>7.96</u>
5	<u>6.00</u>	5.38	2.00	4.45	0.81	6.09	<u>6.79</u>	6.57	2.16	4.45	2.43	6.83
10	4.15	<u>4.28</u>	1.90	4.20	-0.29	4.76	4.76	<u>4.87</u>	1.90	4.20	0.65	5.44
15	3.07	<u>3.80</u>	0.28	<u>3.98</u>	-2.01	4.00	3.50	<u>3.98</u>	0.28	<u>4.00</u>	-1.79	4.09

TABLE 3. The average values of PESQ scores of the enhanced speech signals using different SE algorithms, across all non-stationary noise types (babble, car, and airport noises) and all noise types (WGN, babble, car, and airport noises), for various input SNR levels. A higher PESQ score indicates better performance. The boldface and underlined numbers indicate the best and second-best performances, respectively.

input SNR (dB)	Average PESQ (non-stationary noises)						Average PESQ (all noises)					
	LSA	SMPO	RTE	SKL	GC	GGFC	LSA	SMPO	RTE	SKL	GC	GGFC
-5	1.36	1.51	1.55	-	<u>1.61</u>	1.72	1.39	1.53	<u>1.60</u>	-	1.48	1.69
0	1.71	<u>1.83</u>	1.73	1.38	1.82	1.84	1.76	1.90	1.77	1.38	1.71	<u>1.88</u>
5	2.10	<u>2.22</u>	2.17	1.70	2.06	2.29	2.17	<u>2.28</u>	2.20	1.70	1.95	2.30
10	<u>2.54</u>	2.53	2.44	2.15	2.28	2.55	<u>2.60</u>	2.59	2.53	2.15	2.21	2.61
15	<u>2.95</u>	2.94	2.87	2.62	2.57	2.97	<u>2.96</u>	2.95	2.94	2.62	2.48	2.97

TABLE 4. The average values of STOI measures of the enhanced speech signals using different SE algorithms, across all non-stationary noise types (babble, car, and airport noises) and all noise types (WGN, babble, car, and airport noises), for various input SNR levels. A higher STOI value indicates better performance. The boldface and underlined numbers indicate the best and second-best performances, respectively.

input SNR (dB)	Average STOI (non-stationary noises)				Average STOI (all noises)			
	LSA	SMPO	GC	GGFC	LSA	SMPO	GC	GGFC
-5	0.52	0.53	0.46	0.53	0.52	0.53	0.45	0.53
0	0.61	<u>0.62</u>	0.60	0.63	0.61	<u>0.62</u>	0.60	0.63
5	0.78	0.79	0.71	0.79	0.77	0.79	0.71	0.79
10	0.88	<u>0.89</u>	0.81	0.90	0.88	0.89	0.81	0.89
15	0.92	0.94	0.82	0.94	0.92	0.94	0.82	0.94

TABLE 5. The average values of CD measures of the enhanced speech signals using different SE algorithms, across all non-stationary noise types (babble, car, and airport noises) and all noise types (WGN, babble, car, and airport noises), for various input SNR levels. A lower CD value indicates better performance. The boldface and underlined numbers indicate the best and second-best performances, respectively.

input SNR (dB)	Average CD (non-stationary noises)				Average CD (all noises)			
	LSA	SMPO	GC	GGFC	LSA	SMPO	GC	GGFC
-5	6.76	<u>6.56</u>	8.08	6.40	7.25	<u>7.09</u>	8.22	6.86
0	6.38	<u>6.03</u>	7.81	5.98	6.82	<u>6.54</u>	7.93	6.43
5	5.72	<u>5.38</u>	7.14	5.23	6.17	<u>5.86</u>	7.31	5.23
10	5.04	<u>4.27</u>	6.61	4.22	5.54	4.87	6.78	<u>4.89</u>
15	4.13	3.45	5.99	<u>3.56</u>	4.71	4.06	6.33	<u>4.22</u>

GGFC algorithm outperformed the STFT-based LSA and SMPO counterparts at all input SNR levels.

In terms of the average PESQ performance in Table 3, the GGFC algorithm also offered significant improvement over the GC, RTE, and SKL across all the noise types and SNR levels. The GGFC outperformed the STFT-based LSA and SMPO counterparts, except at 0-dB SNR level when including the WGN. Similar to the average ΔSegSNR measures in Table 2, considering mainly non-stationary noise in practice, the GGFC algorithm offered the best performance at all input SNR levels.

In terms of the average STOI performance in Table 4, whereas the GGFC algorithm outperformed the GC and LSA algorithms, it practically shared the best STOI scores with the SMPO counterpart at all the SNR levels, across the average noise types with and without the stationary WGN.

In terms of the average CD performance in Table 5, it is clear that the GGFC algorithm outperformed both the GC and

LSA algorithms. In comparison to the STFT-based SMPO, it offered the best performance at low to medium SNR levels, and the second-best at high SNR levels when the WGN was included. Under the non-stationary noise types, although the GGFC algorithm was second-best at 15-dB SNR, it outperformed the SMPO at other lower SNR levels, where speech was more corrupted by noise, and hence improvement was more critical.

From the above comparative results, the followings can be deduced. The proposed GGFC algorithm significantly outperformed the WT-based algorithms under all the noise conditions and SNR levels. From the tables, the improvement could be as much as 226% over the RTE algorithms in terms of ΔSegSNR at 0-dB SNR, 36% over the SKL algorithm in terms of the PESQ at 0-dB SNR, 17.8% over the GC algorithm in terms of the STOI at -5-dB SNR, and 33.3% over the GC algorithm in terms of the CD at 15-dB SNR level. In comparison to the well-established STFT-based LSA and SMPO,

the GGFC algorithm practically offered the best overall performance measures across the non-stationary noise types at different SNR levels. Moreover, whereas the second-best performance in terms of the average ΔSegSNR in Table 2 and the PESQ scores in Table 3 were shared among the RTE, LSA, and SMPO algorithms, the GGFC always maintained the best performance, particularly under the non-stationary noise conditions. This clearly demonstrates not only the performance consistency over the SNR levels, but also the robustness across the performance measures, of the GGFC algorithm.

VI. CONCLUSION AND PROSPECTS

The proposed adaptive wavelet thresholding for speech enhancement based on the GG priors and frame-wise context modelling has been demonstrated to provide considerable improvement over other WT-based algorithms. When benchmarked against the well-established STFT-based counterparts, it has also been demonstrated to offer advantages not only in terms of overall performance, but also in terms of performance robustness, across various practical non-stationary noise conditions. Another important advantage includes no adverse effect from the residual musical noise, typically present in the enhanced speech of the STFT-based algorithms.

Because the proposed algorithm provides different optimum threshold values for each individual wavelet coefficient in each subband, independent of the number of bands, it also works well for a wideband noisy speech signal with a higher sampling frequency and higher speech and noise spectral contents. This is in contrast to its WT-based counterparts that make use of the WPT and subband-level thresholding, where higher spectral contents yield lower frequency resolution in each subband with consequent performance reduction.

Since the proposed algorithm has been developed in its intrinsic form, even greater improvement is entirely possible with additional refinement methods. In particular, by virtue of the MMSE framework and the context index as the frame-invariant correspondence similar to the frequency index, it is readily amenable to the incorporation of various enhancement methods already incorporated in the STFT-based algorithms, such as the inclusion of speech presence probability, SNR uncertainty, the use of continuous noise estimation, etc. With the time-frequency multi-resolution characteristic that makes the wavelet signal processing inherently more suitable to handle non-stationary speech and its noise signals, it is expected that the proposed algorithm will give the wavelet thresholding for speech enhancement a renewed impetus in its quest to be a viable alternative to the STFT-based counterparts.

REFERENCES

- [1] F. Weng, P. Angkititrukul, E. E. Shriberg, L. Heck, S. Peters, and J. H. L. Hansen, "Conversational in-vehicle dialog systems: The past, present, and future," *IEEE Signal Process. Mag.*, vol. 33, no. 6, pp. 49–60, Nov. 2016.
- [2] M. Pleva, J. Juhar, S. Ondas, C. R. Hudson, C. L. Bethel, and D. W. Carruth, "Novice user experiences with a voice-enabled human-robot interaction tool," in *Proc. IEEE Int. Conf. Radioelektronika*, Pardubice, Czech Republic, Apr. 2019, pp. 1–5.
- [3] P. Lei, M. Chen, and J. Wang, "Speech enhancement for in-vehicle voice control systems using wavelet analysis and blind source separation," *IET Signal Process.*, vol. 13, no. 4, pp. 693–702, 2019.
- [4] K. Moskvitch, "The Internet of Things 2.0: When things start to listen," *Eng. Technol.*, vol. 12, no. 1, pp. 63–65, Feb. 2017.
- [5] T. Backstrom, "Speech coding, speech interfaces and IOT-opportunities and challenges," in *Proc. 52nd Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, USA, Oct. 2018, pp. 1931–1935.
- [6] R. Haeb-Umbach, S. Watanabe, T. Nakatani, M. Bacchiani, B. Hoffmeister, M. L. Seltzer, H. Zen, and M. Souden, "Speech processing for digital home assistants: Combining signal processing with deep-learning techniques," *IEEE Signal Process. Mag.*, vol. 36, no. 6, pp. 111–124, Nov. 2019.
- [7] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.
- [8] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Orlando, FL, USA, May 2002, p. 44164.
- [9] S. Hayashi, H. Inukai, and M. Sugimoto, "A subtractive-type speech enhancement using the perceptual frequency-weighting function," *IEICE Trans. Fundamentals Electron., Commun. Comput. Sci.*, vols. 92, no. 1, pp. 226–234, 2009.
- [10] C. Li and W.-J. Liu, "A novel multi-band spectral subtraction method based on phase modification and magnitude compensation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Prague, Czech Republic, May 2011, pp. 4760–4763.
- [11] R. Miyazaki, H. Saruwatari, T. Inoue, Y. Takahashi, K. Shikano, and K. Kondo, "Musical-noise-free speech enhancement based on optimized iterative spectral subtraction," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 7, pp. 2080–2094, Sep. 2012.
- [12] T. T. Aung, H. Thumchirdchupong, N. Tangsangumvisai, and A. Nishihara, "Two-microphone subband noise reduction scheme with a new noise subtraction parameter for speech quality enhancement," *IET Signal Process.*, vol. 9, no. 2, pp. 130–142, Apr. 2015.
- [13] S. Kay, *Fundamental of Statistical Signal Processing: Estimation Theory*, 1st ed. Upper Saddle River, NJ, USA: Prentice-Hall, 1993, pp. 344–350.
- [14] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.
- [15] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr. 1985.
- [16] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized gamma priors," *IEEE Trans. Audio, Speech Language Process.*, vol. 15, no. 6, pp. 1741–1752, Aug. 2007.
- [17] Y. Lu and P. C. Loizou, "Estimators of the magnitude-squared spectrum and methods for incorporating SNR uncertainty," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 5, pp. 1123–1137, Jul. 2011.
- [18] C. H. You, S. N. Koh, and S. Rahardja, "Beta-order MMSE spectral amplitude estimation for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 4, pp. 475–486, Jul. 2005.
- [19] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.
- [20] B. Chen and P. C. Loizou, "A Laplacian-based MMSE estimator for speech enhancement," *Speech Commun.*, vol. 49, no. 2, pp. 134–143, Feb. 2007.
- [21] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP J. Adv. Signal Process.*, vol. 2005, no. 7, pp. 1110–1126, Dec. 2005.
- [22] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 845–856, Sep. 2005.
- [23] O. Cappe, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 345–349, Apr. 1994.
- [24] R. Martin, "Statistical methods for the enhancement of noisy speech," in *Speech Enhancement*, J. Benesty, S. Makino, and J. Chen, Eds. New York, NY, USA: Springer-Verlag, 2005, pp. 43–65.
- [25] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Process. Lett.*, vol. 9, no. 4, pp. 113–116, Apr. 2002.

- [26] P. C. Loizou, "Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 857–869, Sep. 2005.
- [27] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, Sep. 1994.
- [28] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 613–627, May 1995.
- [29] D. L. Donoho, I. M. Johnstone, G. Kerkycharian, and D. Picard, "Wavelet shrinkage: Asymptopia?" *J. Roy. Stat. Soc., B*, vol. 57, no. 2, pp. 301–369, 1995.
- [30] S. G. Chang, B. Yu, and M. Vetterli, "Wavelet thresholding for multiple noisy image copies," *IEEE Trans. Image Process.*, vol. 9, no. 9, pp. 1631–1635, Sep. 2000.
- [31] M. Bahoura and J. Rouat, "Wavelet speech enhancement based on the Teager energy operator," *IEEE Signal Process. Lett.*, vol. 8, no. 1, pp. 10–11, Jan. 2001.
- [32] Y. Ghanbari and M. R. Karami-Mollaei, "A new approach for speech enhancement based on the adaptive thresholding of the wavelet packets," *Speech Commun.*, vol. 48, no. 8, pp. 927–940, Aug. 2006.
- [33] G. Lee, S. D. Na, K. Seong, J.-H. Cho, and M. N. Kim, "Speech enhancement algorithm using recursive wavelet shrinkage," *IEICE Trans. Inf. Syst.*, vol. 99, no. 7, pp. 1945–1948, 2016.
- [34] M. Zhao and W.-P. Zhu, "Adaptive wavelet packet thresholding with iterative Kalman filter for speech enhancement," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Montreal, QC, Canada, Nov. 2017, pp. 71–75.
- [35] M. T. Johnson, X. Yuan, and Y. Ren, "Speech signal enhancement through adaptive wavelet thresholding," *Speech Commun.*, vol. 49, no. 2, pp. 123–133, Feb. 2007.
- [36] S. Tabibian, A. Akbari, and B. A. NaserSharif, "Speech enhancement using wavelet thresholding method based on symmetric Kullback–Leibler divergence," *Signal Process.*, vol. 106, pp. 184–197, Jan. 2015.
- [37] M. T. Islam, C. Shahnaz, W.-P. Zhu, and M. O. Ahmad, "Speech enhancement based on student t modeling of teager energy operated perceptual wavelet packet coefficients and a custom thresholding function," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 11, pp. 1800–1811, Nov. 2015.
- [38] M. T. Islam, C. Shahnaz, W.-P. Zhu, and M. O. Ahmad, "Rayleigh modeling of teager energy operated perceptual wavelet packet coefficients for enhancing noisy speech," *Speech Commun.*, vol. 86, pp. 64–74, Feb. 2017.
- [39] Y. Hu and P. C. Loizou, "Speech enhancement based on wavelet thresholding the multitaper spectrum," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 1, pp. 59–67, Jan. 2004.
- [40] K. Ghribi, M. Djendi, and D. Berkani, "A wavelet-based forward BSS algorithm for acoustic noise reduction and speech enhancement," *Appl. Acoust.*, vol. 105, pp. 55–66, Apr. 2016.
- [41] K. Ghribi, M. Djendi, and D. Berkani, "Thresholding wavelet-based forward BSS algorithm for speech enhancement and complexity reduction," in *Proc. 2nd Int. Conf. Natural Language Speech Process. (ICNLSP)*, Algiers, Algeria, Apr. 2018, pp. 1–6.
- [42] S. G. Chang, B. Yu, and M. Vetterli, "Spatially adaptive wavelet thresholding with context modeling for image denoising," *IEEE Trans. Image Process.*, vol. 9, no. 9, pp. 1522–1531, Sep. 2000.
- [43] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.
- [44] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. Boca Raton, FL, USA: CRC Press, 2007, pp. 465–584.
- [45] *MATLAB R2019a*, Mathworks, Natick, MA, USA, 2019
- [46] H. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. Int. Conf. Autom. Speech Recognit. (ASR)*, Paris, France, Apr. 2000, pp. 181–188.
- [47] J. A. Dominguez-Molina, G. Gonzalez-Farias, and R. M. Rodriguez-Dagnino, "A practical procedure to estimate the shape parameter in the generalized Gaussian distribution," Centro de Investigacion en Matematicas, Guanajuato Univ., Guanajuato, Mexico, Tech. Rep., 2003.
- [48] C.-Y. Lu, H. Min, Z.-Q. Zhao, L. Zhu, D.-Sh. Huang, and S. Yan, "Robust and efficient subspace segmentation via least squares regression," in *Computer Vision—ECCV 2012 (Lecture Notes in Computer Science)*, vol. 7578, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Berlin, Germany: Springer, 2012, pp. 347–360.
- [49] S. G. Tanyer and H. Ozer, "Voice activity detection in nonstationary noise," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 4, pp. 478–482, Jul. 2000.
- [50] *Perceptual Evaluation of Speech Quality (PESQ): An Objective Method For End-To-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs*, Standard ITU-T Recommendation P.862, 2001.
- [51] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of Time–Frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.
- [52] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.



PARINA CHAKRABORTY BHATTACHARYA

received the B.Sc. degree (Hons.) in physics and the B.Tech. and M.Tech. degrees in electronics and communication engineering from the University of Calcutta, India, in 2006, 2009, and 2011, respectively. She is currently pursuing the Ph.D. degree in signal processing with the Department of Electrical Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok, Thailand.

Her research interests include speech enhancement, multirate speech processing, and so on.



NISACHON TANGSANGIUMVISAI

(Senior Member, IEEE) was born in Bangkok, Thailand, in 1974. She received the M.Eng. degree in electrical and electronic engineering and the Ph.D. degree in signal processing from the Department of Electrical and Electronics Engineering, Imperial College London, U.K., in 1997 and 2001, respectively.

She was with the Department of Electrical Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok. In 2011, she was an Associate Professor with the Department of Electrical Engineering, Faculty of Engineering, Chulalongkorn University. Her research interests include adaptive signal processing, noise reduction techniques for speech enhancement, and so on. She was a recipient of the Royal Thai Scholarship to study in the U.K., from 1992 to 2001. She received the Research Fellowship from the Japan Society for the Promotion of Science (JSPS) to conduct research at the Tokyo Institute of Technology, Japan, in 2005.



APISAK WORAPISHET

(Senior Member, IEEE) received the B.Eng. degree (Hons.) in electrical engineering from the King Mongkut's Institute of Technology, Ladkrabang, Bangkok, Thailand, in 1990, the M.Eng.Sc. degree in electrical engineering from the University of New South Wales, Kensington, NSW, Australia, in 1995, and the Ph.D. degree in electrical engineering from the Imperial College London, U.K., in 2000.

Since 1990, he has been with the Mahanakorn University of Technology, Bangkok, where he is currently a Professor of electronic engineering. He is also the Director of the Mahanakorn Microelectronics Research Center (MMRC) and a Lecturer with the Mahanakorn Institute of Innovation (MII). His current research interests include RF/microwave passive and active and integrated circuits. He is a member of the Analog Signal Processing Technical Committee (ASPTC), the IEEE Circuits and Systems Society. He also serves as an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS-I: REGULAR PAPERS. He served as the Editor-in-Chief for *ECTI Transactions on Electrical Engineering, Electronics, and Communications*.

• • •