Speech Communication 116 (2020) 12–29

Contents lists available at ScienceDirect

Speech Communication

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

Speech Enhancement Using a Risk Estimation Approach

Jishnu Sadasivan*, Chandra Sekhar Seelamantula, Nagarjuna Reddy Muraka

Department of Electrical Engineering, Indian Institute of Science, Bangalore 560012, India

ARTICLE INFO

Keywords:
Speech enhancement
Unbiased risk estimation
Stein's lemma
Perceptual distortion measure
Objective and subjective assessment

ABSTRACT

The goal in speech enhancement is to obtain an estimate of clean speech starting from the noisy signal by minimizing a chosen distortion measure (risk). Often, this results in an estimate that depends on the unknown clean signal or its statistics. Since access to such priors is limited or impractical, one has to rely on an estimate of the clean signal statistics. In this paper, we develop a risk estimation framework for speech enhancement, in which one optimizes an unbiased estimate of the risk instead of the actual risk. The estimated risk is expressed solely as a function of the noisy observations and the noise statistics. Hence, the corresponding denoiser does not require the clean speech prior. We consider several speech-specific perceptually relevant distortion measures and develop corresponding unbiased estimates. Minimizing the risk estimates gives rise to denoisers, which are nonlinear functions of the a posteriori SNR. Listening tests show that, within the risk estimation framework, Itakura-Saito and weighted hyperbolic cosine distortions are superior than the other measures. Comparisons in terms of perceptual evaluation of speech quality (PESQ), segmental SNR (SSNR), source-to-distortion ratio (SDR), and short-time objective intelligibility (STOI) also indicate a superior performance for these two distortion measures. For SNRs greater than 5 dB, the proposed approach results in better denoising performance — both in terms of objective and subjective assessment — than techniques based on the Wiener filter, log-MSE minimization, and Bayesian nonnegative matrix factorization.

1. Introduction

The goal in speech enhancement is to suppress noise and enhance signal intelligibility and quality. Over the past few decades, several techniques have been developed for noise suppression. The challenges are nonstationarity of the speech signal, distribution of noise, type of noise distortion, noise being signal-dependent or independent, etc. The problem continues to be of significant interest to the speech community particularly considering the enormous increase in the number of smartphone users. An early review of various noise reduction techniques was given by Lim and Oppenheim (1979). Loizou's book on speech enhancement (Loizou, 2007) is a comprehensive reference on the topic. Next, we briefly review speech denoising algorithms before proceeding with the development of the risk estimation framework for speech enhancement.

1.1. Related Literature

Speech enhancement algorithms can be broadly classified as follows: (i) Spectral subtraction algorithms, (ii) Wiener filtering techniques, (iii) Subspace methods, (iv) Statistical model based methods; and (v) Neural network approaches.

1.1.1. Spectral subtraction algorithms

Boll (1979) and Weiss et al. (1974) proposed to subtract an estimate of the noise power spectrum from the noisy signal spectrum, in order to estimate the clean signal spectrum. The fundamental assumption is that the noise is additive and stationary. The noisy signal phase is used in reconstructing the time-domain signal. Weiss et al. (1974) also proposed subtraction techniques in autocorrelation and cepstral domains. Lockwood and Boudy (1992), Kamath and Loizou (2002), Zhang and Zhao (2013) proposed improved versions of spectral subtraction algorithms.

1.1.2. Wiener filtering techniques

These are linear estimators based on the minimum mean-squared error (MMSE) criterion (Loizou, 2007), in which one constructs the Wiener filter using an estimate of the clean and noisy speech power spectra. Lim and Oppenheim (1979) proposed a parametric Wiener filter, which allows for controlling the trade-off between the signal distortion and residual noise. Hu and Loizou (2003, 2004) integrated psychoacoustic constraints into this framework. In Hu and Loizou (2003), they use a perceptual weighting filter to shape the residual noise to make it inaudible. In Hu and Loizou (2004), they constrain the noise spectrum to lie below a preset threshold at each frequency. Chen et al. (2006) quantified

* Corresponding author.
E-mail addresses: sadasivan@iisc.ac.in (J. Sadasivan), css@iisc.ac.in (C.S. Seelamantula), nagarjunareddy100@gmail.com (N.R. Muraka).

https://doi.org/10.1016/j.specom.2019.11.001
Received 21 March 2019; Received in revised form 11 October 2019; Accepted 3 November 2019
Available online 6 November 2019

the amount of noise reduction and analyzed its relation to speech distortion. The Wiener filter requires an estimate of the a priori signal-to-noise ratio (SNR). Scalart and Filho (1996) used a recursive a priori SNR estimator whereas Lim and Oppenheim (1978) iteratively estimated the Wiener filter based on autoregressive modeling of the speech signal. Hansen and Clements (1991) imposed inter- and intra-frame constraints to ensure speech-like characteristics within each iteration. Sreenivas and Kirnapure (1996) proposed a codebook constrained iterative Wiener filter which has a better convergence behavior than the filter proposed in Lim and Oppenheim (1978). Srinivasan et al. (2006, 2007) proposed maximum-likelihood and Bayesian methods for estimating the speech and noise power spectra. Rosenkranz and Puder (2012) proposed adaptation techniques to improve the performance of the codebook approaches reported in Srinivasan et al. (2006, 2007) against model mismatches and unknown noise types. Xia and Bao (2014) utilized a weighted denoising auto-encoder to estimate the clean speech power spectrum from the noisy speech. Deng and Bao (2016) proposed an expectation-maximization (EM) algorithm to estimate the clean speech and noise power spectra.

1.1.3. Subspace techniques

Originally proposed by Ephraim and Van Trees, subspace techniques rely on eigenvalue decomposition of the data covariance matrix (Ephraim and Trees, 1995; Mittal and Phamdo, 2000; Huang and Zhao, 1998; Rezayee and Gazor, 2001) or singular-value decomposition of the data matrix (Dendrinos et al., 1991; Hansen et al., 1999; 1995). The noise eigenvalues/singular values are smaller than those of the noisy signal and denoising happens when the signal is reconstructed from the eigen/singular vectors corresponding to the signal subspace alone. Kalantari et al. (2018) proposed a subspace approach based on spectral domain constraints for the denoising of speech corrupted with colored noise. Jabloun and Champagne (2003) incorporated properties of human audition into the signal subspace approaches.

1.1.4. Statistical model based methods

McAulay and Malpass (1980) proposed a maximum-likelihood (ML) estimator of the clean speech short-time Fourier transform (STFT) magnitude. The clean speech spectra are assumed to be deterministic and the noise is modeled as zero-mean complex Gaussian. The ML estimate of the magnitude spectrum is combined with the noisy phase spectrum in order to reconstruct the speech signal. Ephraim and Malah (1984) proposed a Bayesian MMSE estimator of the short-time spectral amplitude (STSA) by assuming speech and noise to be statistically independent, zero-mean, complex Gaussian random variables. Hendricks et al. (2007) proposed an MMSE estimator of the clean speech discrete Fourier transform (DFT) coefficients assuming a combined stochastic-deterministic model on clean speech and showed that their estimator outperforms the STSA-MMSE estimator obtained with the stochastic model alone (given in Ephraim and Malah, 1984). McCallum and Guillemin (2013) proposed an MMSE-STSA estimator under stochastic-deterministic model assuming a nonzero-mean speech signal. Ephraim (1992) used hidden Markov model (HMMs) to model the dynamics of speech and noise processes. Erkelen et al. (2007) proposed MMSE estimators of clean speech DFT coefficients and DFT magnitudes assuming generalized Gamma distributions on speech. Kund et al. (2008) developed an MMSE estimator by using a Gaussian mixture model (GMM) for the clean speech signal. Martin (2005) proposed an MMSE estimator of the clean speech DFT coefficient assuming a super-Gaussian prior. Lotter and Vary (2005) proposed a maximum a posteriori (MAP) estimator assuming super-Gaussian statistics. Tsao and Lai (2016) proposed a generalized MAP estimator, which possess the ability to adjust the speech distortion based on SNR. Mowlaee et al. (2017) proposed a joint MAP estimator for clean speech spectral amplitude and phase assuming gamma distribution as the amplitude prior and von Mises distribution as the phase prior. Ephraim and Malah (1985) proposed an estimator that minimizes the MSE of log-magnitude spectra as it is perceptually more correlated. Loizou (2005) computed Bayesian estimators for magnitude spectrum using perceptual distortion metrics such as the Itakura-Saito distortion, hyperbolic-cosine distortion, etc. Mohammadiha et al. (2013) use a Bayesian non-negative matrix factorization (NMF) approach to obtain an MMSE estimate of the clean speech DFT magnitude.

1.1.5. Neural network approaches

Recently, Xu et al. (2015) have demonstrated the use of feedforward *deep neural network* (DNN) for learning the non-linear map from the noisy speech to its clean counterpart. Gao et al. (2017) proposed a DNN-based speech enhancement algorithm, which is capable of reducing background noise and speech interference in a speaker-dependent scenario. Eskimez et al. (2018) proposed DNN based speech enhancement approaches in the context of automatic speaker verification systems.

1.2. Our Contributions

The denoising performance of Bayesian techniques depends critically on the clean signal prior assumed — the more accurate the prior, the better is the performance. Since the speech signal is nonstationary and has a lot of variability, obtaining a reliable prior that covers a broad range of use-cases is challenging. These variabilities entail complex stochastic modeling that may also require a large amount of training data for parameter estimation. Further, a mismatch between the training and testing conditions could lead to a degradation in the performance. Also, for a given application, the choice of the denoising algorithm is also dependent on other criteria such as computational constraints, the need and availability of training data, etc. Hence, it would be desirable to have computationally efficient techniques that do not rely explicitly on the prior and require minimal or no training. The goal of this paper is to precisely satisfy this objective.

We introduce the notion of perceptual risk estimation for speech denoising. The random noise considered to be additive (Section 2). Direct minimization of the risk results in estimates that are a function of the clean speech signal. Considering a transform-domain Gaussian observation model, one can develop an unbiased estimate of the MSE based on Stein's lemma (Stein, 1981), which is referred to as Stein's unbiased risk estimator (SURE) in the literature (Blu and Luisier, 2007; 2008; Luisier et al., 2007; Benazza-Benyahia and Pesquet, 2005). The main advantage is that, unlike MSE, SURE does not require knowledge of the unknown clean signal prior. The state-of-the-art image denoising techniques are based on SURE minimization (Blu and Luisier, 2007; 2008; Luisier et al., 2007; Benazza-Benyahia and Pesquet, 2005; Atto et al., 2009). Zheng et al. (2011) proposed a DCT-domain speech enhancement framework for robust speech recognition based on SURE, where the denoising function is chosen as a linear expansion of elementary threshold functions (LET) and the parameters of the LET are obtained by minimizing SURE. Speech denoising in DFT-domain based on first-order SURE-LET was proposed in Muraka and Seelamantula (2012). We proposed generalizations of SURE (which is originally proposed for Gaussian noise) to a wide class of noise distributions and showed its application to speech denoising (Sadasivan and Seelamantula, 2016a; 2016b). Recently, Mai et al. (2018) proposed a combination of block smoothed sigmoid-based shrinkage estimator and MMSE-STSA estimator for denoising where the size of the time-frequency block that is subject to sigmoid shrinkage is optimized using SURE.

Even though MSE is a widely and successfully used distortion measure for signal denoising, in speech processing applications, distortion measures such as Itakura-Saito (IS), hyperbolic-cosine (cosh), weighted cosh, are known to be more perceptually relevant than MSE (Gray et al., 1980). Taking this into account, in this paper, we solve the speech denoising problem within the framework of perceptual risk estimation, wherein we derive unbiased estimates of speech-specific perceptual distortion measures and minimize them to obtain the corresponding denoising functions. We first introduced perceptual risk estimation based single-channel speech enhancement in Muraka and Seelamantula (2011), wherein we compared the MSE with the Itakura–Saito distortion considering additive white Gaussian noise only. In this paper, we significantly expand on that preliminary result and demonstrate how the idea of risk estimation based speech enhancement can be extended to the other widely deployed speech-specific perceptual distortion measures, and also provide a comprehensive performance evaluation in real-world nonstationary noise conditions.

Further, in practice, real-world disturbances generate bounded noise amplitudes and quantization limits the dynamic range. Therefore, we consider the more realistic case of a truncated Gaussian distribution for the samples. The details will be described in Section 2. In order to develop a risk estimator, we make use of Stein's lemma and its higher-order generalization, originally proposed for Gaussian noise (Section 3). The higher-order generalization becomes important in the context of perceptual distortion measures. Correspondingly, we develop the notion of *perceptual risk* optimization for speech enhancement (PROSE) (Section 4). The key advantage of the PROSE framework is that it allows one to replace an ensemble-averaged distortion measure by a practically viable surrogate. We employ a transform-domain *pointwise shrinkage estimator*. The optimum shrinkage estimator obtained by minimizing perceptual risk estimates turns out to be a nonlinear function of the observations and noise statistics. The computational complexity of the shrinkage estimators is low — hence, they are suitable for low-power and real-time applications such as in hearing aids, mobile phone communication applications, etc. We also consider parametric versions, which give additional flexibility to trade-off between residual noise level and speech distortion. It turns out that perceptually optimized denoising functions result in more noise attenuation than MSE. We also carry out objective assessment in terms of segmental signal-to-noise ratio (SSNR), perceptual evaluation of speech quality (PESQ) (ITU-T Rec. P.862, 2001), source-to-distortion ratio (SDR) (Vincent et al., 2006), short-time objective intelligibility (STOI) (Taal et al., 2011), and subjective assessment by means of listening tests and scoring as per ITU-T recommendations (ITU-T Rec. P.835, 2003) (Section 6).

2. Problem Formulation

Consider a short-time frame of noisy speech in which samples of clean speech s_n are distorted additively by noise w_n resulting in the observations:

$$x_n = s_n + w_n, \quad n = 1, 2, \cdots, N, \tag{1}$$

where N is the frame length. The noise samples $\{w_n\}$ are assumed to be zero-mean, bounded, i.i.d. random variables. There is no prior assumed explicitly on the clean signal. Most real-world noise processes are bounded and the presence of a quantizer in a practical data acquisition scenario further limits the dynamic range. Consequently, $\{x_n\}$ are bounded, and $\mathcal{E}\{x_n\} = s_n$, which implies that $\{x_n\}$ are independent, but not identically distributed. The time-domain noise distribution is not restricted to be a Gaussian. Typical speech enhancement approaches work on the DFT domain (Hendriks et al., 2007). We prefer the discrete cosine transform (DCT) as it is real-valued and known to give rise to a more parsimonious representation than the DFT (Soon et al., 1998). Further, Soon et al. established that, for shrinkage estimators, DCT-domain denoising is superior to DFT (Soon et al., 1998). Since our pointwise multiplicative shrinkage estimator belongs to this class, we prefer the DCT to DFT. The DCT representation of (1) is

$$X_k = S_k + W_k, \quad k = 1, 2, \cdot \cdot \cdot, N, \tag{2}$$

where the transform-domain noise $\{W_k\}$ being a linear combination of i.i.d. random variables $\{w_n\}$ has a distribution that approaches a Gaussian by virtue of the *central limit theorem*. However, since $\{w_n\}$ are bounded, $\{W_k\}$ will also be bounded. These two properties taken together make a truncated Gaussian distribution model more appropriate and realistic for the transform coefficients than the standard Gaussian. The noisy samples are *concentrated* about the mean and the deviations from the mean are bounded. The suitability of the truncated Gaussian for modeling real-world processes in general has been advocated by Burkardt (2014).

The goal is to estimate S_k given X_k and noise statistics. Let $d(S_k, \widehat{S}_k)$ denote a distortion measure that quantifies the deviation of the estimate $\widehat{S}_k$ from S_k. The corresponding ensemble averaged distortion or *risk*, as referred to in the statistics literature, is defined as $\mathcal{R} = \mathcal{E}\left\{ d(S_k, \widehat{S}_k) \right\}$, where $\mathcal{E}$ denotes the expectation operator. The estimate is expressed as $\hat{S}_k = f(X_k)$, where f is the *denoising function*, which may not always be linear. We consider pointwise shrinkage, $f(X_k) = a_k X_k, a_k \in [0, 1]$ (Yu et al., 2008), and optimize an estimate of $\mathcal{R}$ with respect to a_k.

A block diagram representation of the proposed method is shown in Fig. 1. To take into account the quasi-stationarity of speech, denoising is performed on a frame-by-frame basis and the enhanced speech is reconstructed using the standard overlap-add synthesis methodology.

3. Risk Estimation Results

We recall a key result from Stein (1981), which is central to the subsequent developments

Lemma 1. *(Stein, 1981) Let W be a $\mathcal{N}(0, \sigma^2)$ real random variable and let $f : (\mathbb{R}) \to (\mathbb{R})$ be an indefinite integral of the Lebesgue measurable function $f^{(1)}$, essentially the derivative of f. Suppose also that $\mathcal{E}\{|f^{(1)}(W)|\} < \infty$. Then, $\mathcal{E}\{W f(W)\} = \sigma^2 \mathcal{E}\{f^{(1)}(W)\}$.*

Stein's lemma facilitates estimation of the mean of *Wf*(*W*) in terms of $f^{(1)}(W)$. Effectively, $\sigma^2 f^{(1)}(W)$ could be used as an unbiased estimate of $\mathcal{E}\{W f(W)\}$. The implications of this apparently simple result can be appreciated when we are required to compute an unbiased estimator of the MSE. Next, we develop a higher-order generalization of Stein's lemma.

Lemma 2. *(Generalized Stein's lemma) Let W be a $\mathcal{N}(0, \sigma^2)$ real random variable and let $f : (\mathbb{R}) \to (\mathbb{R})$ be an n-fold indefinite integral of the Lebesgue measurable function $f^{(n)}$, which is the n^{th} derivative of f. Suppose also that $\mathcal{E}\left\{ |W^{(n+1-k)} f^{(k)}(W)| \right\} < \infty, k = 1, 2, \cdots, n$. Then*

$$\mathcal{E}\left\{ W^{n+1} f(W) \right\} = \sigma^2 \mathcal{E}\left\{ f^{(1)}(W) W^n \right\} + \sigma^2 \, n \mathcal{E}\left\{ f(W) W^{n-1} \right\}.$$

Proof. Let us express $\mathcal{E}\{W^{n+1} f(W)\}$ as $\mathcal{E}\{W g(W)\}$, where $g(W) = W^n f(W)$. Applying Stein's lemma to $\mathcal{E}\{W g(W)\}$, we get that

$$\begin{aligned} \mathcal{E}\left\{ W^{n+1} f(W) \right\} &= \mathcal{E}\{W g(W)\} = \sigma^2 \mathcal{E}\left\{ g^{(1)}(W) \right\}, \\ &= \sigma^2 \mathcal{E}\left\{ f^{(1)}(W) W^n \right\} + \sigma^2 \, n \mathcal{E}\left\{ f(W) W^{n-1} \right\}. \end{aligned}$$

□

Stein's lemma could be applied recursively to each of the terms on the right-hand side, up to a stage where the terms comprise only the derivatives of all orders of *f* up to *n*. Our next set of results is in the context of developing Stein-type lemmas for the practical case of a truncated Gaussian distribution.

Lemma 3. *Let W be a real random variable with probability density function (p.d.f.)*

$$p(w; c, \sigma) = \frac{1}{\sqrt{2\pi}\sigma K} \exp\left(-\frac{w^2}{2\sigma^2} \right) \mathbb{1}_{\{w < |c\sigma|\}}, \tag{3}$$

where K ensures that p integrates to unity, $c \in \mathbb{R}^+$, and $\mathbb{1}$ denotes the indicator function. Let $f : (\mathbb{R}) \to (\mathbb{R})$ be an indefinite integral of the Lebesgue measurable function $f^{(1)}$. Suppose also that $\mathcal{E}\{|f^{(1)}(W)|\} < \infty$ and f does not grow faster than an exponential, then $\mathcal{E}\{W f(W)\} = \sigma^2 \mathcal{E}\{f^{(1)}(W)\} + \mathcal{O}\{\exp(-c^2)\}$ where $\mathcal{O}$ denotes the Big-O notation (Landau symbol).

Proof. Using the property: $-\sigma^2 \frac{\mathrm{d} \exp(-\frac{w^2}{2\sigma^2})}{\mathrm{d}w} = w \exp(-\frac{w^2}{2\sigma^2})$, we write

$$\mathcal{E}\{W f(W)\} = \int_{-\infty}^{+\infty} w f(w) p(w; c, \sigma) \mathrm{d}w,$$

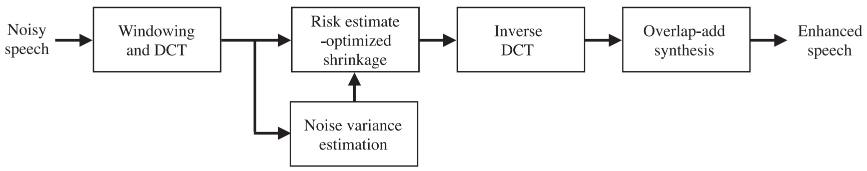


Fig. 1. A block diagram representation of the PROSE methodology.

$$= -\int_{-c\sigma}^{+c\sigma} \sigma^2 f(w) \frac{1}{\sqrt{2\pi}\sigma K} \frac{\mathrm{d}\exp\left(-\frac{w^2}{2\sigma^2}\right)}{\mathrm{d}w}\mathrm{d}w,$$
$$= -\sigma^2 f(w)p(w;c,\sigma)\Big|_{-c\sigma}^{+c\sigma} + \int_{-c\sigma}^{+c\sigma} \sigma^2 f^{(1)}(w)p(w;c,\sigma)\mathrm{d}w$$
$$= \mathcal{O}\{\exp(-c^2)\} + \sigma^2\mathcal{E}\{f^{(1)}(W)\}.$$

□

For even f, the approximation error is zero. Next, we state the counterpart of Lemma 2 for truncated Gaussian distribution, which has a similar proof mechanism.

Lemma 4. *Let W be a real random variable with p.d.f. $p(w;\, c,\, \sigma)$ and let $f : (\mathbb{R}) \to (\mathbb{R})$ be an n-fold indefinite integral of the Lebesgue measurable function $f^{(n)}$, which is the n^{th} derivative of f. Suppose also that $\mathcal{E}\left\{|W^{(n+1-k)} f^{(k)}(W)|\right\} < \infty, k = 1, 2, \cdots, n$ and $f^{(k)}$ does not grow faster than an exponential, then*

$$\mathcal{E}\left\{W^{n+1} f(W)\right\} = \sigma^2\mathcal{E}\left\{f^{(1)}(W)W^n\right\} + \sigma^2 n\mathcal{E}\left\{f(W)W^{n-1}\right\} + \mathcal{O}\left\{\exp(-c^2)\right\}.$$

The approximation error is negligible for large values of c. These results will be handy in computing risk estimators for various perceptual distortion measures.

4. Perceptual Risk Optimization for Speech Enhancement (PROSE)

4.1. Mean-Square Error (MSE)

The squared error is the most commonly employed distortion measure largely because of ease of optimization. The distortion function for squared error in the transform domain is

$$d\left(S_k, \widehat{S}_k\right) = \left(\widehat{S}_k - S_k\right)^2, \text{ where } \widehat{S}_k = f(X_k).$$

The MSE is $\mathcal{R} = \mathcal{E}\{d(S_k, \widehat{S}_k)\}$, which may be expanded as

$$\mathcal{R} = \mathcal{E}\left\{f^2(X_k) - 2f(X_k)X_k + 2f(X_k)W_k\right\} + S_k^2. \tag{4}$$

The last term in (4) does not affect the optimization. Applying Lemma 3 gives $\mathcal{E}\{f(X_k)W_k\} \approx \sigma^2\mathcal{E}\{f^{(1)}(X_k)\}$, and from (4), we get that

$$\mathcal{R} \approx \mathcal{E}\left\{f^2(X_k) - 2f(X_k)X_k + 2\sigma^2 f^{(1)}(X_k)\right\} + S_k^2,$$

from which it can be concluded that

$$\widehat{\mathcal{R}} = f^2(X_k) - 2f(X_k)X_k + 2\sigma^2 f^{(1)}(X_k) + S_k^2, \tag{5}$$

is a nearly unbiased estimator of the MSE. Although $\widehat{\mathcal{R}}$ contains the signal term S_k^2, it does not affect the minimization with respect to f. Consider the pointwise shrinkage estimator $f(X_k) = a_k X_k$, where $a_k \in [0, 1]$. The optimum value of a_k is obtained by minimizing $\widehat{\mathcal{R}}$ subject to the constraint $a_k \in [0,\ 1]$. The Karush–Kuhn–Tucker (KKT) conditions (Fletcher, 1987, pp. 211) for solving this problem are given in Appendix A (cf. (A.1a) – (A.1e)). The optimum a_k that satisfies the KKT conditions[1] is given by

$$a_k = \max\left\{1 - \frac{\sigma^2}{X_k^2}, 0\right\}.$$

The optimum shrinkage estimator becomes:

$$\widehat{S}_k = \max\left\{1 - \frac{1}{\xi_k}, 0\right\}X_k = a_{\mathrm{MSE}}(\xi_k)X_k,$$

where $\xi_k = \frac{X_k^2}{\sigma^2}$ denotes the a posteriori SNR determined based on the noisy signal, and $a_{\mathrm{MSE}}(\xi_k)$ denotes the MSE-related shrinkage function. To impart additional flexibility, we consider parametric refinements, which allow us to trade-off between residual noise level and speech distortion. They are also useful when estimates of the noise variance may not be sufficiently accurate. The parametrically refined version is given by

$$\widehat{S}_k = \max\left\{1 - \frac{\alpha}{\xi_k}, 0\right\}X_k, \tag{6}$$

where α is the parameter, akin to the over-subtraction factor in spectral subtraction algorithms.

4.2. Weighted Euclidean (WE) Distortion

The MSE is perceptually less relevant for speech signals since a large MSE does not always imply poor signal quality. Auditory masking effects render humans more robust to errors at spectral peaks than at the valleys and are taken advantage of in speech/audio compression. One way to introduce differential weighting to spectral peaks and valleys is to consider the weighted Euclidean (WE) distortion:

$$d(S_k, \widehat{S}_k) = \frac{\left(\widehat{S}_k - S_k\right)^2}{S_k}, \text{where } \widehat{S}_k = f(X_k).$$

The measure $d < 0$ when $S_k < 0$, but this is not a problem, because in that case, the cost function must be maximized and not minimized. We shall see that the proposed methodology implicitly takes care of this aspect. Developing d, we get

$$d\left(S_k, \widehat{S}_k\right) = \frac{\widehat{S}_k^2}{S_k} + S_k - 2\widehat{S}_k.$$

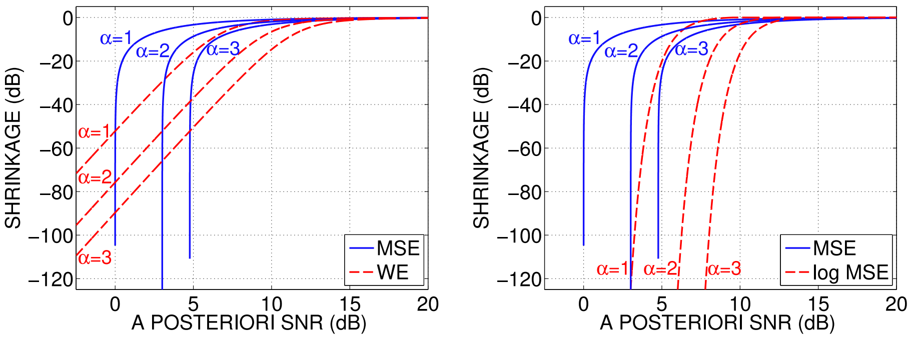
Let us consider a high-SNR scenario, that is, $\left|\frac{W_k}{X_k}\right| < 1$. This event occurs with probability 1 if $|S_k| > 2c\sigma$ (cf. Appendix B). This condition also implies that X_k and S_k have the same sign. In this scenario, the first term is expanded as

$$\frac{\widehat{S}_k^2}{S_k} = \frac{\widehat{S}_k^2}{X_k}\left(1 - \frac{W_k}{X_k}\right)^{-1} = \frac{\widehat{S}_k^2}{X_k}\sum_{n=0}^{\infty}\left(\frac{W_k}{X_k}\right)^n,$$

and the distortion function is rewritten as

$$d(S_k, \widehat{S}_k) = \frac{\widehat{S}_k^2}{X_k}\sum_{n=0}^{\infty}\left(\frac{W_k}{X_k}\right)^n + S_k - 2\widehat{S}_k.$$

[1] The calculations related to the constrained optimization of all the risk estimates considered in this paper are provided in the supporting document.



(a) (b)

Fig. 2. [Color online] A comparison of shrinkage profiles: (a) MSE versus WE; and (b) MSE versus log MSE.

We truncate the infinite sum beyond the fourth-order, in order to maintain high accuracy in the calculations:

$$d(S_k,\widehat{S}_k) \approx \frac{\widehat{S}_k^2}{X_k}\sum_{n=0}^{4}\left(\frac{W_k}{X_k}\right)^n + S_k - 2\widehat{S}_k.$$

Considering $f(X_k) = a_k X_k$, we seek to minimize the risk $\mathcal{R}$:

$$\mathcal{R} = \mathcal{E}\left\{ a_k^2 X_k \sum_{n=0}^{4}\left(\frac{W_k}{X_k}\right)^n \right\} + S_k - 2\mathcal{E}\{a_k X_k\}. \tag{7}$$

To proceed further, we are required to compute expectations of reciprocals of truncated Gaussian random variables, which may not always be finite. A necessary and sufficient condition for the expectations to be finite is that $|S_k| > c\sigma$, which is satisfied in the high-SNR regime. This is an added benefit of working with more realistic distributions such as the truncated Gaussian.

A simplification for the expectation of the sum appearing in the first term of (7) could be made by invoking Lemma 4 (cf. Appendix C). Consequently, we get

$$\mathcal{R} = a_k^2\mathcal{E}\left\{ X_k + \frac{\sigma^2}{X_k} - \frac{\sigma^4}{X_k^3} + 48\frac{\sigma^6}{X_k^5} + 360\frac{\sigma^8}{X_k^7} - 2\frac{X_k}{a_k} \right\} + S_k.$$

The corresponding unbiased risk estimator is obtained as

$$\widehat{\mathcal{R}} = a_k^2\left(X_k + \frac{\sigma^2}{X_k} - \frac{\sigma^4}{X_k^3} + 48\frac{\sigma^6}{X_k^5} + 360\frac{\sigma^8}{X_k^7} \right) - 2a_k X_k + S_k. \tag{8}$$

The goal is to determine $a_k \in [0, 1]$ such that $\widehat{\mathcal{R}}$ is minimized if $S_k > 0$ and maximized if $S_k < 0$. Solving the KKT conditions (A.1a) - (A.1e) corresponding to these two scenarios results in the same optimum a_k. The corresponding estimator is given by

$$\widehat{S}_k = \underbrace{\left(1 + \frac{1}{\xi_k} - \frac{1}{\xi_k^2} + \frac{48}{\xi_k^3} + \frac{360}{\xi_k^4} \right)^{-1}}_{a_{\mathrm{WE}}(\xi_k)} X_k, \tag{9}$$

where $a_{\mathrm{WE}}(\xi_k)$ is the shrinkage factor. Akin to (6), one can define parametric counterparts of a_{WE} by replacing ξ_k with ξ_k/α. Fig. 2(a) compares a_{MSE} and a_{WE} for different values of α. We observe that the attenuation provided by a_{MSE} is considerably smaller than a_{WE} when the a posteriori SNR is greater than zero dB. This indicates that, in the high-SNR regime, minimizing the weighted Euclidean distortion results in higher noise attenuation than the MSE. As α increases, the shrinkage curves shift to the right, implying more attenuation for a given a posteriori SNR.

4.3. Logarithmic Mean-Square Error (log MSE)

Since loudness perception of the peripheral auditory system is logarithmic, one could compare speech spectra by computing the MSE on a logarithmic scale. This property has been used to advantage in vector quantization and speech coding applications (Gray et al., 1980). The log MSE between S_k and $\widehat{S}_k$ is given by

$$d(S_k,\widehat{S}_k) = \left(\log\frac{\widehat{S}_k}{S_k}\right)^2 = \left(\log\widehat{S}_k\right)^2 + \left(\log S_k\right)^2 - 2\log S_k \log\widehat{S}_k. \tag{10}$$

Recall that we consider only non-negative shrinkage functions for denoising, and that under the high SNR assumption both $\widehat{S}_k$ and S_k have the same sign. Consequently, the argument of the logarithm is always positive. The last term in (10) is rewritten as

$$\begin{aligned}\log\widehat{S}_k \log S_k &= \log\widehat{S}_k \log(X_k - W_k), \text{(from (2))},\\ &= \log\widehat{S}_k \log X_k + \log\widehat{S}_k \log\left(1 - \frac{W_k}{X_k}\right),\\ &= \log\widehat{S}_k \log X_k - \log\widehat{S}_k \sum_{n=1}^{\infty}\frac{1}{n}\left(\frac{W_k}{X_k}\right)^n.\end{aligned}$$

Substituting in (10), truncating the series beyond $n = 4$, and considering the expectation results in

$$\mathcal{R} = \mathcal{E}\left\{ \left(\log\widehat{S}_k\right)^2 + \left(\log S_k\right)^2 - 2\log\widehat{S}_k \log X_k + 2\log\widehat{S}_k \sum_{n=1}^{4}\frac{1}{n}\left(\frac{W_k}{X_k}\right)^n \right\}.$$

For the shrinkage estimate $\widehat{S}_k = a_k X_k$, the risk becomes

$$\mathcal{R} = \mathcal{E}\left\{ (\log a_k X_k)^2 + (\log S_k)^2 - 2\log a_k X_k \log X_k + 2\sum_{n=1}^{4}\frac{\log a_k X_k}{n}\left(\frac{W_k}{X_k}\right)^n \right\}.$$

The last term may be simplified as shown in Appendix D, using Lemma 4. Consequently, the unbiased risk estimator is

$$\begin{aligned}\widehat{\mathcal{R}} = &\left(\log S_k\right)^2 - 2\log a_k X_k \log X_k\\ &+ 2\left(\frac{\sigma^2}{X_k^2} - 1.5\frac{\sigma^4}{X_k^4} + 2.17\frac{\sigma^6}{X_k^6} - 159.5\frac{\sigma^8}{X_k^8} \right)\\ &- 2\log a_k X_k\left(0.5\frac{\sigma^2}{X_k^2} - 0.75\frac{\sigma^4}{X_k^4} - 10\frac{\sigma^6}{X_k^6} - 210\frac{\sigma^8}{X_k^8} \right)\\ &+ \left(\log a_k X_k\right)^2.\end{aligned} \tag{11}$$

The optimum value of $a_k \in [0, 1]$ obtained by solving (A.1a) – (A.1e) in this case is:

$$a_k = \min\left\{ \exp\left(0.5\frac{\sigma^2}{X_k^2} - 0.75\frac{\sigma^4}{X_k^4} - 10\frac{\sigma^6}{X_k^6} - 210\frac{\sigma^8}{X_k^8} \right), 1 \right\}.$$

The corresponding estimate of S_k is given by

$$\widehat{S}_k = \underbrace{\min\left\{ \exp\left(\frac{0.5}{\xi_k} - \frac{0.75}{\xi_k^2} - \frac{10}{\xi_k^3} - \frac{210}{\xi_k^4} \right), 1 \right\}}_{a_{\log\,\mathrm{MSE}}(\xi_k)} X_k. \tag{12}$$

The parametrized shrinkage is given by $a_{\log \mathrm{MSE}}\left(\frac{\xi_k}{\alpha}\right)$. A comparison of a_{MSE} and $a_{\log \mathrm{MSE}}$ is shown in Fig. 2(b). At low SNRs, log MSE results in a higher attenuation than MSE.

4.4. Itakura–Saito (IS) Distortion

The IS distortion, although not symmetric, is a popular quality measure used in speech coding due to its perceptual relevance in matching two power spectra. Here, we compute the IS distortion between the DCT coefficients of the noise-free speech and its estimate considering both of them to have the same sign:

$$d\left(S_k, \widehat{S}_k\right) = \frac{\widehat{S}_k}{S_k} - \log \frac{\widehat{S}_k}{S_k} - 1.$$

In the high-SNR scenario, the distortion measure is expanded as

$$\begin{aligned} d\left(S_k, \widehat{S}_k\right) &= \frac{\widehat{S}_k}{X_k}\left(1 - \frac{W_k}{X_k}\right)^{-1} - \log \widehat{S}_k + \log S_k - 1, \\ &= \frac{\widehat{S}_k}{X_k} \sum_{n=0}^{\infty} \left(\frac{W_k}{X_k}\right)^n - \log \widehat{S}_k + \log S_k - 1. \end{aligned}$$

Expressing $\widehat{S}_k = a_k X_k$, and truncating the series beyond $n = 4$, the risk turns out to be

$$\mathcal{R} = \sum_{n=0}^{4} \mathcal{E}\left\{\frac{a_k W_k^n}{X_k^n}\right\} - \mathcal{E}\left\{\log a_k X_k\right\} + \log S_k - 1. \tag{13}$$

The first term is evaluated using Lemma 4 (cf. Appendix E) resulting in the risk

$$\mathcal{R} = \mathcal{E}\left\{a_k\left(1 + 60\frac{\sigma^6}{X_k^6} + 840\frac{\sigma^8}{X_k^8}\right) - \log a_k X_k\right\} + \log S_k - 1.$$

The corresponding unbiased estimator of $\mathcal{R}$ is

$$\widehat{\mathcal{R}} = a_k\left(1 + 60\frac{\sigma^6}{X_k^6} + 840\frac{\sigma^8}{X_k^8}\right) - \log a_k X_k + \log\, S_k - 1, \tag{14}$$

which when optimized with respect to $a_k \in [0, 1]$ (solution of (A.1a) to (A.1e)) gives

$$a_k = \left(1 + \frac{60}{\xi_k^3} + \frac{840}{\xi_k^4}\right)^{-1} \Rightarrow \widehat{S}_k = \underbrace{\left(1 + \frac{60}{\xi_k^3} + \frac{840}{\xi_k^4}\right)^{-1}}_{a_{\mathrm{IS}}(\xi_k)} X_k. \tag{15}$$

Fig. 3(a) shows a comparison of a_{IS} and a_{MSE}. For a posteriori SNR greater than 0 dB, the attenuation is higher in the case of Itakura-Saito distortion.

4.5. Itakura-Saito Distortion Between DCT Power Spectra (IS-II)

We next consider the IS distortion between S_k^2 and $\widehat{S}_k^2$:

$$\begin{aligned} d\left(S_k, \widehat{S}_k\right) &= \frac{\widehat{S}_k^2}{S_k^2} - \log \frac{\widehat{S}_k^2}{S_k^2} - 1, \\ &= \frac{\widehat{S}_k^2}{X_k^2}\left(1 - \frac{W_k}{X_k}\right)^{-2} - \log \widehat{S}_k^2 + \log S_k^2 - 1, \\ &= \frac{\widehat{S}_k^2}{X_k^2} \sum_{n=0}^{\infty} (n+1)\left(\frac{W_k}{X_k}\right)^n - \log \widehat{S}_k^2 + \log S_k^2 - 1. \end{aligned}$$

Considering $\widehat{S}_k = a_k X_k$, and truncating the series beyond $n = 4$, results in the risk

$$\mathcal{R} = \mathcal{E}\left\{a_k^2 \sum_{n=0}^{4} (n+1)\left(\frac{W_k}{X_k}\right)^n\right\} - \mathcal{E}\left\{\log a_k^2 X_k^2\right\} + \log S_k^2 - 1.$$

Simplifying the first term using Lemma 4 gives

$$\begin{aligned} \mathcal{R} = {} & a_k^2 \mathcal{E}\left\{1 + \frac{\sigma^2}{X_k^2} - 3\frac{\sigma^4}{X_k^4} + 360\frac{\sigma^6}{X_k^6} + 4200\frac{\sigma^8}{X_k^8}\right\} \\ & - \mathcal{E}\left\{\log a_k^2 X_k^2\right\} + \log S_k^2 - 1. \end{aligned}$$

An unbiased estimator of $\mathcal{R}$ is

$$\widehat{\mathcal{R}} = a_k^2\left(1 + \frac{\sigma^2}{X_k^2} - 3\frac{\sigma^4}{X_k^4} + 360\frac{\sigma^6}{X_k^6} + 4200\frac{\sigma^8}{X_k^8}\right) - \log a_k^2 X_k^2 + \log S_k^2 - 1. \tag{16}$$

Optimizing $\widehat{\mathcal{R}}$ with respect to $a_k \in [0, 1]$, following (A.1a) – (A.1e), we get

$$\begin{aligned} a_k &= \min\left\{1, \left(1 + \frac{\sigma^2}{X_k^2} - 3\frac{\sigma^4}{X_k^4} + 360\frac{\sigma^6}{X_k^6} + 4200\frac{\sigma^8}{X_k^8}\right)^{-\frac{1}{2}}\right\}, \\ \Rightarrow \widehat{S}_k &= \underbrace{\min\left\{1, \left(1 + \frac{1}{\xi_k} - \frac{3}{\xi_k^2} + \frac{360}{\xi_k^3} + \frac{4200}{\xi_k^4}\right)^{-\frac{1}{2}}\right\}}_{a_{\text{IS-II}}(\xi_k)} X_k, \end{aligned} \tag{17}$$

which we shall refer to as the IS-II estimator. A comparison of the IS and IS-II shrinkage functions is shown in Fig. 3(b). At low SNRs, IS attenuates more than IS-II.

4.6. Hyperbolic Cosine Distortion Measure (COSH)

A symmetrized version of the IS distortion results in the *cosh* measure:

$$d(S_k, \widehat{S}_k) = \cosh\left(\log \frac{S_k}{\widehat{S}_k}\right) - 1 = \frac{1}{2}\left[\frac{S_k}{\widehat{S}_k} + \frac{\widehat{S}_k}{S_k}\right] - 1. \tag{18}$$

The corresponding risk is $\mathcal{R} = \frac{1}{2}\mathcal{E}\{\frac{S_k}{\widehat{S}_k}\} + \frac{1}{2}\mathcal{E}\{\frac{\widehat{S}_k}{S_k}\} - 1$. Substituting $\widehat{S}_k = a_k X_k$, and invoking Lemma 4, the expectations turn out to be

$$\mathcal{E}\left\{\frac{S_k}{a_k X_k}\right\} = \mathcal{E}\left\{\frac{1}{a_k} + \frac{\sigma^2}{a_k X_k^2}\right\}, \quad \text{and}$$

$$\mathcal{E}\left\{\frac{a_k X_k}{S_k}\right\} = \mathcal{E}\left\{a_k\left(1 + 60\frac{\sigma^6}{X_k^6} + 840\frac{\sigma^8}{X_k^8}\right)\right\}.$$

Correspondingly, the unbiased risk estimator is

$$\widehat{\mathcal{R}} = \frac{1}{2}\left(\frac{1}{a_k} + \frac{\sigma^2}{a_k X_k^2} + a_k\left(1 + 60\frac{\sigma^6}{X_k^6} + 840\frac{\sigma^8}{X_k^8}\right)\right) - 1. \tag{19}$$

Optimizing $\widehat{\mathcal{R}}$ with respect to $a_k \in [0, 1]$ following (A.1a)–(A.1e) gives

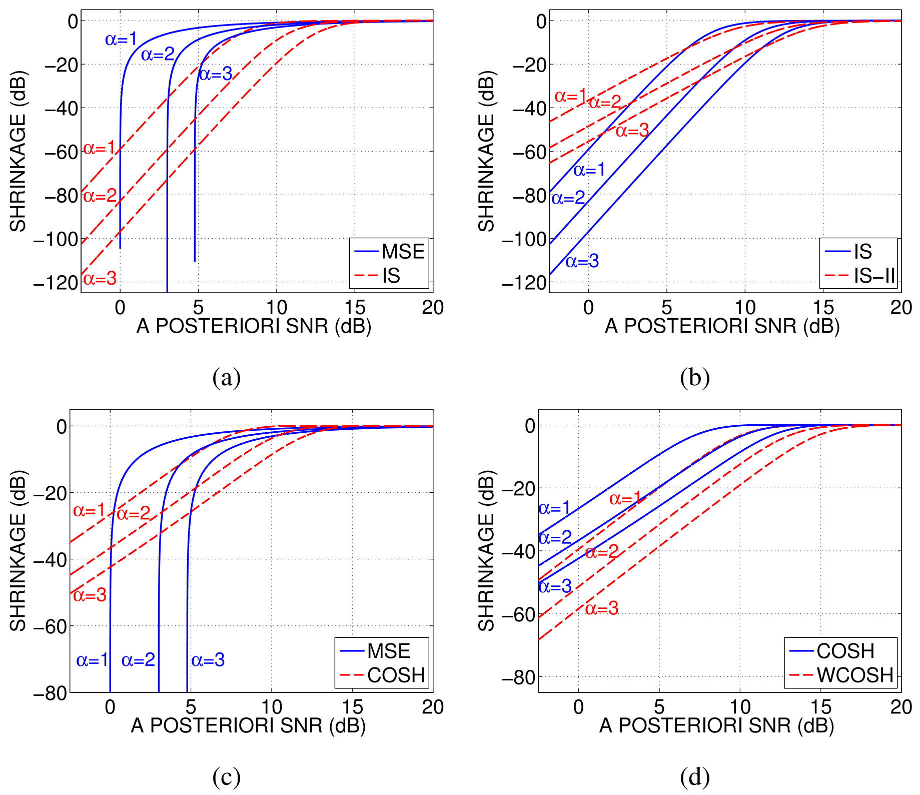
$$\widehat{S}_k = \underbrace{\min\left\{1, \sqrt{\frac{1 + \frac{1}{\xi_k}}{1 + 60\frac{1}{\xi_k^3} + 840\frac{1}{\xi_k^4}}}\right\}}_{a_{\mathrm{COSH}}(\xi_k)} X_k. \tag{20}$$

Fig. 3(c) shows a comparison of a_{MSE} and a_{COSH} versus ξ_k – it is clear that a_{COSH} results in a higher attenuation than a_{MSE}.

4.7. Weighted COSH Distortion (WCOSH)

Similar to the weighted Euclidean measure, we consider the weighted cosh distortion:

$$d(S_k, \widehat{S}_k) = \left(\frac{1}{2}\left[\frac{S_k}{\widehat{S}_k} + \frac{\widehat{S}_k}{S_k}\right] - 1\right)\frac{1}{S_k}. \tag{21}$$



(a)

(b)

(c)

(d)

Fig. 3. [Color online] A comparison of shrinkage profiles: (a) MSE versus IS; (b) IS versus IS-II; (c) MSE versus COSH; and (d) COSH versus WCOSH.

The risk estimator can be computed as in the case of the cosh measure (cf. Appendix F). The optimal estimate is given by

$$\widehat{S}_k = \underbrace{\min\left\{1, \left(1 - \frac{1}{\xi_k} + \frac{3}{\xi_k^2} + \frac{420}{\xi_k^3} + \frac{8400}{\xi_k^4}\right)^{-\frac{1}{2}}\right\}}_{a_{\mathrm{WCOSH}}(\xi_k)} X_k. \tag{22}$$

A comparison of a_{COSH} and a_{WCOSH} shown in Fig. 3(d) indicates that the latter results in a higher noise attenuation at high noise levels.

5. Implementation

For experimental validation, we use the clean speech from the NOIZEUS database (Hu and Loizou, 2007) (8 kHz sampling frequency and 16-bit quantization) and nonstationary noise data from both NOIZEUS and NOISEX-92 (Varga and Steeneken, 1993) databases. Noisy speech is generated by adding noise at a desired global SNR. The noisy speech signal is processed in the DCT domain on a frame-by-frame basis. The frame size is 40 ms, the window function is Hamming, and the overlap between consecutive frames is 75%.

The developments in the PROSE framework assumed knowledge of the noise variance in each bin of a given frame. In practice, the noise variance has to be estimated. For this purpose, we use the likelihood-ratio-test-based voice activity detector (VAD) of Sohn et al. (1999), which was also employed in the extensive comparisons reported by Hu and Loizou (2007). The inverse a posteriori SNR is then estimated using the recursive formula

$$\frac{1}{\widehat{\xi}_k(i)} = \beta \frac{\widehat{\sigma}_k^2(i)}{X_k^2(i)} + (1-\beta)\max\left(1 - \frac{\widehat{S}_k^2(i-1)}{X_k^2(i-1)}, 0\right), \tag{23}$$

where k denotes the index of the DCT coefficient, i denotes the frame index, and β is a smoothing parameter (set to 0.98 in our experiments), $\widehat{\sigma}_k^2(i)$ is the estimate of the noise variance at the k^{th} DCT coefficient in the i^{th} frame, $\widehat{S}_k(i-1)$ is the denoised k^{th} DCT coefficient in the $(i-1)^{th}$ frame. Recursive estimation of the a posteriori SNR helps to reduce the musical noise artifacts in the denoised signal. Further, the minimum value of the shrinkage parameter is chosen to be a small value (0.05 in our experiments) in order to obtain a spectral floor thereby reducing the effect of musical noise. The first few frames are assumed to contain only noise. For the first frame, β is set to unity, and the k^{th} noise variance is estimated by averaging the noise variances of the k^{th} coefficient in the first ten frames. The noise variance is updated in the noise-only frames as classified by the VAD following the recursion:

$$\widehat{\sigma}_k^2(i) = \begin{cases} \eta\widehat{\sigma}_k^2(i-1) + (1-\eta)X_k^2(i), \text{under} \quad \mathcal{H}_0, \\ \widehat{\sigma}_k^2(i-1), \text{under} \quad \mathcal{H}_1, \end{cases}$$

where $\mathcal{H}_0$ and $\mathcal{H}_1$ are the null hypothesis (noise-only) and the alternative hypothesis (signal + noise), respectively. The value of $\eta = 0.98$ following Hu and Loizou (2007). The noisy speech signals are denoised using shrinkage functions corresponding to various distortion measures considered in this paper. We set $\alpha = 1.75$ uniformly across all measures in the PROSE framework as it was found to give better results than $\alpha = 1$.

6. Experimental Results

The denoising performance of PROSE estimators is evaluated using both objective measures and subjective listening tests. For benchmarking, we use three algorithms: (i) Wiener filter method, where a priori SNR is estimated using the decision-directed approach (WFIL) (Scalart and Filho, 1996); (ii) a Bayesian estimator for short-time log-spectral amplitude (LSA) (Ephraim and Malah, 1985); and (iii) Bayesian non-negative matrix factorization algorithm (BNMF) (Mohammadiha et al., 2013)[2], which gives an MMSE estimate of the clean speech DFT magnitude. In the BNMF approach, the

[2] Matlab implementation available online: https://www.uni-oldenburg.de/en/mediphysics-acoustics/sigproc/staff/nasser-mohammadiha/matlab-codes/.

speech bases are learned offline and the noise bases are learned online. BNMF has been shown to be the best among the NMF approaches for speech enhancement (Mohammadiha et al., 2013). In an extensive evaluation carried out by Hu and Loizou (2007)[3], the WFIL and LSA algorithms were shown to perform better than the other techniques. Hu and Loizou's evaluation considered thirteen speech denoising algorithms encompassing four different classes of noise suppression methods (spectral subtraction techniques, subspace methods, statistical-model based approaches, and Wiener filter type algorithms). In terms of intelligibility also, WFIL and LSA techniques turned out to be superior (Ch. 11, pp. 567 of Loizou, 2007).

6.1. *Objective Evaluation*

The denoising performance is quantified in terms of: (i) segmental SNR (SSNR), a local metric, which is the average of the SNRs computed over short segments; (ii) perceptual evaluation of speech quality (PESQ) (ITU-T Rec. P.862, 2001), which is an objective score for assessing end-to-end speech quality in narrowband telephone networks and speech codecs, described in ITU-T Recommendation P.862 (ITU-T Rec. P.862, 2001); (iii) source-to-distortion ratio (SDR), which measures the overall quality of the estimated source signal considering the denoising problem as a source-separation problem (Vincent et al., 2006)[4]; and (iv) short-time objective intelligibility measure (STOI) (Taal et al., 2011)[5], which ranges from 0 to 1 and reflects the correlation between short-time temporal envelope of clean speech and denoised speech. It has been shown to correlate highly with the intelligibility scores obtained through listening tests (Taal et al., 2011). For SSNR, PESQ, and SDR, we report the *gains* achieved by denoising. The SSNR gain is the difference between the output and input SSNR values. The PESQ gain and SDR gain are also computed similarly.

We consider all 30 speech files from the NOIZEUS database (Hu and Loizou, 2007), and four noise types — *White noise, F16 noise, Train noise*, and *Street noise*. The results presented here are obtained after averaging over the entire database for 50 noise realizations.

6.1.1. *White noise*

Fig. 4 compares the performance in *White noise*. For input SNRs in the range of −5 dB to 20 dB, PROSE with COSH, log-MSE, IS, and WE measures results in a higher SSNR gain (cf. Fig. 4(a)). As the input SNR increases, the performance gains offered by PROSE over BNMF, LSA, and WFIL increase. For input SNRs between −5 and 20 dB, the PESQ gain is maximum for WE, log MSE, and IS (cf. Fig. 4(b)), whereas for SNRs below 5 dB, BNMF also has a high PESQ gain. Within the PROSE framework, the PESQ gains are higher for perceptually motivated distortions than the MSE. PROSE with WE, log MSE, COSH, and IS distortions is better than the benchmark techniques in terms of SSNR and PESQ. In terms of SDR as well, PROSE employing perceptual distortion measures exhibits a superior performance (cf. Fig. 4(c)). For input SNRs between 0 and 10 dB, BNMF also shows a high SDR gain.

The STOI scores are shown in Table 1. To analyze the statistical significance of the average STOI score of the best performing technique, we perform a multiple paired comparison between the best performing technique and the others using Tukey's honestly significant difference test (Ott and Longnecker, 2016, Chapter 9) considering 95% confidence level. The algorithms that have high/comparable scores are indicated in boldface.

For input SNRs greater than 5 dB, PROSE is superior to the benchmark techniques in terms of STOI, except in the case of WFIL, where it is comparable. For input SNRs 5 dB or less, BNMF shows higher STOI scores than the other algorithms.

Table 1
Comparison of denoising performance in terms of STOI scores for different input SNRs (*White noise*). For a given input SNR, algorithms whose scores are shown in boldface were found to perform equally well. The other algorithms have an inferior performance.

SNR (dB)	−5	0	5	10	15	20
Input.	0.549	0.648	0.744	0.827	0.891	0.935
MSE	0.555	0.668	0.772	**0.854**	**0.912**	**0.949**
log MSE	0.519	0.642	0.762	**0.854**	**0.914**	**0.951**
WE	0.530	0.648	0.763	**0.855**	**0.914**	**0.951**
COSH	0.543	0.660	0.771	**0.857**	**0.915**	**0.952**
IS	0.528	0.642	0.758	**0.852**	**0.913**	**0.950**
IS-II	0.537	0.648	0.760	**0.852**	**0.912**	**0.950**
WCOSH	0.538	0.645	0.755	**0.845**	**0.910**	**0.948**
BNMF	**0.569**	**0.691**	**0.790**	**0.857**	0.897	0.919
LSA	0.551	0.650	0.751	0.834	0.891	0.933
WFIL	0.554	0.664	0.768	**0.849**	**0.906**	**0.944**

Table 2
Comparison of denoising performance in terms of STOI scores for different input SNRs (*F16 noise*). For a given input SNR, algorithms whose scores are shown in boldface were found to perform equally well. The other algorithms have an inferior performance.

SNR (dB)	−5	0	5	10	15	20
Input	0.550	0.669	0.782	0.870	0.929	0.965
MSE	0.570	0.705	**0.820**	0.899	0.947	**0.974**
log MSE	0.554	0.691	0.815	**0.903**	**0.950**	**0.974**
WE	0.562	0.695	0.815	**0.902**	**0.950**	**0.975**
COSH	0.569	0.705	**0.823**	**0.905**	**0.950**	**0.975**
IS	0.562	0.689	0.808	0.899	**0.949**	**0.974**
IS-II	0.567	0.694	0.809	0.898	**0.948**	**0.974**
WCOSH	0.568	0.689	0.802	0.893	0.946	0.973
BNMF	**0.592**	**0.719**	**0.819**	0.884	0.916	0.933
LSA	0.571	0.681	0.784	0.871	0.927	0.960
WFIL	0.568	0.696	0.806	0.888	0.939	0.970

6.1.2. *F16 noise*

Fig. 5 (a) compares the SSNR gains in *F16 noise*. The trends are similar to the *White noise* scenario. The PESQ scores are shown in Fig. 5(b). For input SNRs between −5 and 20 dB, the PESQ gain is maximum for WE, IS, and IS-II, followed by log-MSE and WCOSH, whereas for SNRs below 5 dB, LSA also shows a high PESQ gain. In terms of SDR, PROSE is better than the other techniques for several choices of the distortion measure (cf. Fig. 5(c)). The SDR trends of PROSE are similar for both *White noise* and *F16 noise*. In terms of SSNR, PESQ, and SDR, the denoising performance of PROSE based on most distortion measures is better than the benchmark techniques. In terms of STOI, for SNRs below 5 dB, BNMF is better (cf. Table 2), whereas for SNRs greater than 5 dB, PROSE is better.

6.1.3. *Street noise*

The denoising results are presented in Fig. 6. The PROSE framework based on WCOSH, WE, IS, and IS-II distortions results in a higher SSNR gain than the other algorithms (cf. Fig. 6(a)). Similar to the *F16 noise* scenario, the margin of improvement in case of PROSE increases with increase in input SNR than the benchmark algorithms. The PESQ gain shows a slightly different trend (Fig. 6(b)). BNMF gives a marginally higher PESQ gain (about 0.05 higher) than PROSE. This is attributed to the training phase in BNMF, which is particularly advantageous in low SNR and nonstationary noise conditions. Within the PROSE family of denoisers, distortion measures other than the MSE result in a higher PESQ score. For all input SNRs, PROSE techniques (except MSE) show a superior denoising performance compared with the benchmark techniques in terms of SDR (cf. Fig. 6(c)). For input SNR in the range 0 to 10 dB,

[3] Matlab implementations of the LSA and WFIL used in Hu and Loizou (2007) are available in the CD-ROM accompanying (Loizou, 2007), which was used in the performance comparisons.

[4] Matlab implementation available at http://bass-db.gforge.inria.fr/bss_eval/.

[5] Matlab implementation available at http://siplab.tudelft.nl/.

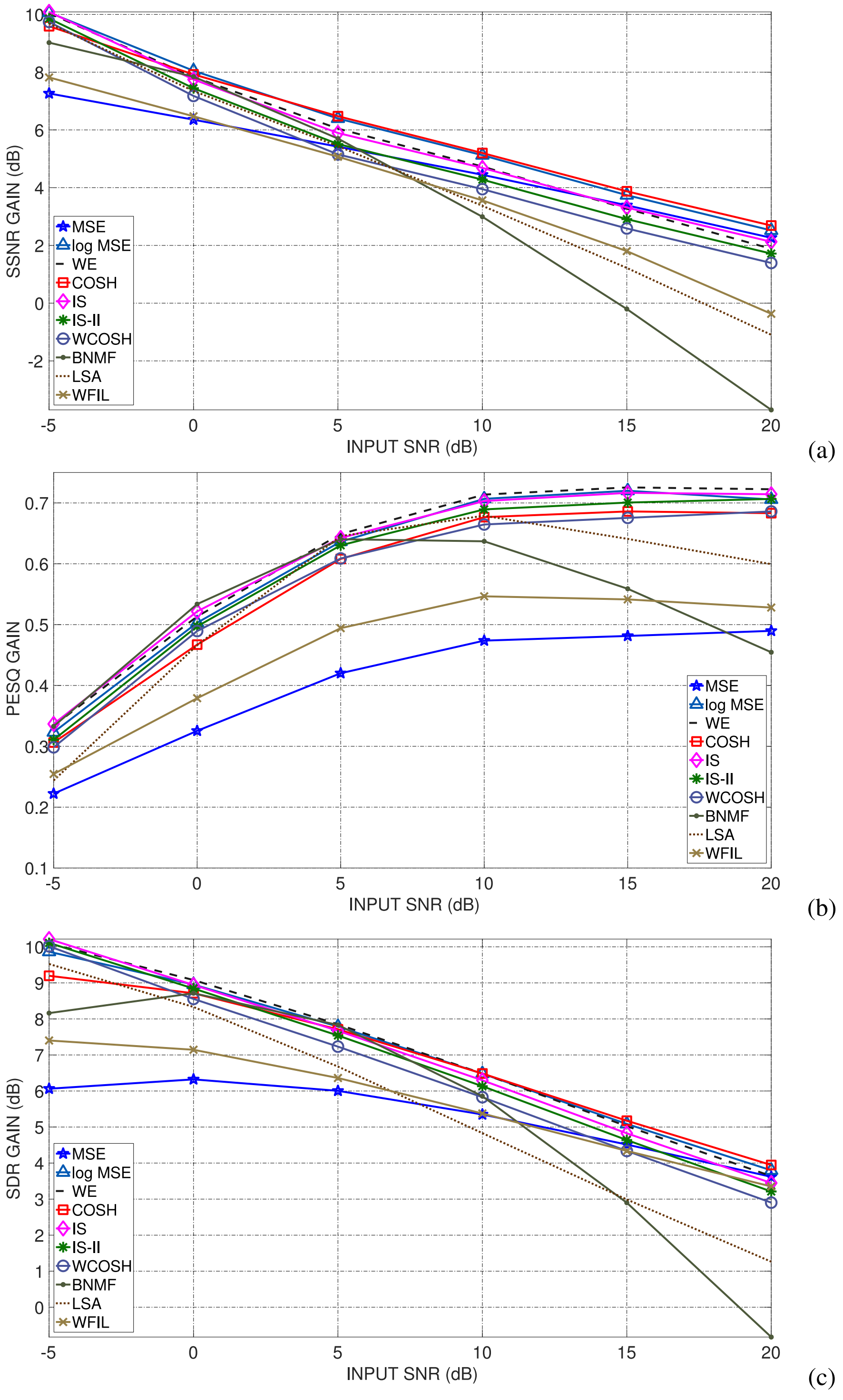


Fig. 4. [Color online] A comparison of denoising performance in *White noise*: (a) SSNR gain; (b) PESQ gain; and (c) SDR gain.

the denoising performance of BNMF is comparable in terms of SDR, but not superior to the best performing PROSE techniques. Table 3 shows the STOI scores. We observe that, for SNRs greater than 5 dB, PROSE algorithms yields a higher STOI score than BNMF, WFIL, and LSA.

6.1.4. *Train noise*

Fig. 7 compares the denoising performance in *Train noise*. Similar to the *Street noise* scenario, PROSE denoisers based on WCOSH, IS-II, and WE yield a higher SSNR gain than BNMF, WFIL, and LSA (cf. Fig. 7(a)). Although the SSNR gain trends are similar to the *Street noise* scenario, the margin of improvement is higher in *Train noise*. This may be because *Street noise* is comparatively more nonstationary than *Train noise*, which may have resulted in less accurate estimates of the noise standard deviation. From Fig. 7(b), we observe that for input SNRs greater than 5 dB, PROSE denoisers with WCOSH and IS-II measures are better than all the other methods. For input SNRs lower than 5 dB, LSA gives a PESQ gain about 0.05 higher than the rest. PESQ gains are also higher in case of *Train noise* than *Street noise*. The denoising performance measured in terms of SDR is shown in Fig. 7(c) — the performance trends are similar to the *Street noise* scenario. Table 4 shows the corresponding STOI scores and the trends are similar to the *Street noise* scenario.

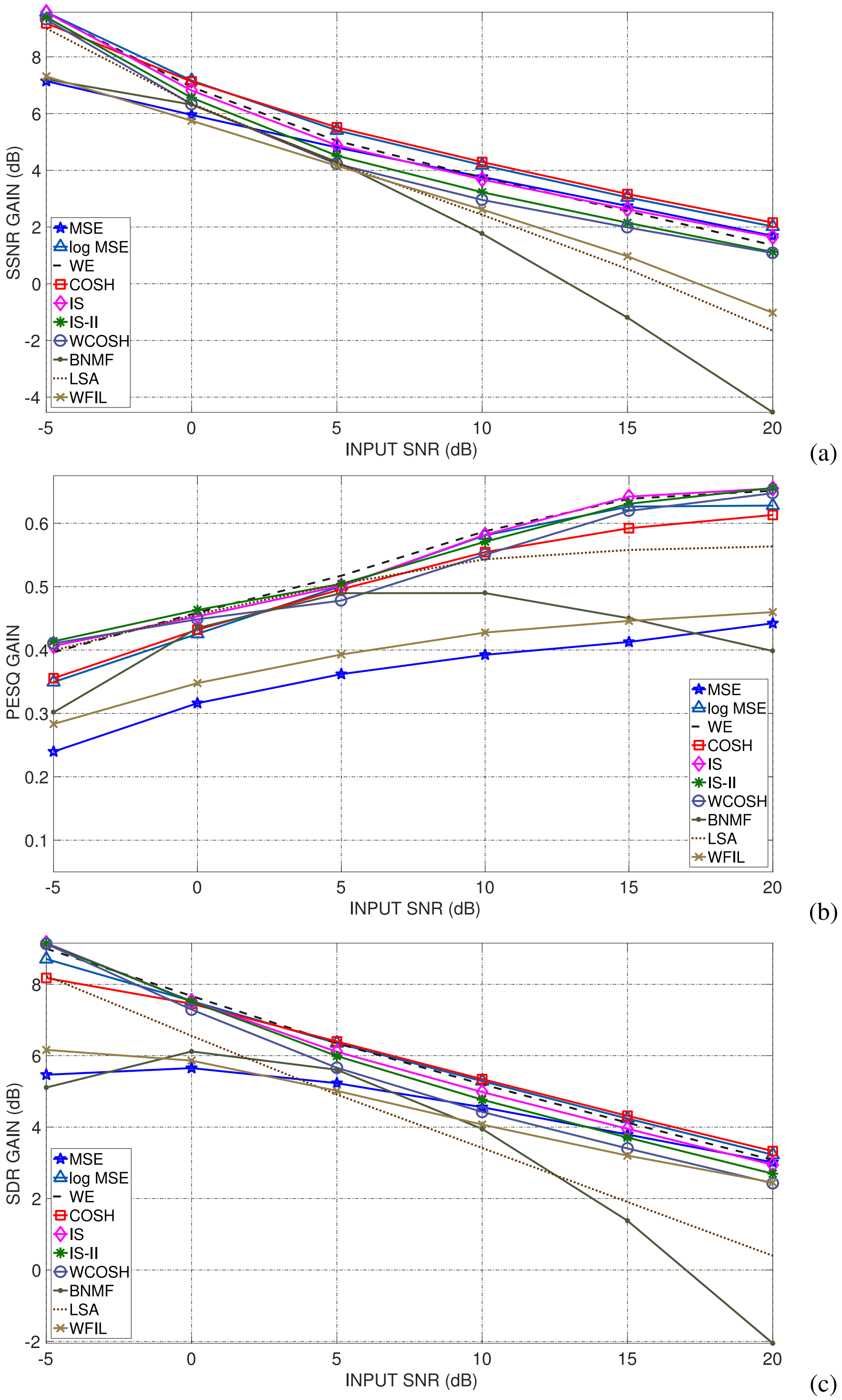


Fig. 5. [Color online] A comparison of denoising performance in *F16 noise*: (a) SSNR gain; (b) PESQ gain; and (c) SDR gain.

To summarize, the PROSE denoisers based on WCOSH, IS-II, IS, and WE show consistently superior denoising performance in terms of SSNR compared with LSA, WFIL, and BNMF. The margin of improvement over LSA, WFIL, and BNMF also increases with input SNR. Within the PROSE family of denoisers, the PESQ gains are higher for perceptual measures than for MSE. In terms of SDR, PROSE denoisers are superior to the benchmark techniques. For nonstationary noises such as the *Street noise*, the BNMF technique is marginally better than PROSE at low SNRs, which may be attributed to the training process and the use of the clean speech dictionary. For SNRs less than 0 dB, the PESQ gain offered by PROSE is not significant, which may be due to the errors in estimating the noise variance in low SNR conditions. For input SNRs greater than 5 dB, the PROSE methodology and WFIL give consistently better STOI scores, unlike LSA and BNMF approaches. The entire repository of denoised speech files under various noise conditions is available online at: http://spectrumee.wix.com/prose.

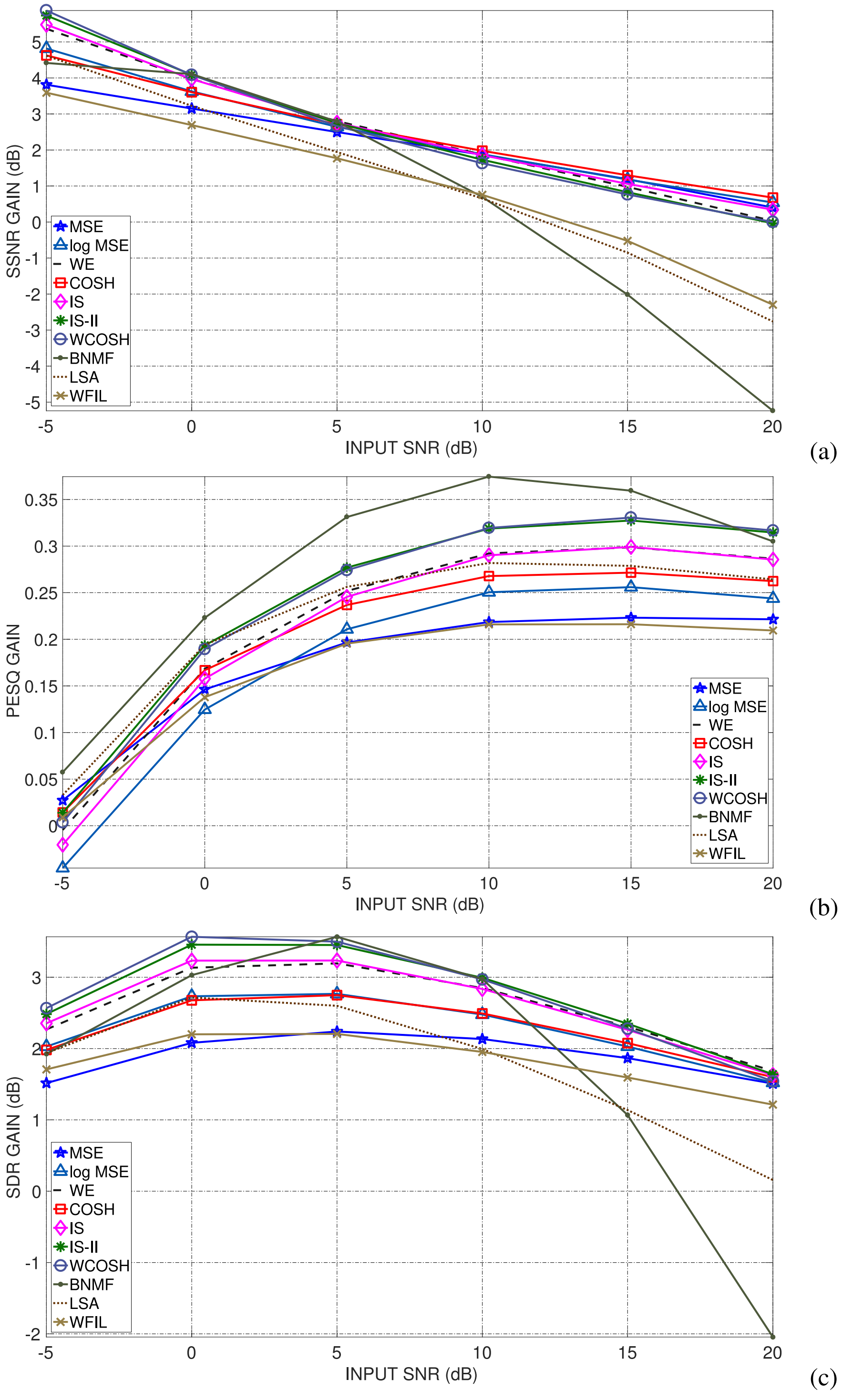


Fig. 6. [Color online] A comparison of denoising performance in *Street noise*: (a) SSNR gain; (b) PESQ gain; and (c) SDR gain.

6.2. Subjective Evaluation

We consider four speech files (two male and two female speakers) at SNRs 0, 10, and 20 dB. Fifteen listeners in the age group of 20 – 35 years endowed with normal hearing were selected for the listening test. The subjects were given a Sennheiser HD650 headphone for listening, and were asked to rate the enhanced speech signal based on the ITU-T recommended P.835 scale (ITU-T Rec. P.835, 2003):

(i) Speech signal distortion (SIG); 1: very distorted, 2: fairly distorted, 3: somewhat distorted, 4: little distorted, 5: not distorted;

(ii) Background intrusiveness (BAK); 1: very intrusive, 2: somewhat intrusive, 3: noticeable but not intrusive, 4: somewhat noticeable, 5: not noticeable; and

(iii) Overall quality (OVRL); 1: bad, 2: poor, 3: fair, 4: good, 5: excellent.

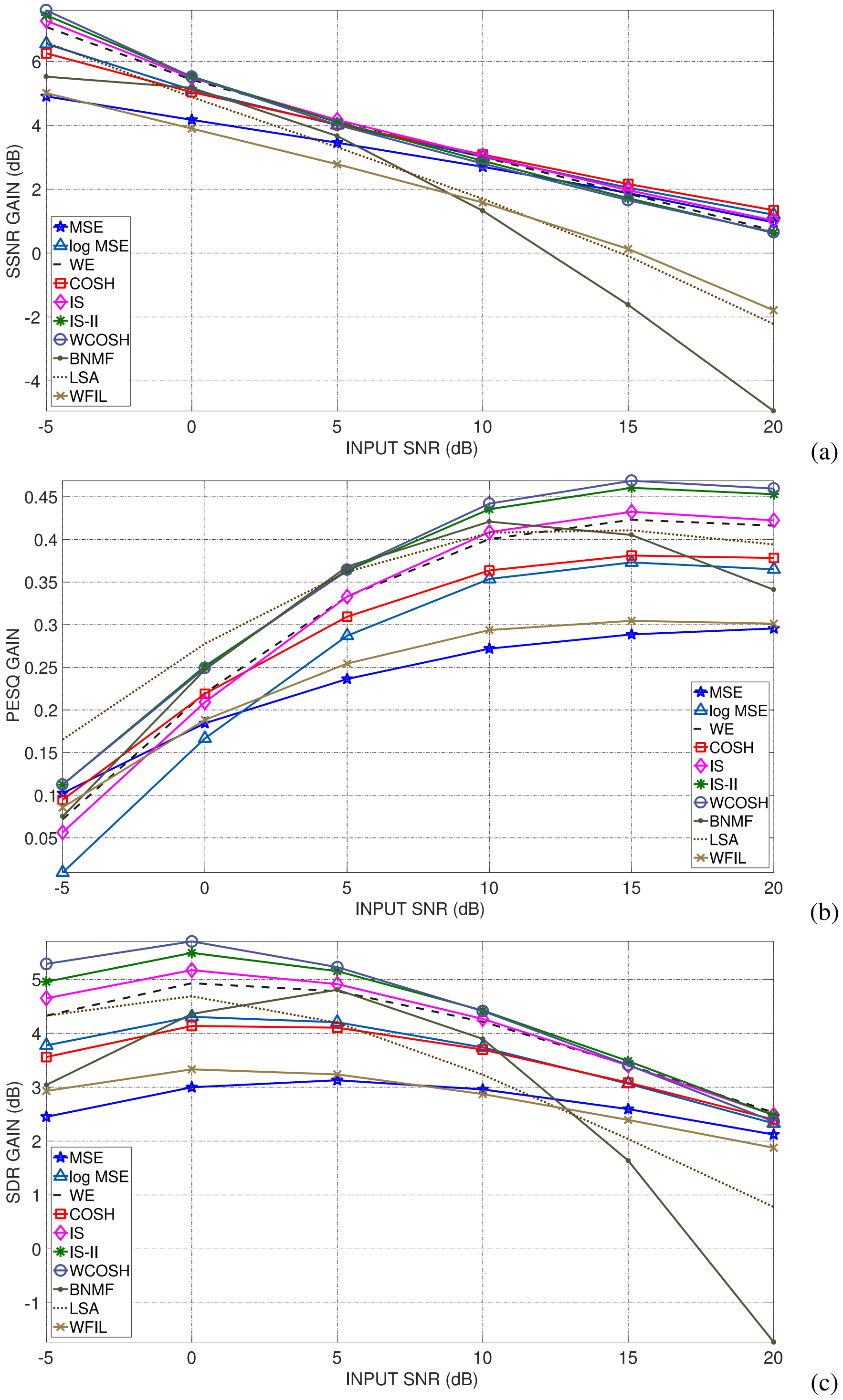


Fig. 7. [Color online] A comparison of denoising performance in *Train noise*: (a) SSNR gain; (b) PESQ gain; and (c) SDR gain.

The listening tests were conducted in four sessions, one for each noise type: *White* noise, *F16* noise, *Street* noise, *Train* noise. Following the ITU-T recommendation, each session was further divided into two subsessions. In the first one, two files (one male and one female speaker) were used with the rating order as SIG – BAK – OVRL and in the second one, the other two files (again one male and one female speaker) are presented and the order of rating was BAK – SIG – OVRL. The two-session test is done to suppress listener's bias. In each subsession, the listeners rated the denoising performance of all the algorithms. Thus, in one subsession, a listener had to grade a total of 2 (speakers) × 3 (SNRs) × 11 (algorithms) = 66 files. The subjects were given sufficient time to relax within and across subsessions, and across SNRs. The denoised speech files were presented in a random order, and the scores were derandomized accordingly before calculating the average scores. Fig. 8 shows the mean scores of SIG, BAK, and OVRL for *White noise, F16 noise, Street noise*, and *Train noise*.

We observe that, among the PROSE algorithms, WCOSH exhibits a consistently higher denoising performance in terms of SIG, BAK, and

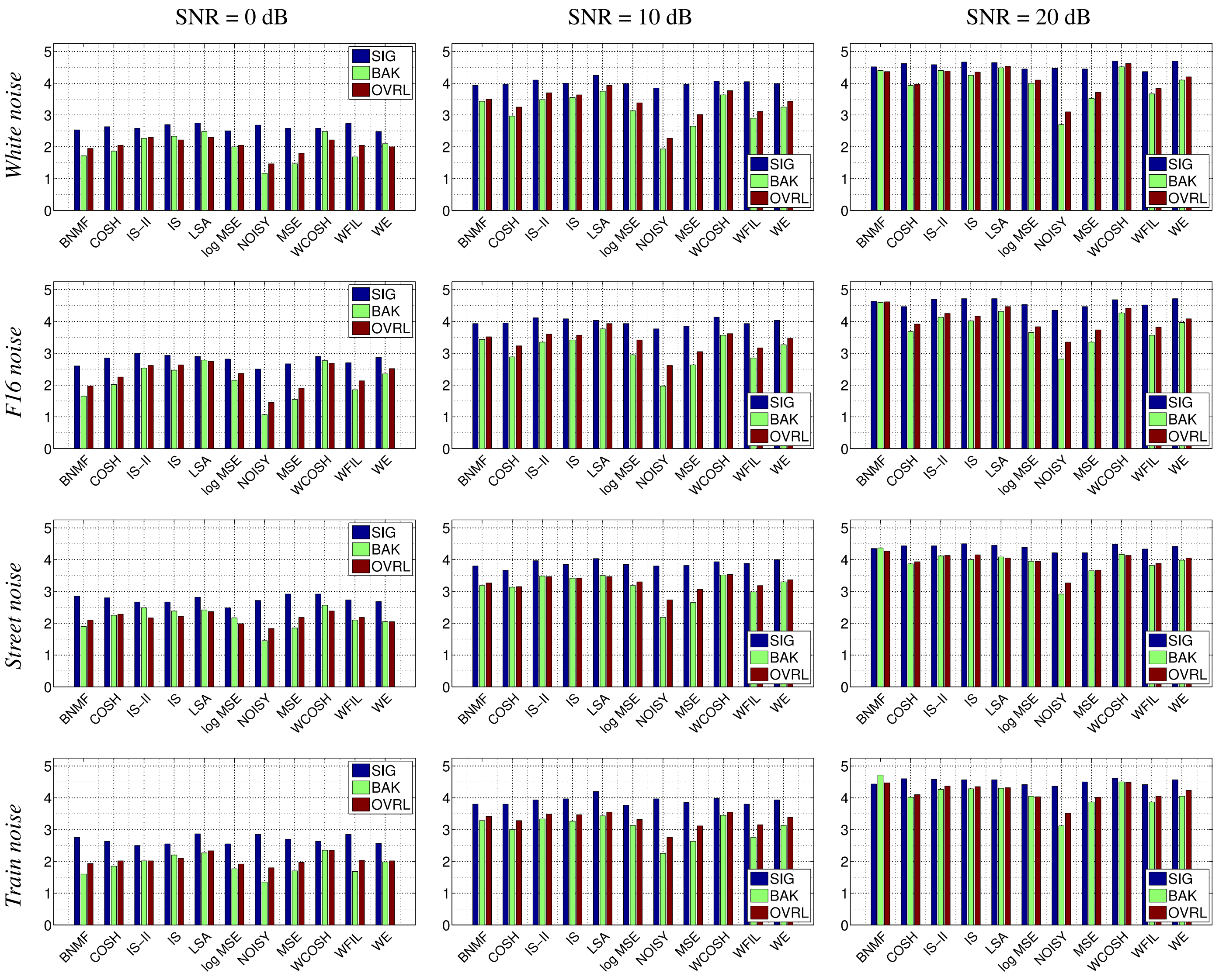


Fig. 8. [Color online] A comparison of the mean values of SIG, BAK, and OVRL ratings of the enhanced signal.

Table 3
Comparison of denoising performance in terms of STOI scores for different input SNRs (*Street noise*). For a given input SNR, algorithms whose scores are shown in boldface were found to perform equally well. The other algorithms have an inferior performance.

SNR (dB)	−5	0	5	10	15	20
Input	**0.545**	0.666	0.781	0.872	0.933	0.967
MSE	**0.543**	**0.676**	**0.796**	**0.886**	**0.942**	**0.973**
log MSE	0.518	0.658	0.785	**0.883**	**0.942**	**0.973**
WE	0.522	0.660	0.787	**0.883**	**0.942**	**0.973**
COSH	0.531	0.669	**0.792**	**0.885**	**0.943**	**0.973**
IS	0.517	0.655	0.782	**0.881**	**0.942**	**0.973**
IS-II	0.515	0.658	0.784	**0.881**	**0.941**	**0.973**
WCOSH	0.522	0.654	0.779	0.878	**0.940**	**0.972**
BNMF	**0.549**	**0.679**	**0.792**	0.869	0.912	0.933
LSA	0.518	0.643	0.763	0.859	0.924	0.962
WFIL	0.534	0.665	0.784	0.876	0.935	0.969

Table 4
Comparison of denoising performance in terms of STOI scores for various input SNRs (*Train noise*). For a given input SNR, algorithms whose scores are shown in boldface were found to perform equally well. The other algorithms have an inferior performance.

SNR (dB)	−5	0	5	10	15	20
Input	**0.559**	0.686	0.801	0.887	0.944	0.976
MSE	**0.557**	**0.699**	**0.817**	**0.900**	**0.951**	**0.979**
log MSE	0.524	0.682	**0.813**	**0.899**	**0.950**	**0.978**
WE	0.528	0.684	**0.815**	**0.901**	**0.951**	**0.978**
COSH	0.542	**0.693**	**0.818**	**0.901**	**0.951**	**0.979**
IS	0.520	0.677	**0.811**	**0.899**	**0.950**	**0.978**
IS-II	0.528	0.680	**0.812**	**0.900**	**0.951**	**0.978**
WCOSH	0.523	0.675	0.807	**0.898**	**0.950**	**0.977**
BNMF	**0.553**	**0.693**	0.808	0.881	0.919	0.937
LSA	0.526	0.666	0.789	0.877	0.935	0.968
WFIL	0.544	0.685	0.804	0.890	0.945	0.975

OVRL scores compared with the other algorithms. Among the techniques considered for benchmarking, LSA shows a consistently higher performance. The performances of LSA and PROSE with WCOSH are comparable. Also, PROSE consistently improves the BAK and OVRL scores compared with the noisy signal. Listening results reveal that, for all the algorithms considered, the extent of improvement in SIG scores is not significant compared with the improvement in BAK and OVRL scores. Within the PROSE family, WCOSH, IS-II, and IS show high BAK and

Table 5
Results obtained from the statistical analysis of BAK scores. Algorithms indicated using a ⋄, symbol were found to perform equally well.

Noise type	Input SNR	MSE	log MSE	WE	COSH	IS	IS-II	WCOSH	BNMF	LSA	WFIL
White noise	0 dB			⋄,		⋄,	⋄,	⋄,		⋄,	
	10 dB					⋄,	⋄,	⋄,	⋄,	⋄,	
	20 dB					⋄,	⋄,	⋄,	⋄,	⋄,	
F16 noise	0 dB					⋄,	⋄,	⋄,		⋄,	
	10 dB					⋄,		⋄,	⋄,	⋄,	
	20 dB							⋄,	⋄,	⋄,	
Train noise	0 dB			⋄,		⋄,	⋄,	⋄,	⋄,	⋄,	
	10 dB		⋄,	⋄,		⋄,	⋄,	⋄,	⋄,	⋄,	
	20 dB							⋄,	⋄,		
Street noise	0 dB		⋄,	⋄,	⋄,	⋄,	⋄,	⋄,		⋄,	⋄,
	10 dB		⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	
	20 dB						⋄,	⋄,	⋄,	⋄,	

Table 6
Results obtained from the statistical analysis of SIG scores. Algorithms indicated using a ⋄, symbol were found to perform equally well.

Noise type	Input SNR	MSE	log MSE	WE	COSH	IS	IS-II	WCOSH	BNMF	LSA	WFIL
White noise	0 dB	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,
	10 dB	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,
	20 dB	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,
F16 noise	0 dB	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,
	10 dB	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,
	20 dB	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,
Train noise	0 dB	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,
	10 dB	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,
	20 dB	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,
Street noise	0 dB	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,
	10 dB	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,
	20 dB	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,

Table 7
Results obtained from the statistical analysis of OVRL scores. Algorithms indicated using a ⋄, symbol were found to perform equally well.

Noise type	Input SNR	MSE	log MSE	WE	COSH	IS	IS-II	WCOSH	BNMF	LSA	WFIL
White noise	0 dB		⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,
	10 dB					⋄,	⋄,	⋄,		⋄,	
	20 dB					⋄,	⋄,	⋄,	⋄,	⋄,	
F16 noise	0 dB		⋄,	⋄,		⋄,	⋄,	⋄,		⋄,	
	10 dB					⋄,	⋄,	⋄,	⋄,	⋄,	
	20 dB							⋄,	⋄,	⋄,	
Train noise	0 dB	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,
	10 dB		⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,
	20 dB			⋄,		⋄,	⋄,	⋄,	⋄,	⋄,	
Street noise	0 dB	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,
	10 dB		⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	
	20 dB	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,	⋄,

OVRL scores, whereas MSE results in a lower score, compared with the other algorithms.

To assess the statistical significance of the average score obtained by the highest performing algorithm, we conducted multiple paired comparisons between the algorithm that possesses the highest average score against all the other algorithms based on Tukey's honestly significant difference (HSD) test (Ott and Longnecker, 2016, Chapter 9), as suggested in ITU-T Rec. P.835 (2003). The confidence level considered in the statistical test is 95%. Tables 5–7 show the results of the analysis corresponding to BAK, SIG, and OVRL scores, respectively, where the algorithms that perform equally well are indicated by the ⋄, symbol. We observe that, for all three SNR conditions and for all four noise types, among the algorithms compared, WCOSH shows a consistently higher denoising performance in terms of BAK score. In most listening scenarios, IS, IS-II, LSA, and BNMF also show comparable performance (cf. Table 5). In the case of SIG scores, the average scores did not exhibit any statistically significant difference in performance (cf. Table 6). We observe that, among the algorithms compared, WCOSH and LSA perform equally well and exhibit the highest denoising performance in terms of OVRL scores, followed by IS, IS-II, and BNMF (cf. Table 7).

BNMF is robust at low SNRs for nonstationary noise, which is mainly due to the intensive training phase and the use of the clean speech dictionary. On the other hand, PROSE does not require training and does not rely on a clean speech dictionary.

We would like to emphasize that the PROSE estimators are relatively simpler functions of a posteriori SNR, and require much

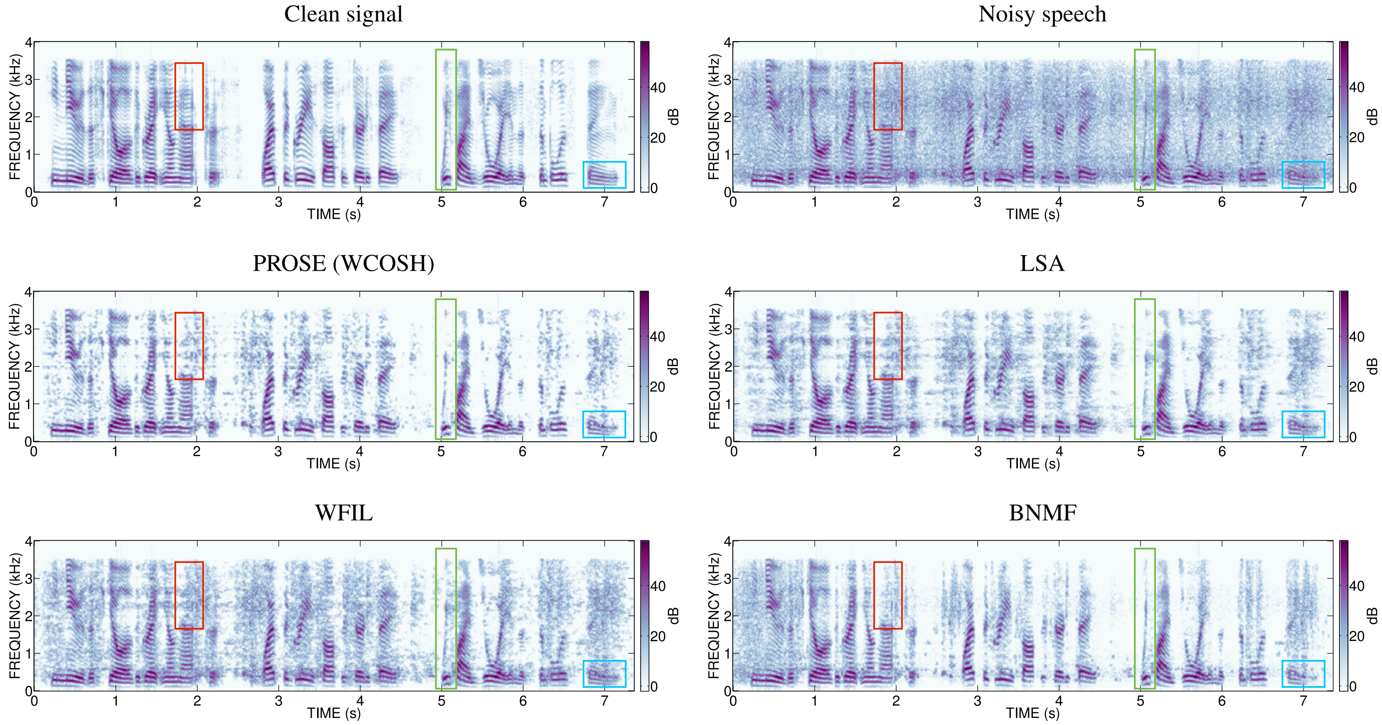


Fig. 9. [Color online] Spectrograms of clean speech, noisy speech, and speech denoised using various approaches. The utterance is: "He knew the skill of the great young actress. Wipe the grease off his dirty face. We find joy in the simplest things."

less computation compared with the denoising functions proposed for LSA (cf. Eq. (20) in Ephraim and Malah, 1985) and BNMF (cf. Eq. (10) in Mohammadiha et al., 2013).

6.3. Spectrograms

Fig. 9 shows the spectrograms of clean, noisy (SNR = 10 dB), and enhanced signals for the case of train noise. We observe that WFIL and LSA result in significant residual noise, whereas BNMF has relatively lower residual noise. BNMF recovers clean speech spectra with less distortion in some regions, compared with the other algorithms (cf. green box), but it introduces distortions in the other parts (highlighted by the blue box, for instance). PROSE with WCOSH results in superior noise suppression with minimal speech distortion than WFIL, LSA, and BNMF. Both PROSE and BNMF introduce a small amount of musical noise, in particular, in the silence regions. Regions that are submerged in noise are difficult to recover by any algorithm (red box, for instance), which explains why the improvements in SIG scores are lower than improvements in BAK and OVRL for all the techniques.

7. Conclusion

We introduced the notion of unbiased risk estimation within a perceptual framework (abbreviated PROSE) for performing single-channel speech enhancement. The analytical developments are based on Stein's lemma and its generalized version introduced in this paper, which proved to be efficient for obtaining unbiased estimates of the distortion measures. We have also established the optimality of the shrinkage parameters considering the Karush–Kuhn–Tucker conditions. Validation on several speech signals in real-world nonstationary noise scenarios, and comparisons with benchmark techniques showed that, for input SNR greater than 5 dB, the proposed PROSE method results in better denoising performance. Within the PROSE family, estimators based on Itakura-Saito distortion and weighted cosh distortion resulted in superior denoising performance. Among the risk estimation based techniques, the quality of the denoised speech is higher (measured in terms of PESQ and subjective listening scores) for perceptual risk-based techniques than the MSE based one.

Performance evaluation of various denoising techniques by Loizou (cf. Loizou, 2007, Chapter 11) revealed that the best performing algorithm in terms of speech quality (evaluated based on SIG, BAK, and OVRL scores) may not be the best performing one in terms of intelligibility for listeners with normal hearing. For high SNR conditions (greater than 5 dB), the proposed risk estimation based speech denoising algorithms exhibit a higher denoising performance both in terms of speech quality and intelligibility (measured in terms of STOI).

The PROSE methodology is ideal from an implementation perspective because the shrinkage estimators are easy to compute. It also does not require a training phase, thus making it ideal for deployment in practical applications particularly those involving hearing aids, mobile devices, etc.

To estimate and update the noise statistics, we employed a voice-activity detector, which comes into action only when a frame contains only noise. A noise estimation technique that continually updates the noise statistics even in speech regions could be used to further enhance the performance of the PROSE approach. The parameter α, which controls the amount of speech distortion and noise attenuation could be optimized for each DCT subband separately based on psycho-acoustic criteria.

The PROSE framework proposed in this paper could be integrated with state-of-the-art deep learning approaches to develop unsupervised denoising strategies. Such an attempt was recently shown to be successful in the context of image denoising (Metzler et al., 2018). It would be possible to develop a similar strategy for speech denoising as well — this is a research problem in its own right and requires a separate investigation.

Declaration of Competing Interest

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

Acknowledgments

We would like to thank Subhadip Mukherjee for technical discussions and the subjects for participating in the listening test. We would also like to thank (Mohammadiha et al., 2013) and (Taal et al., 2011) for providing the Matlab implementations of their algorithms online, which facilitated the comparisons.

Appendix A. Karush–Kuhn–Tucker Conditions

The goal is to solve the optimization problem

$$\min_{a_k} \widehat{\mathcal{R}} \text{ subject to } a_k \in [0, 1].$$

The corresponding Lagrangian is $\mathbf{L}(a_k, \lambda_1, \lambda_2) = \widehat{\mathcal{R}} + \lambda_1(a_k - 1) - \lambda_2 a_k$, where $\lambda_1, \lambda_2 \in \mathbb{R}^+ \cup \{0\}$. The KKT conditions are as follows:

$$\text{C1}: \ \frac{\mathrm{d}\mathbf{L}\big(a_k, \lambda_1, \lambda_2\big)}{\mathrm{d}a_k} = 0, \tag{A.1a}$$

$$\text{C2}: \ \lambda_1(a_k - 1) = 0, -\lambda_2 a_k = 0, \tag{A.1b}$$

$$\text{C3}: \ a_k \in [0, 1], \tag{A.1c}$$

$$\text{C4}: \ \lambda_1 \geq 0, \lambda_2 \geq 0, \text{ and} \tag{A.1d}$$

$$\text{C5}: \ \frac{\mathrm{d}^2\mathbf{L}\big(a_k, \lambda_1, \lambda_2\big)}{\mathrm{d}a_k^2} > 0. \tag{A.1e}$$

Solving (A.1a) –(A.1e) gives the optimum shrinkage parameter a_k. The derivations for all the distortion measures are provided in the supporting document.

Appendix B. The High-SNR Scenario

The a priori SNR is defined as SNR $:= \frac{S_k{}^2}{\sigma^2}$. By high SNR, we mean that SNR $> 4c^2$. Consider the probability:

$$\begin{aligned}
\text{Prob}\big\{|W_k| < |X_k|\big\} &= \text{Prob}\big\{X_k{}^2 - W_k{}^2 > 0\big\},\\
&= \text{Prob}\big\{(X_k + W_k)(X_k - W_k) > 0\big\},\\
&= \text{Prob}\big\{S_k(X_k + W_k) > 0\big\},\\
&= \text{Prob}\left\{W_k < -\frac{S_k}{2}\right\}, \text{if} \quad S_k < 0\\
&= \text{Prob}\left\{W_k > -\frac{S_k}{2}\right\}, \text{if} \quad S_k > 0.
\end{aligned} \tag{B.1}$$

Therefore, $\text{Prob}\{|W_k| < |X_k|\} = \text{Prob}\{W_k < \frac{|S_k|}{2}\} = 1$ if $|S_k| > 2c\sigma \Rightarrow$ a priori SNR $> 4c^2$.

Appendix C. Weighted Euclidean Distance

Consider

$$\mathcal{E}\left\{a_k^2 X_k \sum_{n=0}^{4}\left(\frac{W_k}{X_k}\right)^n\right\} = a_k^2 \mathcal{E}\left\{X_k + W_k + \frac{W_k^2}{X_k} + \frac{W_k^3}{X_k^2} + \frac{W_k^4}{X_k^3}\right\}. \tag{C.1}$$

Using Lemma 4, we have that

$$\begin{aligned}
&\mathcal{E}\left\{\frac{W_k^2}{X_k}\right\} = \sigma^4 \mathcal{E}\left\{\frac{2}{X_k^3}\right\} + \sigma^2 \mathcal{E}\left\{\frac{1}{X_k}\right\},\\
&\mathcal{E}\left\{\frac{W_k^3}{X_k^2}\right\} = \sigma^6 \mathcal{E}\left\{\frac{-24}{X_k^5}\right\} + \sigma^4 \mathcal{E}\left\{\frac{-6}{X_k^3}\right\}, \text{and}\\
&\mathcal{E}\left\{\frac{W_k^4}{X_k^3}\right\} = \sigma^8 \mathcal{E}\left\{\frac{360}{X_k^7}\right\} + \sigma^6 \mathcal{E}\left\{\frac{72}{X_k^5}\right\} + 3\sigma^4 \mathcal{E}\left\{\frac{1}{X_k^3}\right\}.
\end{aligned}$$

Substituting the preceding expressions in (C.1), we get

$$\mathcal{E}\left\{a_k^2 X_k \sum_{n=0}^{4}\left(\frac{W_k}{X_k}\right)^n\right\} = a_k^2 \mathcal{E}\left\{X_k + \frac{\sigma^2}{X_k} - \frac{\sigma^4}{X_k^3} + 48\frac{\sigma^6}{X_k^5} + 360\frac{\sigma^8}{X_k^7}\right\}.$$

Appendix D. Log Mean-Square Error

Consider $\mathcal{E}\left\{\sum_{n=1}^{4} \frac{\log a_k X_k}{n}\left(\frac{W_k}{X_k}\right)^n\right\} = \mathcal{E}\left\{\sum_{n=1}^{4} J_n(X_k) W_k^n\right\}$,

where $J_n(X_k) = \log a_k X_k/(nX_k^n)$. Applying Lemma 4 gives

$$\begin{aligned}
&\mathcal{E}\{J_1(X_k)W_k\} = \sigma^2 \mathcal{E}\left\{J_1^{(1)}(X_k)\right\},\\
&\mathcal{E}\{J_2(X_k)W_k^2\} = \sigma^4 \mathcal{E}\left\{J_2^{(2)}(X_k)\right\} + \sigma^2 \mathcal{E}\{J_2(X_k)\},\\
&\mathcal{E}\{J_3(X_k)W_k^3\} = \sigma^6 \mathcal{E}\left\{J_3^{(3)}(X_k)\right\} + 3\sigma^4 \mathcal{E}\left\{J_3^{(1)}(X_k)\right\}, \text{and}\\
&\mathcal{E}\{J_4(X_k)W_k^4\} = \mathcal{E}\left\{\sigma^8 J_4^{(4)}(X_k) + 6\sigma^6 J_4^{(2)}(X_k) + 3\sigma^4 J_4(X_k)\right\},\\
&\text{where } J_1^{(1)} = \frac{1}{X_k^2} - \frac{\log a_k X_k}{X_k^2}, J_2^{(2)} = \frac{-5}{2X_k^4} + \frac{3\log a_k X_k}{X_k^4},\\
&J_3^{(1)} = \frac{1}{3X_k^4} - \frac{\log a_k X_k}{X_k^4}, J_3^{(3)} = \frac{47}{3X_k^6} - \frac{20\log a_k X_k}{X_k^6},\\
&J_4^{(2)} = \frac{-9}{4X_k^6} + \frac{5\log a_k X_k}{X_k^6}, \text{and } J_4^{(4)} = -\frac{638}{4X_k^8} + \frac{210\log a_k X_k}{X_k^8}.
\end{aligned}$$

Appendix E. Itakura–Saito Distortion

With reference to (13), consider the truncated approximation:

$$\sum_{n=0}^{\infty} \mathcal{E}\{a_k W_k^n / X_k^n\} \approx \sum_{n=0}^{4} \mathcal{E}\{a_k W_k^n / X_k^n\}. \tag{E.1}$$

Applying Lemma 4, we have that

$$\begin{aligned}
&\mathcal{E}\left\{\frac{W_k}{X_k}\right\} = \sigma^2 \mathcal{E}\left\{\frac{-1}{X_k^2}\right\}, \mathcal{E}\left\{\frac{W_k^2}{X_k^2}\right\} = \mathcal{E}\left\{\sigma^4 \frac{6}{X_k^4} + \sigma^2 \frac{1}{X_k^2}\right\},\\
&\mathcal{E}\left\{\frac{W_k^3}{X_k^3}\right\} = \sigma^6 \mathcal{E}\left\{\frac{-60}{X_k^6}\right\} + 3\sigma^4 \mathcal{E}\left\{\frac{-3}{X_k^4}\right\}, \text{and}\\
&\mathcal{E}\left\{\frac{W_k^4}{X_k^4}\right\} = \sigma^8 \mathcal{E}\left\{\frac{840}{X_k^8}\right\} + 6\sigma^6 \mathcal{E}\left\{\frac{20}{X_k^6}\right\} + 3\sigma^4 \mathcal{E}\left\{\frac{1}{X_k^4}\right\}.
\end{aligned}$$

The final expression for (E.1) is given by

$$\sum_{n=0}^{4} \mathcal{E}\{a_k W_k^n / X_k^n\} = a_k \mathcal{E}\left\{840\frac{\sigma^8}{X_k^8} + 60\frac{\sigma^6}{X_k^6} + 1\right\}. \tag{E.2}$$

Appendix F. Weighted COSH Distance

We provide certain simplifications for the expectation terms in the risk estimator for weighted cosh measure:

$$\mathcal{R} = \mathcal{E}\left\{d(S_k, \widehat{S}_k)\right\} = \frac{1}{2}\mathcal{E}\left\{\frac{1}{\widehat{S}_k} + \frac{\widehat{S}_k}{S_k^2}\right\} - \frac{1}{S_k}. \tag{F.1}$$

The second term in (F.1) is approximated as

$$\frac{\widehat{S}_k}{X_k^2}\left(1 - \frac{W_k}{X_k}\right)^{-2} \approx \frac{\widehat{S}_k}{X_k^2}\left(1 + 2\frac{W_k}{X_k} + 3\frac{W_k^2}{X_k^2} + 4\frac{W_k^3}{X_k^3} + 5\frac{W_k^4}{X_k^4}\right).$$

Substituting $\widehat{S}_k = a_k X_k$ and taking expectation, we get that

$$\mathcal{E}\left\{\frac{\widehat{S}_k}{S_k^2}\right\} = \mathcal{E}\left\{\frac{a_k X_k}{X_k^2}\left(1 + 2\frac{W_k}{X_k} + 3\frac{W_k^2}{X_k^2} + 4\frac{W_k^3}{X_k^3} + 5\frac{W_k^4}{X_k^4}\right)\right\}.$$

Simplified expressions for the individual terms in the above equation are given below:

$$\mathcal{E}\left\{\frac{W_k}{X_k^2}\right\} = \sigma^2\mathcal{E}\left\{\frac{-2}{X_k^3}\right\}, \mathcal{E}\left\{\frac{W_k^2}{X_k^3}\right\} = \mathcal{E}\left\{\sigma^4\frac{12}{X_k^5} + \sigma^2\frac{1}{X_k^3}\right\},$$

$$\mathcal{E}\left\{\frac{W_k^3}{X_k^4}\right\} = \sigma^6\mathcal{E}\left\{\frac{-120}{X_k^7}\right\} + 3\sigma^4\mathcal{E}\left\{\frac{-4}{X_k^5}\right\}, \text{and}$$

$$\mathcal{E}\left\{\frac{W_k^4}{X_k^5}\right\} = \sigma^8\mathcal{E}\left\{\frac{1680}{X_k^9}\right\} + 6\sigma^6\mathcal{E}\left\{\frac{30}{X_k^7}\right\} + 3\sigma^4\mathcal{E}\left\{\frac{1}{X_k^5}\right\}.$$

Substituting these expressions in (F.1), we get

$$\mathcal{R} = \mathcal{E}\left\{\frac{a_k}{2X_k}\left(1 - \frac{\sigma^2}{X_k^2} + 3\frac{\sigma^4}{X_k^4} + 420\frac{\sigma^6}{X_k^6} + 8400\frac{\sigma^8}{X_k^8}\right) + \frac{1}{2a_kX_k} - \frac{1}{S_k}\right\}. \tag{F.2}$$

The quantity inside the braces is an unbiased estimate of $\mathcal{R}$.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi 10.1016/j.specom.2019.11.001.

References

Atto, A.M., Pastor, D., Mercier, G., 2009. Smooth adaptation by sigmoid shrinkage. EURASIP J. Image Video Process. 1–16.

Benazza-Benyahia, A., Pesquet, J.C., 2005. Building robust wavelet estimators for multicomponent images using stein's principle. IEEE Trans. Image Process. 14, 1814–1830.

Blu, T., Luisier, F., 2007. The SURE-LET approach to image denoising. IEEE Trans. Image Process. 16 (11), 2778–2786.

Blu, T., Luisier, F., 2008. SURE-LET multichannel image denoising: Interscale orthonormal wavelet thresholding. IEEE Trans. Image Process. 17 (4), 482–492.

Boll, S., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process. 27 (2), 113–120.

Burkardt, J., 2014. The Truncated Normal Distribution. Department of Scientific Computing Website, Florida State University.

Chen, J., Benesty, J., Huang, Y., Doclo, S., 2006. New insight into noise reduction wiener filter. IEEE Trans. Speech Audio Process. 14 (4), 1218–1234.

Dendrinos, M., Bakamidis, S., Carayannis, G., 1991. Speech enhancement from noise: A regenerative approach. Speech Commun. 10 (1), 45–57.

Deng, F., Bao, C., 2016. Speech enhancement based on AR model parameters estimation. Speech Commun. 79, 30–46.

Ephraim, Y., 1992. A bayesian estimation approach for speech enhancement using hidden markov models. IEEE Trans. Signal Process. 40 (4), 725–735.

Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum mean-squared error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. 32 (6), 1109–1121.

Ephraim, Y., Malah, D., 1985. Speech enhancement using a minimum mean-squared error log-spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. 33 (2), 443–445.

Ephraim, Y., Trees, H.L.V., 1995. A signal subspace approach for speech enhancement. IEEE Trans. Speech Audio Process. 3 (4), 251–266.

Erkelen, J.S., Hendriks, R.C., Heusdens, R., Jensen, J., 2007. Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors. IEEE Trans. Audio Speech Lang. Process. 15 (6), 1741–1752.

Eskimez, S.E., Soufleris, P., Duan, Z., Heinzelman, W., 2018. Front-end speech enhancement for commercial speaker verification systems. Speech Commun. 95, 101–113.

Fletcher, R., 1987. Practical Methods of Optimization, second John Wiley and Sons, New York.

Gao, T., Du, J., Dai, L.R., Lee, C.H., 2017. A unified DNN approach to speaker-dependent simultaneous speech enhancement and speech separation in low SNR environments. Speech Commun. 95, 28–29.

Gray, R.M., Buzo, A., Gray Jr., A.H., Matsuyama, Y., 1980. Distortion measures for speech processing. IEEE Trans. Acoust. Speech Signal Process. 28, 367–376.

Hansen, J.H.L., Clements, M.A., 1991. Constrained iterative speech enhancement with application to speech recognition. IEEE Trans. Signal Process. 39 (4), 795–805.

Hansen, P.S.K., Hansen, P.C., Hansen, S.D., Sorensen, J.A., 1995. Reduction of broad-band noise in speech by truncated QSVD. IEEE Trans. Speech Audio Process. 3 (6), 439–448.

Hansen, P.S.K., Hansen, P.C., Hansen, S.D., Sorensen, J.A., 1999. Experimental comparison of signal subspace based noise reduction methods. Proc. IEEE Int. Conf. Acoust. Speech Signal Process. 1, 101–104.

Hendricks, R.C., Heusdens, R., Jensen, J., 2007. An MMSE estimator for speech enhancement under a combined stochastic-deterministic speech model. IEEE Trans. Audio, Speech, Language Process. 15 (2), 406–415.

Hendriks, R.C., Gerkmann, T., Jensen, J., 2007. DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement: A Survey of the State-of-the-Art. Morgan & Claypool.

Hu, Y., Loizou, P., 2003. A perceptually motivated approach for speech enhancement. IEEE Trans. Speech Audio Process. 11 (5), 457–465.

Hu, Y., Loizou, P., 2004. Incorporating psycho-acoustical model in frequency domain speech enhancement. IEEE Signal Process. Lett. 11 (1), 270–273.

Hu, Y., Loizou, P., 2007. Subjective comparison and evaluation of speech enhancement algorithms. Speech Commun. 49, 588–601.

Huang, J., Zhao, Y., 1998. An energy-constrained signal subspace method for speech enhancement and recognition in colored noise. Speech Commun. 26 (3), 165–181.

ITU-T Rec., P.862, 2001. Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. Int. Telecommun. Union.

ITU-T Rec., P.835, 2003. Subjective test methodology for evaluating speech communication systems that include noise suppression algorithms. Int. Telecommun. Union.

Jabloun, F., Champagne, B., 2003. Incorporating the human hearing properties in the signal subspace approach for speech enhancement. IEEE Trans. Speech Audio Process. 11 (6), 700–708.

Kalantari, M., Gooran, S.R., Kanan, H.R., 2018. Improved embedded pre-whitening subspace approach for enhancing speech contaminated by colored noise. Speech Commun. 99, 12–26.

Kamath, S., Loizou, P., 2002. A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. Proc. IEEE Int. Conf. Acoust. Speech Signal Process. 4, 4164–4167.

Kundu, A., Chatterjee, S., Murthy, A.S., Sreenivas, T.V., 2008. GMM based bayesian approach to speech enhancement in signal/transform domain. Proc. IEEE Int. Conf. Acoust. Speech and Signal Process. 4893–4896.

Lim, J.S., Oppenheim, A.V., 1978. All-pole modeling of degraded speech. IEEE Trans. Acoust. Speech Signal Process. 26 (3), 197–210.

Lim, J.S., Oppenheim, A.V., 1979. Enhancement and bandwidth compression of noisy speech. Proc. IEEE 67 (12), 1586–1604.

Lockwood, P., Boudy, J., 1992. Experiments with a non-linear spectral subtractor (NSS), hidden Markov models and the projections, for robust speech recognition in cars. Speech Commun. 11 (2), 215–228.

Loizou, P., 2007. Speech Enhancement — Theory and Practice. CRC Press.

Loizou, P.C., 2005. Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum. IEEE Trans. Speech Audio Process. 13 (5), 857–869.

Lotter, T., Vary, P., 2005. Speech enhancement by maximum a posteriori estimation using super-gaussian speech model. EURASIP J. Appl. Signal Process. 7, 1110–1126.

Luisier, F., Blu, T., Unser, M., 2007. A new SURE approach to image denoising: interscale orthonormal wavelet thresholding. IEEE Trans. Image Process. 16, 593–606.

Mai, V.K., Pastor, D., Aïssa-El-Bey, A., Bidan, R.L., 2018. Semi-parametric joint detection and estimation for a speech enhancement based on minimum mean-square error. Speech Commun. 102, 27–38.

Martin, R., 2005. Speech enhancement based on minimum mean-square error estimation and superGaussian priors. EEE Trans. Speech Audio Process. 13 (5), 845–856.

McAulay, R.J., Malpass, M.L., 1980. Speech enhancement using a soft decision noise suppression filter. IEEE Trans. Acoust. Speech Signal Process. 28 (2), 137–145.

McCallum, M., Guillemin, B., 2013. Stochastic-deterministic MMSE STFT speech enhancement with general a priori information. IEEE Trans. Audio Speech Lang. Process. 21 (7), 1445–1457.

Metzler, C.A., Mousavi, A., Heckel, R., Baraniuk, R.G., 2018. Unsupervised learning with stein's unbiased risk estimator. [stat.ML]. arXiv: 1805.10531.

Mittal, U., Phamdo, N., 2000. Signal/noise KLT based approach for enhancing speech degraded by colored noise. IEEE Trans. Speech Audio Process. 8 (2), 159–167.

Mohammadiha, N., Smaragdis, P., Leijon, A., 2013. Supervised and unsupervised speech enhancement using nonnegative matrix factorization. IEEE Trans. Audio Speech Lang. Process. 21 (10), 2140–2151.

Mowlaee, P., Stahl, J., Kulmer, J., 2017. Iterative joint MAP single-channel speech enhancement given non-uniform phase prior. Speech Commun. 86, 85–96.

Muraka, N.R., Seelamantula, C.S., 2011. A risk-estimation-based comparison of mean-square error and itakura-saito distortion measures for speech enhancement. Proc. Interspeech 349–352.

Muraka, N.R., Seelamantula, C.S., 2012. A risk-estimation-based formulation for speech enhancement and its relation to wiener filtering. Proc. Int. Conf. Signal Process. Commun. 1–5.

Ott, R.L., Longnecker, M., 2016. An introduction to statistical methods and data analysis. CENGAGE Learn.. Seventh edition

Rezayee, A., Gazor, S., 2001. An adaptive KLT approach for speech enhancement. IEEE Trans. Speech Audio Process. 9 (2), 87–95.

Rosenkranz, T., Puder, H., 2012. Improving robustness of codebook-based noise estimation approaches with delta codebooks. IEEE Trans. Audio Speech Lang. Process 20 (4), 1177–1188.

Sadasivan, J., Seelamantula, C.S., 2016. A novel risk-estimation-theoretic framework for speech enhancement in nonstationary and non-gaussian noise conditions. Proc. Interspeech. 3718–3722.

Sadasivan, J., Seelamantula, C.S., 2016. An unbiased risk estimator for gaussian mixture noise distributions —application to speech denoising. Proc. IEEE Int. Conf. Acoust. Speech Signal Process. 4513–4517.

Scalart, P., Filho, J.V., 1996. Speech enhancement based on a priori signal to noise estimation. Proc. IEEE Int. Conf. Acoust. Speech Signal Process. 2, 629–632.

Sohn, J., Kim, N.S., Sung, W., 1999. A statistical model-based voice activity detection. IEEE Signal Process. Lett. 6 (1), 1–3.

Soon, I.Y., Koh, S.N., Yeo, C.K., 1998. Noisy speech enhancement using discrete cosine transform. Speech Commun. 24, 249–257.

Sreenivas, T.V., Kirnapure, P., 1996. Codebook constrained wiener filtering for speech enhancement. IEEE Trans. Speech Audio Process. 4 (5), 383–389.

Srinivasan, S., Samuelsson, J., Kleijn, W., 2006. Codebook driven short-term predictor parameter estimation for speech enhancement. IEEE Trans. Audio Speech Lang. Process. 14 (1), 163–176.

Srinivasan, S., Samuelsson, J., Kleijn, W., 2007. Codebook-based Bayesian speech enhancement for nonstationary environments. IEEE Trans. Audio Speech Lang. Process. 15 (2), 441–452.

Stein, C.M., 1981. Estimation of the mean of a multivariate normal distribution. Ann. Statist. 9 (6), 1135–1151.

Taal, C.H., Heusdens, R., Jensen, J., 2011. An algorithm for intelligibility prediction of time-frequency weighted noisy speech. IEEE Trans. Audio Speech Lang. Process. 19 (7), 2125–2136.

Tsao, Y., Lai, Y., 2016. Generalized maximum a posteriori spectral amplitude estimation for speech enhancement. Speech Commun. 76, 112–126.

Varga, A., Steeneken, H.J.M., 1993. Assessment for automatic speech recognition: ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems. Speech Commun. 12 (3), 247–251.

Vincent, E., Gribonval, R., Févotte, C., 2006. Performance measurement in blind audio source separation. IEEE Trans. Acoust. Speech Signal Process. 14, 1462–1469.

Weiss, M., Aschkensy, E., Parson, T., 1974. Study and the development of the INTEL techniques for improving speech intelligibility. Technical Report NSC-FR/4023. Nicolet Scientific Corporation.

Xia, B., Bao, C., 2014. Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification. Speech Commun. 60, 13–29.

Xu, Y., Du, J., Dai, L.R., Lee, C.H., 2015. A regression approach to speech enhancement based on deep neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. 23 (1), 7–19.

Yu, D., Droppo, J., Wu, J., Gong, Y., Acero, A., 2008. Robust speech recognition using a cepstral minimum-mean-square-error-motivated noise suppressor. IEEE Trans. Audio Speech Lang. Process. 16 (5), 1061–1070.

Zhang, Y., Zhao, Y., 2013. Real and imaginary modulation spectral subtraction for speech enhancement. Speech Commun. 55, 509–522.

Zheng, N., Li, X., Blu, T., Lee, T., 2011. SURE-MSE speech enhancement for robust speech recognition. Int. Symp. Chin. Spoken Language Process. 271–274.