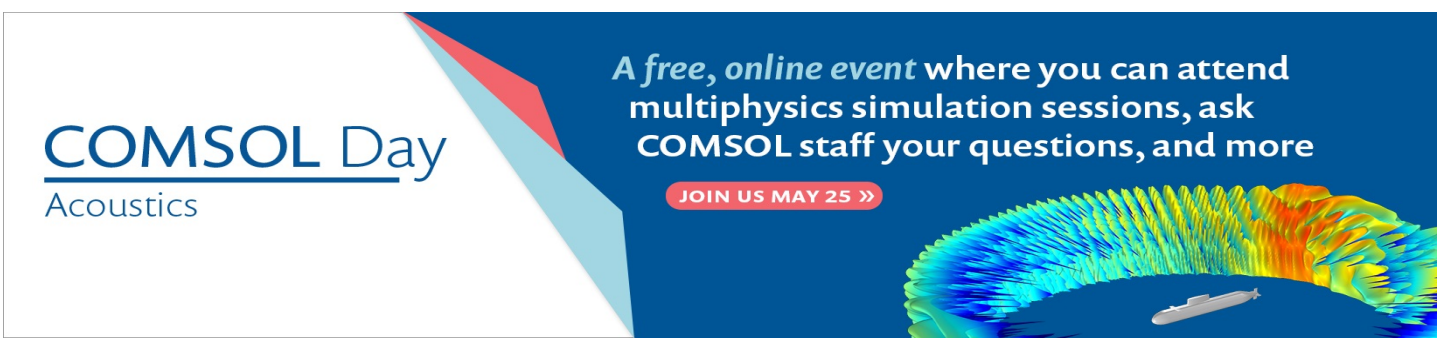
Spectral distortion level resulting in a just-noticeable difference between an *a priori* signal-to-noise ratio estimate and its instantaneous case

Aaron Nicolson, and Kuldip K. Paliwal

Citation: The Journal of the Acoustical Society of America **148**, 1879 (2020); doi: 10.1121/10.0002113
View online: https://doi.org/10.1121/10.0002113
View Table of Contents: https://asa.scitation.org/toc/jas/148/4
Published by the Acoustical Society of America



JASA ARTICLE

Spectral distortion level resulting in a just-noticeable difference between an *a priori* signal-to-noise ratio estimate and its instantaneous case

Aaron Nicolson[a] and Kuldip K. Paliwal[b]
Signal Processing Laboratory, Griffith University, Brisbane, Queensland 4111, Australia

ABSTRACT:
Minimum mean-square error (MMSE) approaches to speech enhancement are widely used in the literature. The quality of enhanced speech produced by an MMSE approach is directly impacted by the accuracy of the employed *a priori* signal-to-noise ratio (SNR) estimator. In this paper, the *a priori* SNR estimate spectral distortion (SD) level that results in a just-noticeable difference (JND) in the perceived quality of MMSE approach enhanced speech is found. The JND SD level is indicative of the accuracy that an *a priori* SNR estimator must exceed to have no impact on the perceived quality of MMSE approach enhanced speech. To measure the JND SD level, listening tests are conducted across five SNR levels, five noise sources, and two MMSE approaches [the MMSE short-time spectral amplitude (MMSE-STSA) estimator and the Wiener filter]. A statistical analysis of the results indicates that the JND SD level increases with the SNR level, is higher for the MMSE-STSA estimator, and is not impacted by the type of background noise. Following the literature, a significant improvement in *a priori* SNR estimation accuracy is required to reach the JND SD level. © *2020 Acoustical Society of America*. https://doi.org/10.1121/10.0002113

(Received 2 March 2020; revised 23 August 2020; accepted 15 September 2020; published online 7 October 2020)
[Editor: Michael I. Mandel] Pages: 1879–1889

I. INTRODUCTION

Minimum mean-square error (MMSE) approaches to speech enhancement, such as the MMSE short-time spectral amplitude (MMSE-STSA) estimator (Ephraim and Malah, 1984), are widely used in the literature. Their performance is highly dependent upon the accuracy of the used *a priori* SNR estimator. One measure used to determine the accuracy of an *a priori* SNR estimate is spectral distortion (SD) (Nicolson and Paliwal, 2019). SD is computed between an *a priori* SNR estimate and its instantaneous case. The instantaneous *a priori* SNR is computed from the unobserved clean speech and noise of the observed noisy speech.

A deep learning approach to *a priori* SNR estimation was recently proposed, which attained significantly lower SD levels than that of previous *a priori* SNR estimators (Nicolson and Paliwal, 2019), such as the decision-directed (DD) approach (Ephraim and Malah, 1984) and the harmonic regeneration noise reduction (HRNR) technique (Plapous *et al.*, 2005). The recent improvement in *a priori* SNR estimation accuracy has enabled MMSE approaches to outperform recent deep learning approaches to speech enhancement (Nicolson and Paliwal, 2019; Nikzad *et al.*, 2020; Roy *et al.*, 2020a,b; Zhang *et al.*, 2020).

It is clear that *a priori* SNR estimation has improved significantly as of late. What is unknown, however, is the SD level required by an MMSE approach to produce enhanced speech indistinguishable to that of the instantaneous case. Determining this SD level thus forms the purpose of this study, as it will provide a target level of accuracy for *a priori* SNR estimation research. We further define the target level of accuracy as the *a priori* SNR estimate SD level that results in a just-noticeable difference (JND) in the perceived quality of MMSE approach enhanced speech. That is, the SD level that causes a JND between: (1) the enhanced speech produced by an MMSE approach utilising the instantaneous *a priori* SNR and (2) the enhanced speech produced by an MMSE approach utilising the estimated *a priori* SNR. For an *a priori* SNR estimator to have no impact on the perceived quality of MMSE approach enhanced speech, it must attain an SD level lower than that of the JND SD level.

In this study, a series of listening tests are conducted to measure the JND SD level. For other psychoacoustic JNDs recently reported in the literature, please see Agus *et al.* (2018), Alkahtani (2019), Boucher *et al.* (2019), Chappel *et al.* (2016), and Nadiroh and Arifianto (2018). We investigate the JND SD level over a range of conditions. These include multiple SNR levels, MMSE approaches, and noise sources. Two popular MMSE approaches are tested, namely, the MMSE-STSA estimator and Wiener filter (WF) (Loizou, 2013). We also compare the SD levels of current *a priori* SNR estimators in the literature to that of the JND SD level.

This paper is organised as follows. In Sec. II, the analysis, modification, and synthesis (AMS) framework and the MMSE approaches are described. The experimental setup for the listening tests is described in Sec. III. A statistical

[a]Author to whom correspondence should be addressed: aaron.nicolson@griffithuni.edu.au, ORCID: 0000-0002-7163-1809.
[b]ORCID: 0000-0002-3553-3662.

analysis is performed in Sec. IV to determine how each condition impacts the JND SD level. Also included in Sec. IV is a comparison between the SD levels of current *a priori* SNR estimators in the literature to that of the JND SD levels. Conclusions are drawn in Sec. V.

II. BACKGROUND

A. Analysis, modification, and synthesis framework

The short-time Fourier analysis, modification, and synthesis (AMS) framework is used here for speech enhancement (Allen, 1977; Allen and Rabiner, 1977). The AMS framework consists of three stages: (1) the analysis stage, where noisy speech undergoes short-time Fourier transform (STFT) analysis, (2) the modification stage, where the noisy speech magnitude spectrum is modified, and (3) the synthesis stage, where the enhanced speech is synthesised by applying the inverse STFT. A block diagram of the AMS framework is shown in Fig. 1.

In the time-domain, the noisy speech signal, $x[n]$, is given by

$$x[n] = s[n] + d[n], \tag{1}$$

where $s[n]$ and $d[n]$ denote the clean speech and uncorrelated additive noise, respectively, and n denotes the discrete-time index. The noisy speech is analysed frame-wise using the running STFT (Vary and Martin, 2006),

$$X[l,k] = \sum_{n=0}^{N_d-1} x[n+lN_s]w[n]e^{-j2\pi nk/N_d}, \tag{2}$$

where l denotes the frame index, k denotes the discrete-frequency index, N_d denotes the frame duration in discrete-time samples, N_s denotes the frame shift in discrete-time samples, and $w[n]$ is an analysis window function. In polar form, the noisy speech spectrum is expressed as

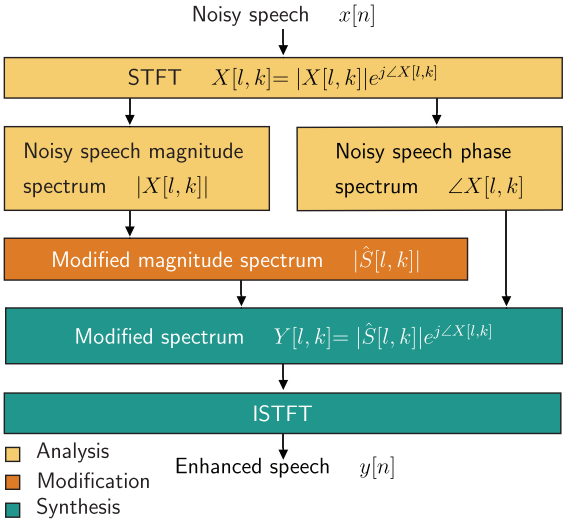


FIG. 1. (Color online) Short-time Fourier AMS speech enhancement framework.

$$X[l,k] = |X[l,k]|e^{j\angle X[l,k]}, \tag{3}$$

where $|X[l,k]|$ and $\angle X[l,k]$ denote the noisy speech magnitude and phase spectra, respectively. Similarly, the clean speech magnitude and phase spectra are denoted as $|S[l,k]|$ and $\angle S[l,k]$, respectively, and the noise magnitude and phase spectra are denoted as $|D[l,k]|$ and $\angle D[l,k]$, respectively.

The modified magnitude spectrum is then formed by enhancing the noisy speech magnitude spectrum. The modified magnitude spectrum is an estimate of the clean speech magnitude spectrum, and is denoted by $|\hat{S}[l,k]|$. The modified spectrum is constructed by combining the modified magnitude spectrum with the noisy speech phase spectrum

$$Y[l,k] = |\hat{S}[l,k]|e^{j\angle X[l,k]}. \tag{4}$$

The synthesis stage involves applying the inverse STFT to the modified spectrum. First, the inverse discrete Fourier transform (DFT) is applied to the modified spectrum

$$y_f[l,n] = \frac{1}{N_d}\sum_{k=0}^{N_d-1} Y[l,k]e^{j2\pi nk/N_d}, \tag{5}$$

where $y_f[l,n]$ is the framed enhanced speech. The least-squares overlap-add method is subsequently applied to produce the final enhanced speech (Crochiere, 1980; Griffin and Jae Lim, 1984),

$$y[n] = \frac{\sum_{l=-\infty}^{\infty} w[n-lN_s]y_f[l,n-lN_s]}{\sum_{l=-\infty}^{\infty} w^2[n-lN_s]}, \tag{6}$$

where $w[n]$ is a synthesis window function.

In this study, the Hamming window function is used for analysis and synthesis, with a frame-duration of 32 ms ($N_d = 512$) and a frame-shift of 16 ms ($N_s = 256$). The 257-point single-sided noisy speech magnitude spectrum, which includes both the DC frequency component and the Nyquist frequency component is modified.

B. MMSE approaches to speech enhancement

In this investigation, we aim to determine how the selected MMSE approach impacts the JND SD level. The two MMSE approaches that are evaluated include the MMSE-STSA estimator and the WF. The procedure to enhance speech using the MMSE-STSA estimator and the WF is described in this subsection. The MMSE-STSA estimator optimally estimates [in the mean squared error (MSE) sense] the magnitude spectrum of the clean speech (Ephraim and Malah, 1984). Similarly, the WF approach optimally estimates (in the MSE sense) the complex DFT coefficients of the clean speech (Loizou, 2013). The MMSE-STSA and WF approaches estimate the magnitude spectrum of the clean speech by applying a gain function, $G[l,k]$, to $|X[l,k]|$:

https://doi.org/10.1121/10.0002113

$$|\hat{S}[l,k]| = G[l,k]|X[l,k]|. \tag{7}$$

The result of Eq. (7) is used to compute the modified spectrum in Eq. (4).

An important parameter used to compute the gain function of an MMSE approach is the *a priori* SNR. The *a priori* SNR (McAulay and Malpass, 1980) of a noisy speech spectral component is defined as

$$\xi[l,k] = \frac{\lambda_s[l,k]}{\lambda_d[l,k]}, \tag{8}$$

where $\lambda_s[l,k] = \mathrm{E}\{|S[l,k]|^2\}$ is the variance of the clean speech spectral component, and $\lambda_d[l,k] = \mathrm{E}\{|D[l,k]|^2\}$ is the variance of the noise spectral component. The MMSE-STSA estimator gain is given by

$$G_{\mathrm{MMSE-STSA}}[l,k] = \frac{\sqrt{\pi}}{2}\frac{\sqrt{\nu[l,k]}}{\gamma[l,k]}\exp\left(\frac{-\nu[l,k]}{2}\right) \times \left((1+\nu[l,k])I_0\left(\frac{\nu[l,k]}{2}\right) + \nu[l,k]I_1\left(\frac{\nu[l,k]}{2}\right)\right), \tag{9}$$

where $I_0(\cdot)$ and $I_1(\cdot)$ denote the modified Bessel functions of zero and first order, respectively, and $\nu[l,k]$ is given by

$$\nu[l,k] = \frac{\xi[l,k]}{\xi[l,k]+1}\gamma[l,k]. \tag{10}$$

Here, $\gamma[l,k]$ denotes the *a posteriori* SNR and is defined as

$$\gamma[l,k] = \frac{|X[l,k]|^2}{\lambda_d[l,k]}. \tag{11}$$

The WF approach gain function is given by

$$G_{\mathrm{WF}}[l,k] = \frac{\xi[l,k]}{\xi[l,k]+1}. \tag{12}$$

For the listening tests, the JND is to be found between the estimated *a priori* SNR, $\hat{\xi}[l,k]$, and a reference. The reference can be computed from the clean speech and noise in Eq. (1), as they are known completely during the listening tests. There are two reference options.

Option 1: Equation (8) computed using the instantaneous values $|S[l,k]|^2$ and $|D[l,k]|^2$ in place of $\lambda_s[l,k]$ and $\lambda_d[l,k]$, respectively, i.e., use the instantaneous *a priori* SNR as the reference.

Option 2: Compute the clean speech and noise power spectral densities (PSDs), $\lambda_s[l,k] = \mathrm{E}\{|S[l,k]|^2\}$ and $\lambda_d[l,k] = \mathrm{E}\{|D[l,k]|^2\}$, for Eq. (8) using first-order recursive smoothing, e.g., $\hat{\lambda}_s[l,k] = \alpha\hat{\lambda}_s[l-1,k] + (1-\alpha)|S[l,k]|^2$.

Through an informal listening test, it was found that using first-order recursive smoothing (with a smoothing factor ranging from 0.1 to 0.9) produces enhanced speech that exhibits significantly more speech distortion and reverberation than the enhanced speech of the instantaneous *a priori* SNR. This is due to spectral smearing caused by the first-order recursive smoothing algorithm. Objective scores for the WF using options 1 and 2 are given in Table I. It can be seen that the enhanced speech of option 1, the instantaneous *a priori* SNR, produces higher objective quality and intelligibility scores than option 2 at different α values. Hence, the instantaneous case is used for the listening tests. Additionally, the instantaneous value $|D[l,k]|^2$ can be used in the place of $\lambda_d[l,k]$ in Eq. (11), giving the instantaneous *a posteriori* SNR.

To uphold the statistical properties of the WF and the MMSE-STSA estimator, the DFT coefficients (real and imaginary parts) of the clean speech and the noise must be statistically independent Gaussian random variables with zero mean. However, finding the true probability distribution of the clean speech DFT coefficients is difficult, as speech is neither a stationary nor an ergodic process (Ephraim and Malah, 1984). This also applies to non-stationary noise sources. Measuring their probability distribution by examining long-term behavior has been suggested (Martin, 2002; Porter and Boll, 1984). However, it is argued that histograms of the DFT coefficients, obtained using a large amount of data, reflect the relative frequency rather than the true probability density of the DFT coefficients (Ephraim and Malah, 1984). Hence, we assume that the DFT coefficients of the clean speech used in this study (as described in Sec. III A) are statistically independent Gaussian random variables with zero mean and variances that are time-varying (this assumption is also made for the noise source DFT coefficients). This is identical to the assumption made by Ephraim and Malah (1984).

III. EXPERIMENT SETUP

A. Clean speech and noise recordings

Determining how the noise source impacts the JND SD level constitutes one part of this investigation. Therefore, we utilise recordings of five different noise sources to produce the noisy speech for the listening tests. Along with additive white Gaussian noise (AWGN), four real-world noise sources, including two non-stationary and two coloured, are

TABLE I. Enhanced speech objective quality and intelligibility scores (higher is better) using the WF with options 1 and 2. The perceptual evaluation of speech quality (PESQ) metric is used to obtain the objective quality scores (Rix *et al.*, 2001). The short-time objective intelligibility (STOI) metric (Taal *et al.*, 2011) is used to obtain the objective intelligibility scores (in %). The test set described in Sec. III E is used to obtain the objective scores. The objective scores are averaged over all conditions.

Reference	PESQ	STOI
Option 1	2.97	95.5
Option 2; $\alpha = 0.1$	2.80	95.4
Option 2; $\alpha = 0.5$	2.06	93.5
Option 2; $\alpha = 0.9$	1.46	86.0

included. The two real-world non-stationary noise sources include *voice babble* from the RSG-10 noise dataset (Steeneken and Geurtsen, 1988) and *street music* (recording No. 26 270) from the Urban Sound dataset (Salamon *et al.*, 2014). The two real-world coloured noise sources include *F16* and *factory* (welding) from the RSG-10 noise dataset (Steeneken and Geurtsen, 1988). These noise sources were chosen because speech enhancement methods in the literature are typically evaluated using real-world non-stationary and coloured noise sources (Nicolson and Paliwal, 2019; Nikzad *et al.*, 2020; Zhang *et al.*, 2020). The clean speech recordings from the TSP speech corpus (Kabal, 2002) are used to produce the noisy speech for the listening tests (only adult speakers are used). A total of 1378 clean speech recordings are available for the listening tests, with a minimum duration of 1.3 s and a maximum duration of 4.8 s. The clean speech and noise recordings are single-channel, with a sampling frequency of 16 kHz (recordings with a greater sampling frequency are downsampled). The noisy speech is created by mixing the selected clean speech and noise recording at a specified SNR level. The SNR levels used for the listening tests include –5, 0, 5, 10, and 15 dB.

B. Spectral distortion

Spectral distortion (SD) is a measure of *a priori* SNR estimation accuracy,

$$D = \sqrt{\frac{1}{N_d/2+1}\sum_{k=0}^{N_d/2}(\xi_{\mathrm{dB}}[l,k] - \hat{\xi}_{\mathrm{dB}}[l,k])^2}. \tag{13}$$

It is defined as the root-mean-square difference between the *a priori* SNR estimate (dB), $\hat{\xi}_{\mathrm{dB}}[l,k] = 10\log_{10}(\hat{\xi}[l,k])$, and the instantaneous *a priori* SNR (dB), $\xi_{\mathrm{dB}}[l,k] = 10\log_{10}(\xi[l,k])$, over the spectral components of the lth frame (Nicolson and Paliwal, 2019; Paliwal and Atal, 1993).

C. Stimuli generation

To find the JND SD level, the amount of distortion present in the estimated *a priori* SNR must be controlled from trial-to-trial. For a trial, the following stimuli pair is produced for a given noisy speech signal.

Stimulus 1: Enhanced speech produced by an MMSE approach utilising the instantaneous *a priori* SNR.
Stimulus 2: Enhanced speech produced by an MMSE approach utilising an *a priori* SNR estimate with an SD level of D.

The instantaneous *a priori* SNR is first computed from the clean speech and noise that form the noisy speech. The instantaneous *a priori* SNR is then used with an MMSE approach to create stimulus 1. The *a priori* SNR estimate for stimulus 2 is created by adding a random number to each component of the instantaneous *a priori* SNR (dB),

$$\hat{\xi}_{\mathrm{dB}}[l,k] = \xi_{\mathrm{dB}}[l,k] + z, \tag{14}$$

where z is a realisation of random variable Z. The distribution of Z is set so that the SD of $\hat{\xi}_{\mathrm{dB}}[l,k]$ is D.

The distribution of Z must emulate the distribution of the distortion produced by current *a priori* SNR estimators. Shown in Fig. 2 (top) are distributions of the distortion produced by the Deep Xi framework employing a residual long short-term memory (ResLSTM) network (Nicolson and Paliwal, 2019) and a residual bidirectional long short-term memory (ResBiLSTM) network (Nicolson and Paliwal, 2019; Nicolson, 2020b). The distortion is calculated by subtracting the estimated *a priori* SNR (dB) from the instantaneous *a priori* SNR (dB), $\xi_{\mathrm{dB}}[l,k] - \hat{\xi}_{\mathrm{dB}}[l,k]$. Figure 2 (bottom) shows the quantile–quantile (Q–Q) plots of the distortion of Deep Xi-ResLSTM and Deep Xi-ResBiLSTM versus a standard normal distribution. The Q–Q plots produce an approximately straight line—especially within two standard deviations of the mean (95% of the distribution)—indicating that the distortion follows a normal distribution.

Deep Xi-ResLSTM exhibits a small amount of bias ($\mu = 2.08$), while Deep Xi-ResBiLSTM exhibits even less bias ($\mu = -0.85$). This suggests that a more complex deep neural network (DNN) exhibits less bias, given that the ResBiLSTM network is more complex than the ResLSTM network. This is demonstrated by Neal *et al.* (2018), where it was found that both bias and variance decreases as the complexity of the DNN grows. With the assumption that more complex DNNs will be used to improve *a priori* SNR estimation in the future, the bias of *a priori* SNR estimators will decrease, i.e., the mean (μ) of their distortion will approach zero. Hence, we assume that Z is distributed

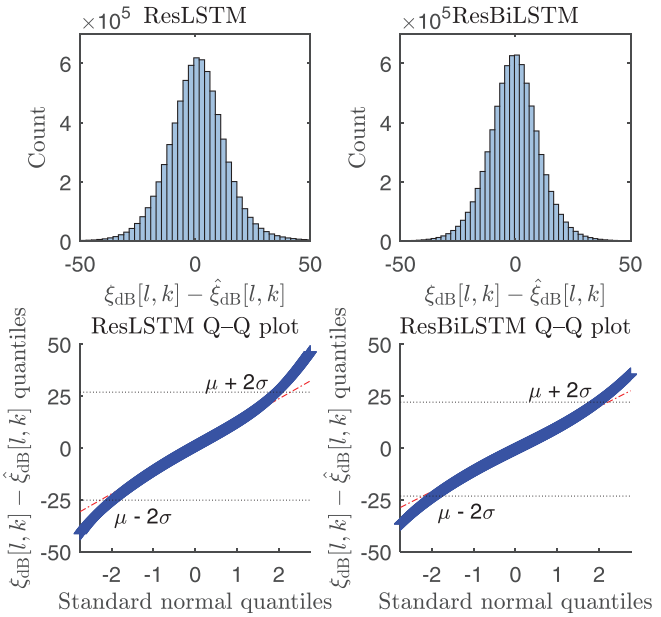


FIG. 2. (Color online) (Top) Histogram of the distortion of current *a priori* SNR estimators. (Bottom) Quantile–quantile (Q–Q) plot of the distortion versus a standard normal distribution. The mean and standard deviation of the distortion is denoted by μ and σ, respectively. The distortion is found over all spectral components of the test set described in Sec. III E.

https://doi.org/10.1121/10.0002113

normally with zero mean. To determine the variance of Z, Eq. (14) is first substituted into Eq. (13). With the knowledge that Z has zero mean, we find that the variance of Z is D^2. Thus, Z is distributed normally with zero mean and a variance of D^2: $Z \sim \mathcal{N}(0, D^2)$. The *a priori* SNR estimate with an SD level of D is artificially created using Eq. (14), which is subsequently used with an MMSE approach to create stimulus 2.

D. Listening tests

The procedure used for the JND SD level listening tests is described here. The results of the listening tests are used to determine how the SNR level, the selected MMSE approach, and the noise source impacts the JND SD level. In order to analyse how each of these factors impacts the JND SD level, listeners record JND SD levels for 30 conditions, as described in Table II (one condition per row). Each condition is determined by the employed MMSE approach, noise source, and SNR level. Results for the first five conditions in Table II enable us to determine how the SNR level impacts the JND SD level. Results for the first ten conditions in Table II enable us to determine how the selected MMSE approach impacts the JND SD level. Results for the last 25 conditions in Table II enable us to determine if the type of background noise impacts the JND SD level.

TABLE II. Conditions for each of the tests. Each session consisted of 10 tests, with 30 total conditions tested over the three sessions.

		Condition		
Session	Listeners	MMSE approach	Noise source	SNR level
1	23	WF	AWGN	−5 dB
		WF	AWGN	0 dB
		WF	AWGN	5 dB
		WF	AWGN	10 dB
		WF	AWGN	15 dB
		MMSE-STSA	AWGN	−5 dB
		MMSE-STSA	AWGN	0 dB
		MMSE-STSA	AWGN	5 dB
		MMSE-STSA	AWGN	10 dB
		MMSE-STSA	AWGN	15 dB
2	20	MMSE-STSA	*Voice babble*	−5 dB
		MMSE-STSA	*Voice babble*	0 dB
		MMSE-STSA	*Voice babble*	5 dB
		MMSE-STSA	*Voice babble*	10 dB
		MMSE-STSA	*Voice babble*	15 dB
		MMSE-STSA	*F16*	−5 dB
		MMSE-STSA	*F16*	0 dB
		MMSE-STSA	*F16*	5 dB
		MMSE-STSA	*F16*	10 dB
		MMSE-STSA	*F16*	15 dB
3	20	MMSE-STSA	*Street music*	−5 dB
		MMSE-STSA	*Street music*	0 dB
		MMSE-STSA	*Street music*	5 dB
		MMSE-STSA	*Street music*	10 dB
		MMSE-STSA	*Street music*	15 dB
		MMSE-STSA	*Factory* (welding)	−5 dB
		MMSE-STSA	*Factory* (welding)	0 dB
		MMSE-STSA	*Factory* (welding)	5 dB
		MMSE-STSA	*Factory* (welding)	10 dB
		MMSE-STSA	*Factory* (welding)	15 dB

For each of the conditions described in Table II, at least 20 listeners recorded a JND SD level. This is double the amount of listeners that participated in previous speech enhancement JND studies (Chappel *et al.*, 2016; Wójcicki and Loizou, 2012). Each listener contributes a maximum of one JND SD level per condition. Each listener records a JND SD level for each condition over three separate sessions (testing all 30 conditions in one session would cause fatigue). Each session is completed in approximately 10–20 min. Each participant is given at least a 20 min break before attempting another session. 20 listeners participated in all three sessions (13 male and seven female, aged between 18 and 41), where 10 of the listeners had prior music/signal processing experience. Three listeners participated in only the first session (one male and two female, aged between 18 and 35). Each listener possessed normal hearing. Each session is conducted in a quiet room using closed circumaural headphones (Sennheiser HD280 PRO) at a comfortable listening level. Before starting the first session, each listener participates in a practice test, to familiarise themselves and to adjust the volume to a comfortable level. The authors of this study did not participate in the listening tests.

Each listener completes ten tests during a session, one test for each condition. The order of the conditions for a session is randomised for each listener. The noisy speech used for a test is created on the fly by mixing a random section of a recording of the condition's noise source with a randomly selected clean speech recording, at the condition's SNR level. For each test in a session, a listener completes multiple trials. For each trial, a stimuli pair is presented to the listener. The stimuli pair, as described previously, includes: Stimulus 1, the enhanced speech produced by an MMSE approach utilising the instantaneous *a priori* SNR and stimulus 2, the enhanced speech produced by an MMSE approach utilising an *a priori* SNR estimate with an SD level of D. During a test, the condition and the noisy speech remain the same from trial-to-trial, only D of stimulus 2 changes.

Stimulus 1 is played first as a reference to the listener, followed by 200 ms of silence, and then stimulus 2. The entirety of each stimulus was played to the listener. The duration of each stimulus is between 1.3 and 4.8 s. The SD level, D, is changed adaptively from trial-to-trial, in order to find the SD level that is noticeable by the listener 50% of the time (the JND SD level) (Booth and Freeman, 1993). Once the stimuli pair for a trial has been presented, the listener selects one of three options. The first option is selected if the stimuli pair has no perceivable difference in speech quality. The second option is selected if there is a perceivable difference in speech quality. The third option allows the listener to re-listen to the stimuli pair for the trial. The next trial is presented to the listener if one of the first two options is selected.

An example of how the SD level, D, is changed adaptively from trial-to-trial is shown in Fig. 3. The up-down method (Levitt, 1971) is used to adaptively control the SD level from trial-to-trial, and has been used in many psychoacoustical studies (Buck *et al.*, 2012; Wójcicki and Loizou, 2012). The initial value for the up-down method is found by using the ascending method of limits (Levitt, 1971). After the initial value for the up-down method is found, six total runs are used for each test, ensuring that at least six reversals are completed, following (Wetherill and Levitt, 1965). An SD level step size of 0.5 dB is used for the ascending method of limits and the up-down method. The midpoint of every second run is used as the mid-run estimate of the JND SD level (the initial run for the ascending method of limits is excluded, i.e., the midpoint of runs 2, 4, and 6, as shown in Fig. 3, are used as the mid-run estimates) (Levitt, 1971). This gives a total of three mid-run estimates for each test, which are averaged to give the listener's JND SD level for the test/condition.

E. *A priori* SNR estimator test set

The test set from Nicolson and Paliwal (2019) is used here to evaluate *a priori* SNR estimators in the literature, which we refer to as the *a priori* SNR estimator test set henceforth [available online (Nicolson, 2020a)]. The noisy speech for the *a priori* SNR estimator test set was created using a subset of the clean speech and noise recordings described in Sec. III A. Four of the five noise sources from Sec. III A were used, specifically, *voice babble*, *F16*, *street music*, and *factory* (welding). Ten clean speech recordings were randomly selected without replacement from the TSP speech corpus (Kabal, 2002) (only adult speakers were used) for each of the four noise recordings. To generate the noisy speech, a random section of the noise recording was mixed with the clean speech at five SNR levels: $\{-5, 0, 5, 10, 15\}$ dB. This created a test set of 200 noisy speech signals. The noisy speech was single channel, with a sampling frequency of 16 kHz. The *a priori* SNR estimator test set is not used for the listening tests, rather, it is used to compare the performance of *a priori* SNR estimators in the literature to that of the JND SD levels.

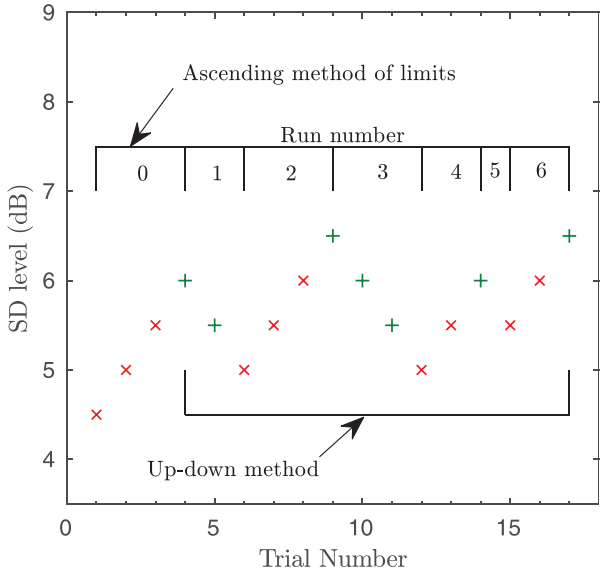


FIG. 3. (Color online) A red cross indicates that a difference is not perceived between the stimuli pair for a trial. A green plus indicates that a difference is perceived between the stimuli pair for a trial. The ascending method of limits is used for the initial run, to find the initial SD level for the up-down method. The up-down method then controls the SD level for six runs. Mid-run estimates are found from the midpoint of runs 2, 4, and 6.

IV. RESULTS AND DISCUSSION

The JND SD levels ascertained from the listening tests are shown in Fig. 4. For each of the conditions described in Table II, at least 20 JND SD levels are recorded. Each listener contributes a maximum of one JND SD level per condition. 20 listeners participated in all three sessions and three listeners participated in only the first session. In this section, we perform a statistical analysis to determine how the SNR level, MMSE approach, and noise source impacts the JND SD level—before presenting the final JND SD levels in Sec. IV F. Each of the following subsections is summarised as follows.

Section IV A: determines if the JND SD level is impacted by the SNR level.

Section IV B: determines if the JND SD level is impacted by the selected MMSE approach.

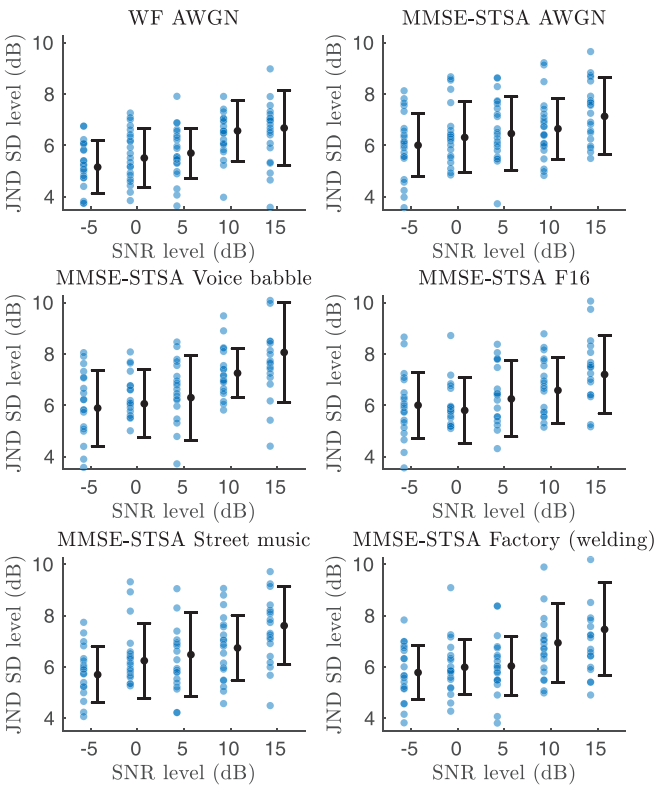


FIG. 4. (Color online) JND SD levels attained for each condition. The black dot and the error bar indicate the mean and standard deviation, respectively, of the JND SD level for the corresponding condition. Each condition comprises of a noise source, an SNR level, and an MMSE approach. The conditions are described in Table II.

https://doi.org/10.1121/10.0002113

Section IV C: determines if the JND SD level is impacted by the noise source.

Section IV D: indicates why the JND SD level is impacted by the SNR level.

Section IV E: indicates why the JND SD level is impacted by the selected MMSE approach.

Section IV F: determines the final JND SD level.

Section IV G: presents a comparison of the final JND SD level to the SD levels of current *a priori* SNR estimators.

A. How the SNR level impacts the JND SD level

Presented in Table III are the statistics of the JND SD level at each SNR level. The statistics are computed over the JND SD levels for the WF and for AWGN. The dispersion at each SNR level is moderately consistent (Bartlett's test, $p \gg 0.05$), with a standard deviation (s) ranging from 0.98 to 1.46 and an interquartile range (IQR) ranging from 0.99 to 1.69. However, the mean at each SNR level is significantly different [one-way analysis of variance (ANOVA), $p \leq 0.05$]. Despite the dispersion at each SNR level, there exists a moderate positive relationship between the SNR level and the JND SD level (Pearson correlation coefficient, $r = 0.38$). In Sec. IV D, we indicate why the JND SD level increases with the SNR level.

B. How the MMSE approach impacts the JND SD level

The distribution of the JND SD level for each MMSE approach at each SNR level is shown in Fig. 5. Only the JND SD levels for AWGN are considered. As shown in Table IV, there is a significant difference between the mean JND SD level of the MMSE-STSA estimator and the WF at an SNR level of −5, 0, and 5 dB (two-sample t-test, $p \leq 0.05$ for SNR levels −5, 0, and 5 dB). However, there is no significant difference between the mean JND SD level of the MMSE estimator and the WF at an SNR level of 10 and 15 dB (two-sample t-test, $p > 0.05$ for SNR levels 10 and 15 dB). This indicates that the choice of MMSE approach impacts the JND SD level at an SNR level of 5 dB or less. In Sec. IV E, we indicate why there is a significant difference for each MMSE approach at an SNR level of 5 dB or lower.

TABLE III. Statistics of the JND SD level as the SNR level increases. The statistics for each SNR level are computed over the JND SD levels for the WF and for AWGN. The mean ($\bar{x}$), the 95% confidence interval (CI), the standard deviation (s), the interquartile range (IQR), and the sample size (N) are the given statistics.

	SNR level				
Statistic	−5	0	5	10	15
$\bar{x}$	5.15	5.51	5.70	6.57	6.68
95% CI	±0.42	±0.48	±0.40	±0.48	±0.61
s	1.03	1.17	0.98	1.18	1.46
IQR	1.26	1.69	1.05	0.99	1.23
N	23	23	23	23	23

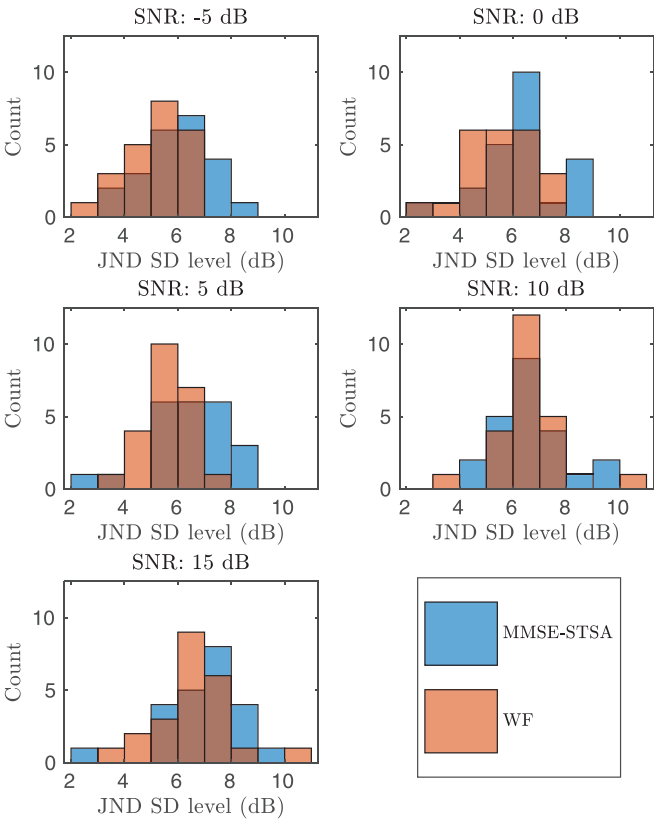


FIG. 5. (Color online) Histograms of the JND SD levels for the MMSE-STSA estimator and the WF at each tested SNR level. Only the JND SD levels for AWGN are considered. Each histogram has a sample size (N) of 23.

C. How the noise source impacts the JND SD level

For this subsection, only JND SD levels for the MMSE-STSA estimator and from listeners that completed all three sessions are included. This provides a balanced sample size for each of the noise sources. The spread of the JND SD level for each of the noise sources is shown in Fig. 6. It can be seen that the median JND SD level is similar for each of the noise sources. The result of a one-way ANOVA test between the JND SD levels of each noise source is given in Table V. It can be observed that there is no significant difference between the mean JND SD level of the noise sources at each SNR level (one-way ANOVA, $p > 0.05$ for each SNR level). Moreover, there is no significant difference between the dispersion of the JND SD level of the noise sources at each SNR level (Barlett's test, $p > 0.05$ for each

TABLE IV. Tests to determine if there is a significant difference between the JND SD levels of the MMSE-STSA estimator and WF at each SNR level. Only the JND SD levels for AWGN are considered. The p-value for the two-sample t-test and the number of samples (N) are the given statistics.

	SNR level				
Statistic	−5	0	5	10	15
Two-sample t-test, p	0.01	0.04	0.04	0.81	0.30
N	23	23	23	23	23

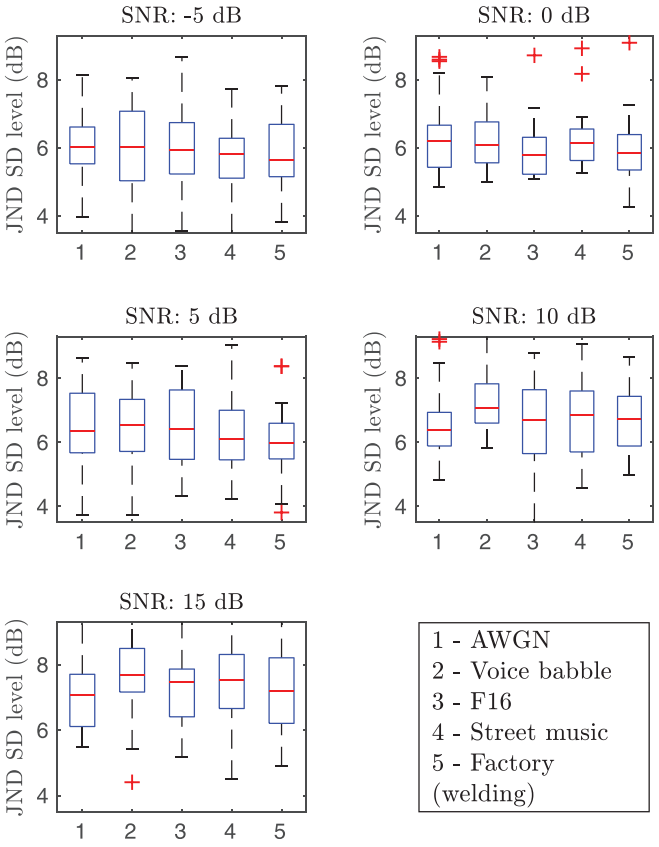


FIG. 6. (Color online) Boxplots of the JND SD level for each noise source. Only JND SD levels for the MMSE-STSA estimator and from listeners that completed all three sessions are included. A sample size (N) of 20 is used for each boxplot. Each subplot corresponds to a different SNR level. The central red mark indicates the median, and the bottom and top edges of the blue box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted individually using the red "+" symbol.

SNR level). This indicates that the noise source has no impact on the JND SD level, at least for the five tested.

D. Why the JND SD level is impacted by the SNR level

As shown in Sec. IV A, the JND SD level increases with the SNR level. To explain this occurrence, we evaluate the gain functions of the MMSE-STSA estimator and the WF. The gain as a function of the *a priori* SNR is shown in Fig. 7 (top). It can be observed that the gradient of the gain

TABLE V. Statistics of the JND SD level for the noise sources at each SNR level. The statistics are computed over the JND SD levels for the MMSE-STSA estimator and for the listeners that completed all three sessions. The p-value for a one-way ANOVA test, the p-value for a Bartlett's test, and the number of samples (N) are the given statistics.

	SNR level				
Statistic	−5	0	5	10	15
One-way ANOVA, p	0.90	0.81	0.91	0.40	0.29
Bartlett's, p	0.57	0.67	0.58	0.39	0.71
N	20	20	20	20	20

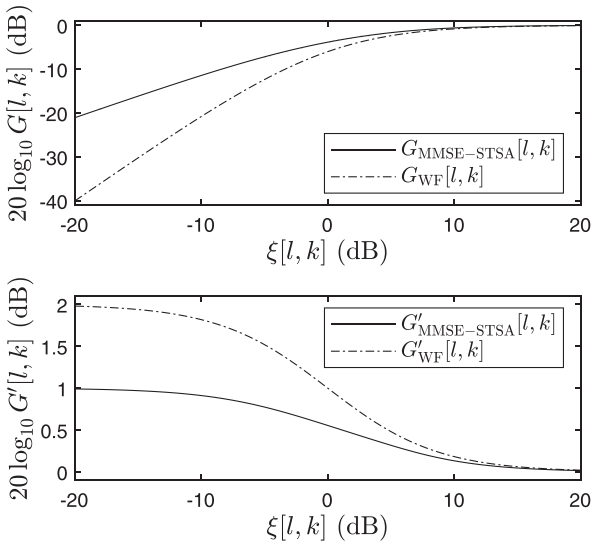


FIG. 7. (Top) Gain as a function of the *a priori* SNR for the MMSE-STSA estimator and the WF. (Bottom) Derivative of the gain as a function of the *a priori* SNR for the MMSE-STSA estimator and the WF. The *a posteriori* SNR is computed using its maximum likelihood (ML) estimate, $\hat{\gamma}[l,k] = \xi[l,k] + 1$, from Nicolson and Paliwal (2019).

decreases as the *a priori* SNR increases, as shown in Fig. 7 (bottom). The lower the gradient, the lower the resultant gain distortion that a set SD level will cause. Therefore, a set SD level applied to a greater instantaneous *a priori* SNR will result in a lower gain distortion. The amount of gain distortion adversely impacts the quality of the resultant enhanced speech. This indicates that the JND SD level is the SD level that causes sufficient gain distortion to produce a JND. The gradient of the gain indicates that the JND SD level will increase with the instantaneous *a priori* SNR. This is because a greater SD level is required to produce sufficient gain distortion to hear a JND at greater instantaneous *a priori* SNRs.

Further insights can be obtained by examining the distribution of the instantaneous *a priori* SNR as the SNR level increases. The distribution (in dB) as the SNR level increases is presented in Figs. 8(a)–8(e). As shown in Fig. 8(f), the mean instantaneous *a priori* SNR increases with the SNR level, while the standard deviation remains unchanged. This causes the amount of gain distortion for a set SD level to decrease on average as the SNR level increases. This indicates that the JND SD level increases with the SNR level, which is consistent with the results in Table III. The JND SD level is thus directly impacted by the gradient of the gain function. In summary, the JND SD level increases with the SNR level because (1) a set SD level applied to a greater instantaneous *a priori* SNR will cause a lower gain distortion and (2) the mean instantaneous *a priori* SNR increases with the SNR level.

E. Why the JND SD level is impacted by the MMSE approach

Observing the gradient of the gain functions in Fig. 7 (bottom) provides insight as to why there is a significant

https://doi.org/10.1121/10.0002113

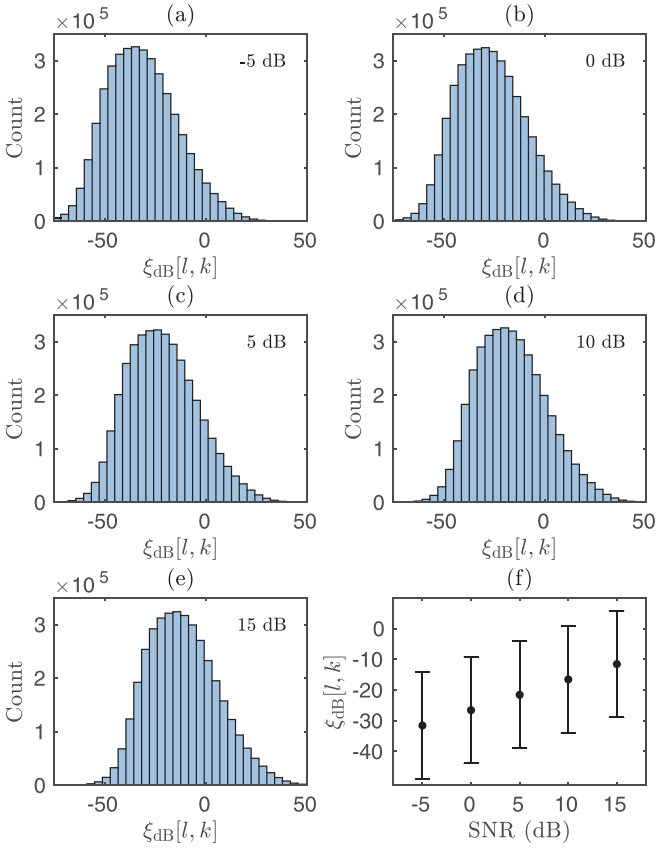


FIG. 8. (Color online) (a)–(e) Distribution of the instantaneous *a priori* SNR (dB) as the SNR level increases. The distribution is found over 100 randomly selected clean speech recordings described in Sec. III A. Each clean speech recording is corrupted with AWGN at five different SNR levels: $\{-5, 0, 5, 10, 15\}$ dB. (f) Mean and standard deviation of the instantaneous *a priori* SNR (dB) as the SNR level increases.

difference in the JND SD levels of the MMS-STSA estimator and the WF at an SNR level of 5 dB or less. As established in Sec. IV D, the gradient of the gain function has an impact on the JND SD level as the SNR level changes. At lower SNR levels, where the mean instantaneous *a priori* SNR is also lower [Fig. 8(f)], the gradients of both gain functions are different, causing a significant difference between their JND SD levels. At greater SNR levels, where the mean instantaneous *a priori* SNR is also greater, the gradients of both gain functions are similar, causing a smaller difference between their JND SD levels. This gives reason as to why the JND SD levels for the WF are significantly lower at SNR levels of 5 dB or less.

F. Final JND SD levels

As determined in Secs. IV A and IV B, the SNR level and the selected MMSE approach has a significant impact on the JND SD level. In this subsection, we present the final JND SD levels, as given by the mean MMSE-STSA estimator and WF JND SD levels at each of the tested SNR levels. The analysis in Sec. IV C indicates that the noise source has no significant impact on the JND SD level. Thus, the final JND SD levels are independent of the noise source. The mean MMSE-STSA estimator JND SD level at each tested SNR level is presented in Table VI. These are the SD levels that an *a priori* SNR estimator must exceed to have no impact on the perceived quality of MMSE-STSA estimator enhanced speech. The statistics are computed over the JND SD levels of all the noise sources. This provides a larger sample size to compute the statistics for the MMSE-STSA estimator.

The mean MMSE-STSA estimator JND SD level is significantly different at each SNR level (one-way ANOVA, $p \leq 0.05$). This is due to a moderate positive relationship between the mean MMSE-STSA estimator JND SD level and the SNR level (Pearson correlation coefficient, $r = 0.37$). The dispersion of the mean MMSE-STSA estimator JND SD level at each SNR level is significantly different (Bartlett's test, $p \leq 0.05$), with the dispersion tending to increase with the SNR level [see standard deviation (s) in Table VI]. This is consistent with reports from listeners that it was more difficult to track the JND SD level at higher SNR levels. The mean WF JND SD level at each tested SNR level is presented in Table III. These are the SD levels that an *a priori* SNR estimator must exceed to have no impact on the perceived quality of WF enhanced speech. As the WF is only tested with AWGN, its statistics are computed from a smaller sample size than that of the MMSE-STSA estimator. This impacts the 95% confidence interval (CI) of the mean WF JND SD levels. As can be observed, the CI for the WF is greater (Table III) than that of the MMSE-STSA estimator (Table VI) for each SNR level.

G. SD levels of *a priori* SNR estimators

Here, we compare the SD levels of current and previous *a priori* SNR estimators found in the work by Nicolson and Paliwal (2019) (the used test set is described in Sec. III E). The previous *a priori* SNR estimators include the DD approach (Ephraim and Malah, 1984), the two-step noise reduction (TSNR) technique (Plapous *et al.*, 2004), HRNR (Plapous *et al.*, 2005), and selective cepstro-temporal smoothing (SCTS) (Breithaupt *et al.*, 2008). Each uses the MMSE noise power spectral density (PSD) estimator by Gerkmann and Hendriks (2012). The current estimators

TABLE VI. Statistics of the JND SD level for the MMSE-STSA estimator as the SNR level increases. The statistics are computed over all noise sources at each SNR level. This means that multiple JND SD levels for each participant are used at each SNR level. The mean ($\bar{x}$), the 95% confidence interval (CI), the standard deviation (s), the interquartile range (IQR), and the sample size (N) are the given statistics.

	SNR level				
Statistic	−5	0	5	10	15
$\bar{x}$	5.88	6.09	6.31	6.83	7.48
95% CI	±0.24	±0.25	±0.28	±0.24	±0.32
s	1.22	1.30	1.46	1.25	1.67
IQR	1.51	1.09	1.70	1.58	1.86
N	103	103	103	103	103

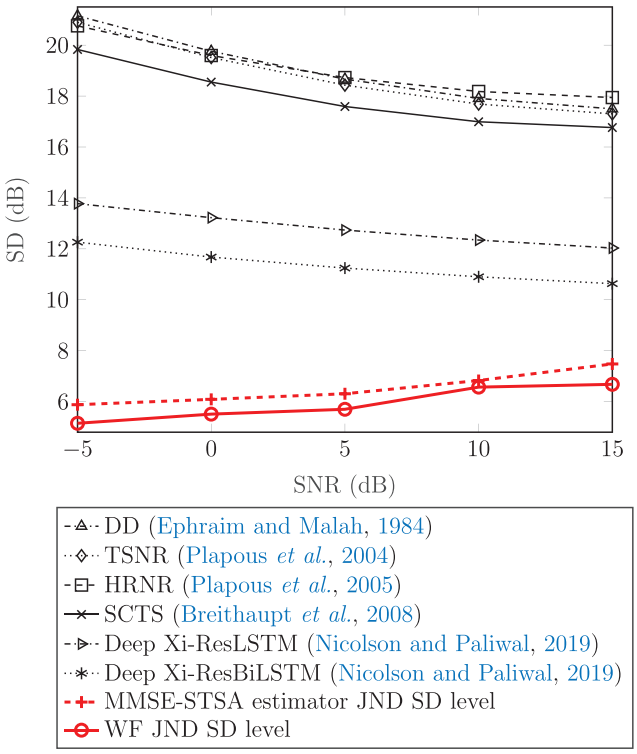


FIG. 9. (Color online) SD levels attained by each of the *a priori* SNR estimators in Nicolson and Paliwal (2019). The red plots are the JND SD levels for the MMSE-STSA and the WF. The test set described in Sec. III E is used to obtain the SD levels. SD levels are averaged over all noise sources.

include Deep Xi-ResLSTM and Deep Xi-ResBiLSTM. The SD levels are averaged over all noise sources [*voice babble*, *F16*, *street music*, and *factory* (welding)], and compared to the JND SD levels of the MMSE-STSA estimator and the WF, as shown in Fig. 9. It can be seen that the *a priori* SNR estimator with the highest accuracy (Deep Xi-ResBiLSTM) produces an SD level that is substantially greater than the JND SD level of the MMSE-STSA estimator and the WF at each SNR level. A significant improvement in *a priori* SNR estimation accuracy is thus required to surpass the JND SD levels of the MMSE-STSA estimator and the WF at each SNR level.

Objective quality and intelligibility scores for the MMSE-STSA estimator and the WF using different *a priori* SNRs are given in Table VII. It can be observed that the objective scores decrease when the instantaneous *a priori* SNR is corrupted to the JND SD level. Although the objective scores produced by Deep Xi-ResBiLSTM outperform noisy speech, they are significantly worse than that of the instantaneous case corrupted to the JND SD level. This is consistent with the SD levels presented in Fig. 9. The objective scores for the instantaneous *a priori* SNR corrupted to the JND SD level indicate the speech enhancement performance that is attainable by an *a priori* SNR estimator that is capable of exceeding the JND SD level.

H. Future recommendations

In this study, the JND SD level for the MMSE-STSA estimator and the WF are presented. However, there are other commonly used MMSE approaches in the literature, including the MMSE log-spectral amplitude (MMSE-LSA) estimator (Ephraim and Malah, 1985) and the square-root WF (SRWF) (Lim and Oppenheim, 1979). As shown in Sec. IV B, the selected MMSE approach has a significant impact on the JND SD level. Therefore, the JND SD levels for other MMSE approaches will need to be investigated. Moreover, only five noise sources were considered in this study. To claim concretely that the noise source has no impact on the JND SD level would require a larger set of noise sources.

V. CONCLUSION

In this study, the *a priori* SNR estimate SD level that results in a JND in the perceived quality of MMSE approach enhanced speech is found through a series of listening tests. An *a priori* SNR estimator will have no impact on the

TABLE VII. Enhanced speech objective quality and intelligibility scores (higher is better) for the MMSE-STSA estimator and the WF. Scores are given for the estimated *a priori* SNR from Deep Xi-ResBiLSTM, the instantaneous *a priori* SNR corrupted to the JND SD level, and the instantaneous *a priori* SNR. The mean opinion score of the listening quality objective (MOS-LQO) is used as the objective quality metric, where the wideband perceptual evaluation of quality (Wideband PESQ) is the objective model used to obtain the MOS-LQO score (Morioka *et al.*, 2005). The short-time objective intelligibility (STOI) metric (Taal *et al.*, 2011) is used to obtain the objective intelligibility scores (in %). The test set described in Sec. III E is used to obtain the objective scores. The objective scores are averaged over all noise sources. The *a posteriori* SNR for Deep Xi-ResBiLSTM is computed using the maximum likelihood (ML) estimate $\hat{\gamma}[l,k] = \xi[l,k] + 1$, as in Nicolson and Paliwal (2019). The instantaneous *a posteriori* SNR is used with the instantaneous *a priori* SNR corrupted to the JND SD level, and the instantaneous *a priori* SNR.

		SNR level									
		MOS-LQO					STOI				
Method	*A priori* SNR	−5	0	5	10	15	−5	0	5	10	15
Noisy speech	—	1.04	1.06	1.12	1.28	1.59	59.3	71.3	82.0	90.2	95.3
WF	Deep Xi-ResBiLSTM	1.17	1.35	1.65	2.14	2.68	67.4	81.7	90.4	95.1	97.5
WF	JND SD level	2.04	2.44	2.85	3.31	3.65	89.4	92.4	95.1	97.0	98.4
WF	Instantaneous	2.16	2.53	2.96	3.38	3.74	90.9	93.7	96.0	97.7	98.7
MMSE-STSA	Deep Xi-ResBiLSTM	1.19	1.42	1.74	2.18	2.69	67.4	81.0	89.7	94.7	97.3
MMSE-STSA	JND SD level	2.14	2.43	2.88	3.24	3.63	89.9	92.6	94.9	96.8	98.1
MMSE-STSA	Instantaneous	2.58	2.89	3.24	3.61	3.89	93.1	94.8	96.4	97.8	98.7

perceived quality of MMSE approach enhanced speech if it is able to attain SD levels lower than that of the JND SD level. Thus, the JND SD level is a target level of accuracy for *a priori* SNR estimation research. A statistical analysis indicates that the SNR level, along with the selected MMSE approach, has a significant impact on the JND SD level. The JND SD level increases with the SNR level and the JND SD level of the MMSE-STSA estimator is higher than that of the WF at each SNR level. Moreover, there was no statistically significant difference between the JND SD levels of the five tested background noise sources at each SNR level. Following the literature, it is determined that a significant improvement in *a priori* SNR estimation accuracy is required to reach the JND SD level.

Agus, N., Anderson, H., Chen, J.-M., Lui, S., and Herremans, D. (**2018**). "Perceptual evaluation of measures of spectral variance," J. Acoust. Soc. Am. **143**(6), 3300–3311.
Alkahtani, F. (**2019**). "Acoustic manifestations of 'narrow focus' in Apurímac Quechua vowels," J. Acoust. Soc. Am. **146**(4), 3008.
Allen, J. (**1977**). "Short term spectral analysis, synthesis, and modification by discrete Fourier transform," IEEE Trans. Acoust. Speech Sign. Process. **25**(3), 235–238.
Allen, J. B., and Rabiner, L. R. (**1977**). "A unified approach to short-time Fourier analysis and synthesis," Proc. IEEE **65**(11), 1558–1564.
Booth, D., and Freeman, R. (**1993**). "Discriminative feature integration by individuals," Acta Psychol. **84**(1), 1–16.
Boucher, M. A., Rychtarikova, M., Zelem, L., Pluymers, B., and Desmet, W. (**2019**). "Reverberation time and audibility in phased geometrical acoustics using plane or spherical wave reflection coefficients," J. Acoust. Soc. Am. **145**(4), 2681–2690.
Breithaupt, C., Gerkmann, T., and Martin, R. (**2008**). "A novel *a priori* SNR estimation approach based on selective cepstro-temporal smoothing," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4897–4900.
Buck, A., Blevins, M. G., Wang, L. M., and Peng, Z. (**2012**). "Measurements of the just noticeable difference for reverberation time using a transformed up–down adaptive method," J. Acoust. Soc. Am. **132**(3), 2060.
Chappel, R., Schwerin, B., and Paliwal, K. (**2016**). "Phase distortion resulting in a just noticeable difference in the perceived quality of speech," Speech Commun. **81**, 138–147.
Crochiere, R. (**1980**). "A weighted overlap-add method of short-time Fourier analysis/synthesis," IEEE Trans. Acoust. Speech Sign. Process. **28**(1), 99–102.
Ephraim, Y., and Malah, D. (**1984**). "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," IEEE Trans. Acoust. Speech Sign. Process. **32**(6), 1109–1121.
Ephraim, Y., and Malah, D. (**1985**). "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," IEEE Trans. Acoust. Speech Sign. Process. **33**(2), 443–445.
Gerkmann, T., and Hendriks, R. C. (**2012**). "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," IEEE Trans. Audio Speech Lang. Process. **20**(4), 1383–1393.
Griffin, D., and Jae Lim (**1984**). "Signal estimation from modified short-time Fourier transform," IEEE Trans. Acoust. Speech Sign. Process. **32**(2), 236–243.
Kabal, P. (**2002**). "TSP speech database," technical report.
Levitt, H. (**1971**). "Transformed up–down methods in psychoacoustics," J. Acoust. Soc. Am. **49**(2B), 467–477.
Lim, J. S., and Oppenheim, A. V. (**1979**). "Enhancement and bandwidth compression of noisy speech," Proc. IEEE **67**(12), 1586–1604.
Loizou, P. C. (**2013**). *Speech Enhancement: Theory and Practice*, 2nd ed. (CRC Press, Boca Raton, FL, USA).
Martin, R. (**2002**). "Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. I-253–I-256.
McAulay, R., and Malpass, M. (**1980**). "Speech enhancement using a soft-decision noise suppression filter," IEEE Trans. Acoust. Speech Sign. Process. **28**(2), 137–145.
Morioka, C., Kurashima, A., and Takahashi, A. (**2005**). "Proposal on objective speech quality assessment for wideband IP telephony," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005 (ICASSP '05)*, Vol. 1, pp. 49–52.
Nadiroh, A., and Arifianto, D. (**2018**). "Just noticeable difference of masker to enhance privacy in an open-plan office," J. Acoust. Soc. Am. **144**(3), 1661.
Neal, B., Mittal, S., Baratin, A., Tantia, V., Scicluna, M., Lacoste-Julien, S., and Mitliagkas, I. (**2018**). "A modern take on the bias-variance tradeoff in neural networks," arXiv:1810.08591 [cs.LG].
Nicolson, A. (**2020a**). "Test set from 10.1016/j.specom.2019.06.002," IEEE Dataport, http://dx.doi.org/10.21227/0ppr-yy46.
Nicolson, A. (**2020b**). "Deep Xi: A deep learning approach to *a priori* SNR estimation," https://github.com/anicolson/DeepXi.
Nicolson, A., and Paliwal, K. K. (**2019**). "Deep learning for minimum mean-square error approaches to speech enhancement," Speech Commun. **111**, 44–55.
Nikzad, M., Nicolson, A., Gao, Y., Zhou, J., Paliwal, K. K., and Shang, F. (**2020**). "Deep residual-dense lattice network for speech enhancement," in *AAAI Conference on Artificial Intelligence*.
Paliwal, K. K., and Atal, B. S. (**1993**). "Efficient vector quantization of LPC parameters at 24 bits/frame," IEEE Trans. Speech Audio Process. **1**(1), 3–14.
Plapous, C., Marro, C., Mauuary, L., and Scalart, P. (**2004**). "A two-step noise reduction technique," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 289–292.
Plapous, C., Marro, C., and Scalart, P. (**2005**). "Speech enhancement using harmonic regeneration," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005 (ICASSP '05)*, Vol. 1, pp. 157–160.
Porter, J., and Boll, S. (**1984**). "Optimal estimators for spectral restoration of noisy speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP '84*, Vol. 9, pp. 53–56.
Rix, A. W., Beerends, J. G., Hollier, M. P., and Hekstra, A. P. (**2001**). "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing* (Cat. No. 01CH37221), Vol. 2, pp. 749–752.
Roy, S. K., Nicolson, A., and Paliwal, K. K. (**2020a**). "A deep learning-based Kalman filter for speech enhancement," in *Proceedings of Interspeech 2020*.
Roy, S. K., Nicolson, A., and Paliwal, K. K. (**2020b**). "Deep learning with augmented Kalman filter for single-channel speech enhancement," in *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*.
Salamon, J., Jacoby, C., and Bello, J. P. (**2014**). "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM International Conference on Multimedia*, MM'14, Association for Computing Machinery, New York, pp. 1041–1044.
Steeneken, H. J., and Geurtsen, F. W. (**1988**). "Description of the RSG-10 noise database," Report No. IZF 1988-3, TNO Institute for Perception, Soesterberg, the Netherlands.
Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (**2011**). "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," IEEE Trans. Audio Speech Lang. Process. **19**(7), 2125–2136.
Vary, P., and Martin, R. (**2006**). *Digital Speech Transmission: Enhancement, Coding and Error Concealment* (Wiley, Hoboken, NJ, USA).
Wetherill, G. B., and Levitt, H. (**1965**). "Sequential estimation of points on a psychometric function," Br. J. Math. Stat. Psychol. **18**(1), 1–10.
Wójcicki, K. K., and Loizou, P. C. (**2012**). "Channel selection in the modulation domain for improved speech intelligibility in noise," J. Acoust. Soc. Am. **131**(4), 2904–2913.
Zhang, Q., Nicolson, A., Wang, M., Paliwal, K. K., and Wang, C. (**2020**). "DeepMMSE: A deep learning approach to MMSE-based noise power spectral density estimation," IEEE/ACM Trans. Audio Speech Lang. Process. **28**, 1404–1415.