Interspeech 2018
2-6 September 2018, Hyderabad

Speech Enhancement Using the Minimum-Probability-of-Error Criterion

Jishnu Sadasivan[1], *Subhadip Mukherjee*[2], *and Chandra Sekhar Seelamantula*[2]

[1]Department of Electrical Communication Engineering, [2]Department of Electrical Engineering,
Indian Institute of Science, Bangalore 560012, India

{sadasivan, subhadipm, chandrasekhar}@iisc.ac.in

Abstract

We propose a novel speech denoising framework by minimizing the probability of error (PE), which measures the deviation probability of the estimate from its true value. To develop the minimum PE (MPE) criterion, one requires the knowledge of the noise probability density function (p.d.f.), which may not be available in a parametric form in speech denoising applications. Therefore, we adopt two approaches for modeling the noise p.d.f.: (i) Gaussian modeling based on adaptive variance estimation; and (ii) a Gaussian mixture model (GMM) in view of its approximation capabilities. We consider discrete cosine transform (DCT) domain shrinkage, where the optimum shrinkage parameter is obtained by minimizing an estimate of the PE. A performance assessment for real-world noise types shows that for input signal-to-noise ratios (SNR) greater than 5 dB, the proposed MPE-based point-wise shrinkage estimators outperform three benchmark techniques in terms of segmental SNR and short-time objective intelligibility (STOI) scores.

Index Terms: Minimum probability of error, Speech denoising, Gaussian mixture model, point-wise shrinkage estimator.

1. Introduction

Ambient acoustic noise introduces unwanted disturbances in speech signals leading to a degradation in speech quality, thereby affecting the downstream processing in speech communication systems and limiting the ability of listeners to understand and concentrate. Therefore, it is imperative to suppress noise and enhance the quality and intelligibility of speech. A typical approach to speech denoising is to minimize an appropriate distortion measure, also referred to as the *risk* function in statistics literature, to obtain an estimate of the clean signal. However, direct minimization of risk requires the knowledge of the underlying clean signal or its statistics, which is difficult to obtain in practice. Hence, one needs to rely on the estimate of the clean signal statistics. Generative processes of speech signals exhibit wide variabilities based on speaker, phonemes and their durations, language, etc., which render the speech signal a non-stationary stochastic process. Therefore, estimating the clean speech prior is difficult, since it necessitates intricate stochastic modeling and requires a rigorous training phase.

Speech enhancement techniques can be broadly categorized into (i) spectral subtraction techniques [1–3], which involve the subtraction of the noise spectrum from the spectrum of noisy speech; (ii) Wiener filtering [4–6], which relies on the estimates of the power-spectra of clean speech and noise; (iii) subspace techniques [7], wherein one utilizes the properties of the signal and noise subspaces; and (iv) statistical model-based approaches, which are setup in a Bayesian framework and rely on an estimate of the clean signal prior [8–16]. Recently, Xu et al. demonstrated the use of *deep neural networks* for learning the nonlinear map from noisy speech to clean speech [17, 18].

In this paper, emphasis is placed on developing a non-Bayesian technique for speech denoising. Our formulation relies only on the noise statistics in its entirety, unlike the mean-squared error (MSE) formulations, in which the first- and second-order statistics suffice [19]. A statistical model is not assumed on the clean speech signal. The key deviation with respect to the state-of-the-art lies in the choice of the distortion measure. We do not employ the standard MSE metric or a perceptual distortion metric. Instead, we consider a novel criterion for denoising, namely the probability of error (PE), which measures the probability of deviation between the ground-truth signal and its estimate. This criterion requires one to know or at least estimate the noise p.d.f., and places no statistical assumptions on the clean signal. The PE criterion is measured in the short-time discrete-cosine transform (DCT) domain. We rely on the parsimony of representation and energy compaction of clean speech in the DCT basis. Soon et al. showed that the DCT is superior to the discrete Fourier transform (DFT) for speech denoising [20]. The noise, however, is non-sparse in the DCT basis. This representation therefore justifies the use of a point-wise shrinkage estimator for denoising.

Since the oracle PE requires one to know the ground-truth, we approximate it by a surrogate function that depends solely on the noisy observations, leading to a practically realizable estimate. Since denoising entails a reduction in noise variance in each spectral band, the PE is minimized with respect to the shrinkage parameter over the interval $[0, 1]$, which is a great convenience as far as optimization is concerned. We develop two different variants of the PE risk, both point-wise, but one is an instantaneous estimator whereas the other incorporates temporal smoothing. Since the key objective is to combat real-world noise types (*street*, *train*, and *F16* noise), which are non-stationary and whose distributions are not available a priori, we adopt two p.d.f. models, one based on the Gaussian and the other employing a GMM. Experimental results are presented on the real-world noise types for various input signal-to-noise ratios (SNRs) and compared with the state of the art.

2. MPE for Speech Denoising

Consider the additive observation model

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

where s_n denotes the clean signal and x_n is the observation corrupted by noise samples w_n, which are independent and identically distributed (i.i.d.) with zero mean and variance σ^2. Short-time discrete cosine transform (DCT) domain processing is considered for denoising within the proposed MPE formalism. The short-time DCT representation of (1) takes the form

$$X_{k,i} = S_{k,i} + W_{k,i},\ k = 1, \cdots, K, \text{ and } i = 1, \cdots, M, \tag{2}$$

where k and i denote the DCT coefficient and the speech frame indices, respectively. For estimating $S_{k,i}$, we develop a point-

10.21437/Interspeech.2018-1294

wise estimator $\widehat{S}_{k,i} = a_{k,i} X_{k,i}$, where the shrinkage factor $a_{k,i} \in [0, 1]$ is selected optimally based on the MPE criterion.

2.1. MPE criteria for point-wise shrinkage

Since the estimator is point-wise, we drop the indices k and i to maintain brevity of notation, and define the PE as

$$\mathcal{R} = \mathbb{P}\left(\left|\widehat{S} - S\right| > \epsilon\right), \tag{3}$$

where $\epsilon > 0$ is a predefined tolerance parameter. Substituting $\widehat{S} = aX = a(S + W)$, the risk in (3) evaluates to

$$\begin{aligned}\mathcal{R}(a, S) &= \mathbb{P}(|a(S + W) - S| > \epsilon) \\ &= 1 - F\left(\frac{\epsilon - (a-1)S}{a}\right) + F\left(-\frac{\epsilon + (a-1)S}{a}\right), \end{aligned} \tag{4}$$

where $F(\cdot)$ is the cumulative distribution function (c.d.f.) of the noise in the DCT domain. Since $\mathcal{R}$ depends on the ground-truth S, it is impractical to optimize it directly over a, as the estimator would be unrealizable. Therefore, we minimize an estimate of $\mathcal{R}$, which is obtained by replacing S in (4) with its noisy counterpart X. Such an estimate $\widehat{\mathcal{R}}$ takes the form

$$\widehat{\mathcal{R}}(a, X) = 1 - F\left(\frac{\epsilon - (a-1)X}{a}\right) + F\left(-\frac{\epsilon + (a-1)X}{a}\right),$$

and correspondingly, the optimal shrinkage is obtained as $a_{\text{opt}} = \arg\min_{0 \le a \le 1} \widehat{\mathcal{R}}$, by performing a grid-search over $[0, 1]$ with a grid-spacing of 0.01.

We consider two types of shrinkage estimators. The first one, referred to as MPE-1, applies different shrinkage factors to each spectral coefficient $\{X_{k,i}\}$ in the i^{th} frame. The optimal shrinkage is selected coefficient-wise as $a_{k,i}^{\text{opt}} = \arg\min_{0 \le a \le 1} \widehat{\mathcal{R}}(a, X_{k,i})$. In the second variant, which we refer to as MPE-2, a single shrinkage factor is applied to a group of coefficients bunched along i, resulting in an estimator of the form $\widehat{S}_{k,i} = a_{k,i}^{\text{opt}} X_{k,i}$, where

$$a_{k,i}^{\text{opt}} = \arg\min_{0 \le a \le 1} \sum_{t=-\tau}^{+\tau} \widehat{\mathcal{R}}(a, X_{k,i-t}). \tag{5}$$

The parameter τ determines the extent of temporal averaging.

2.2. Approximating unknown noise distributions

In real-world speech denoising scenarios, the noise distribution is often not known a priori in parametric form. In such scenarios, one has to model the noise p.d.f. appropriately. We consider two approaches for noise modeling: In the first one, we use a Gaussian, whose variance is estimated adaptively from the noisy speech signal, whereas in the second approach, we employ a GMM-based model. The effectiveness of the models will be validated experimentally.

2.2.1. Gaussian model and adaptive variance estimation

This approach relies on the assumption that the time-domain noise samples within a frame are i.i.d. random variables. Since the DCT coefficients are linear combinations of i.i.d. random variables, considering the frame length to be sufficiently large, we invoke the *central limit theorem*, which assures that each DCT coefficient $W_{k,i}$ is approximately Gaussian distributed. A stochastic model based voice-activity detector (VAD) [26] is employed to estimate the variance of $W_{k,i}$. Going by the recommendation in [20], we use the following recursion to estimate the noise variance adaptively:

$$\hat{\sigma}_{k,i}^2 = \begin{cases} \eta \hat{\sigma}_{k,i-1}^2 + (1 - \eta) X_{k,i}^2, & \text{if } i^{\text{th}} \text{ frame is noise-only,} \\ \hat{\sigma}_{k,i-1}^2, & \text{otherwise,} \end{cases}$$

where $\eta = 0.98$. Essentially, the noise variance is updated if the VAD identifies that the frame under consideration corresponds to noise alone. In the sequel, the point-wise shrinkage estimators MPE-1 and MPE-2 for the Gaussian noise model are referred to as MPE-1-G and MPE-2-G, respectively.

2.2.2. Noise modeling using GMM

The motivation for using GMM stems from the fact that it can approximate any p.d.f. with a finite number of discontinuities sufficiently accurately [21]. The L-component GMM p.d.f. with parameters $\{\alpha_m, \theta_m, \sigma_m\}_{m=1}^{L}$ is given by

$$f(W) = \sum_{m=1}^{L} \frac{\alpha_m}{\sigma_m \sqrt{2\pi}} \exp\left(-\frac{(W - \theta_m)^2}{2\sigma_m^2}\right), \tag{6}$$

and the corresponding PE risk turns out to be

$$\begin{aligned}\widehat{\mathcal{R}} = \sum_{m=1}^{L} \alpha_m \Bigg[&Q\left(\frac{\epsilon - (a-1)X - a\theta_m}{a\sigma_m}\right) + \\ &Q\left(\frac{\epsilon + (a-1)X + a\theta_m}{a\sigma_m}\right)\Bigg], \end{aligned} \tag{7}$$

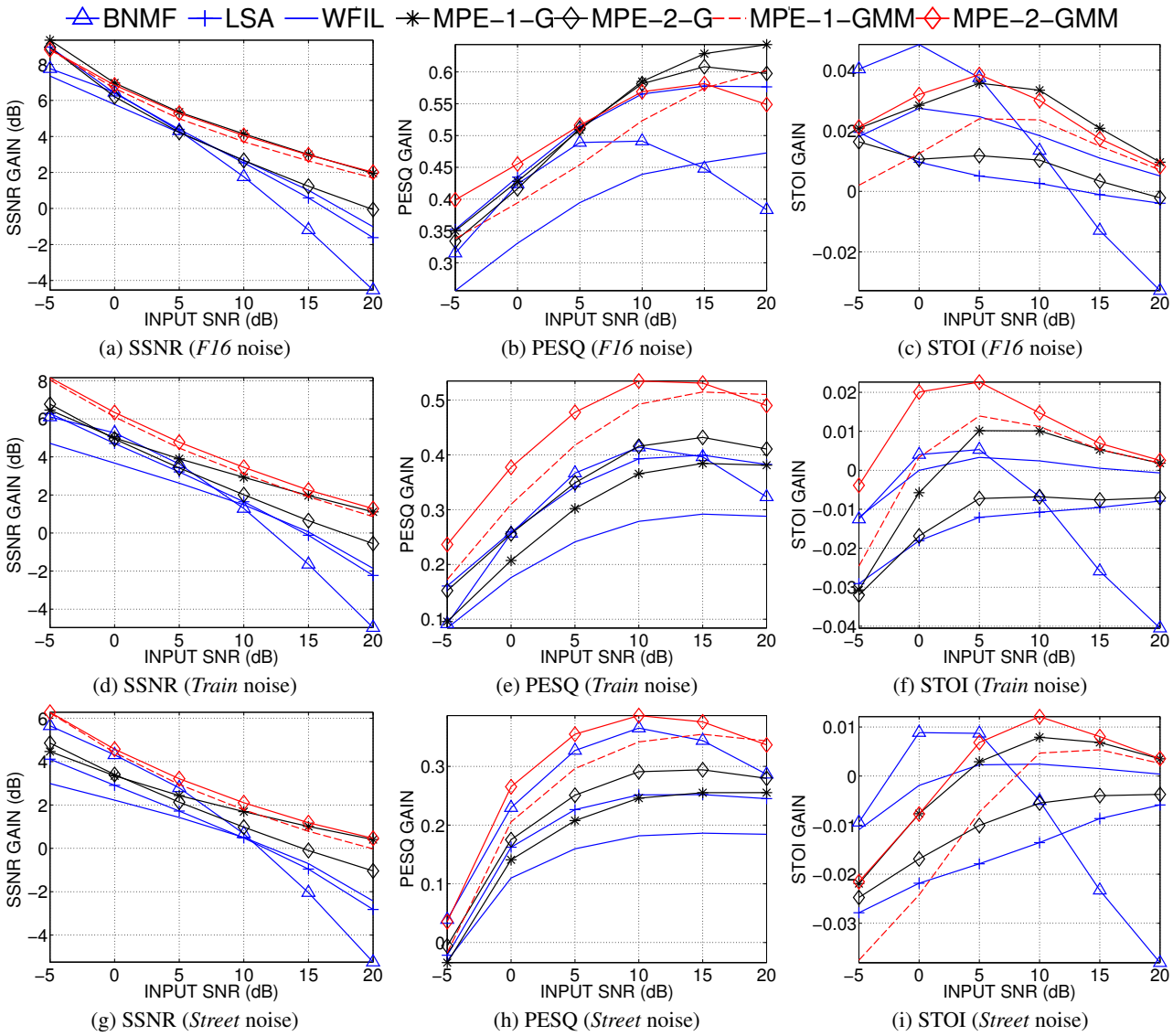
where $Q(u) = \frac{1}{\sqrt{2\pi}} \int_u^\infty \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t$. The number of GMM components M is selected following the Bayesian information criterion (BIC) [23]. For each subband, the parameters of the GMM are estimated using the expectation-maximization (EM) algorithm [22] based on training data corresponding exclusively to noise. The GMM-based p.d.f. modeling, when used in conjunction with the MPE-1 and MPE-2 estimators, are referred to as MPE-1-GMM and MPE-2-GMM, respectively. The noise samples during training and testing are taken to be different.

3. Simulation Results

Clean speech recordings from the Noizeus database (8 kHz sampling frequency) [24] are used in our experiments. The noise samples are taken from both Noizeus (*train* and *street* noises) and Noisex-92 (for *F16* noise; downsampled to 8 kHz) databases [25]. We consider frame-by-frame processing, with a Hamming window, frame length of 40 ms, and an overlap of 75% between consecutive frames. The value of τ in MPE-2 in (5) is set to 3, and we choose $\epsilon = 3\,\sigma$, where σ is the noise standard deviation.

We perform a comparative assessment of the MPE-based techniques with three benchmarking algorithms under different noise conditions[1]. The algorithms chosen for the comparison are: (i) Wiener filter technique, which uses a decision-directed approach for a priori SNR estimation (WFIL) [6]; (ii) log-spectral amplitude estimator (LSA), which minimizes the mean-squared error (MSE) of the logarithm of clean speech spectral amplitude [9]; and (iii) Bayesian non-negative matrix

[1]Example speech files are available at `https://spectrumee.wixsite.com/mpe-se`.



(a) SSNR (*F16* noise) (b) PESQ (*F16* noise) (c) STOI (*F16* noise)

(d) SSNR (*Train* noise) (e) PESQ (*Train* noise) (f) STOI (*Train* noise)

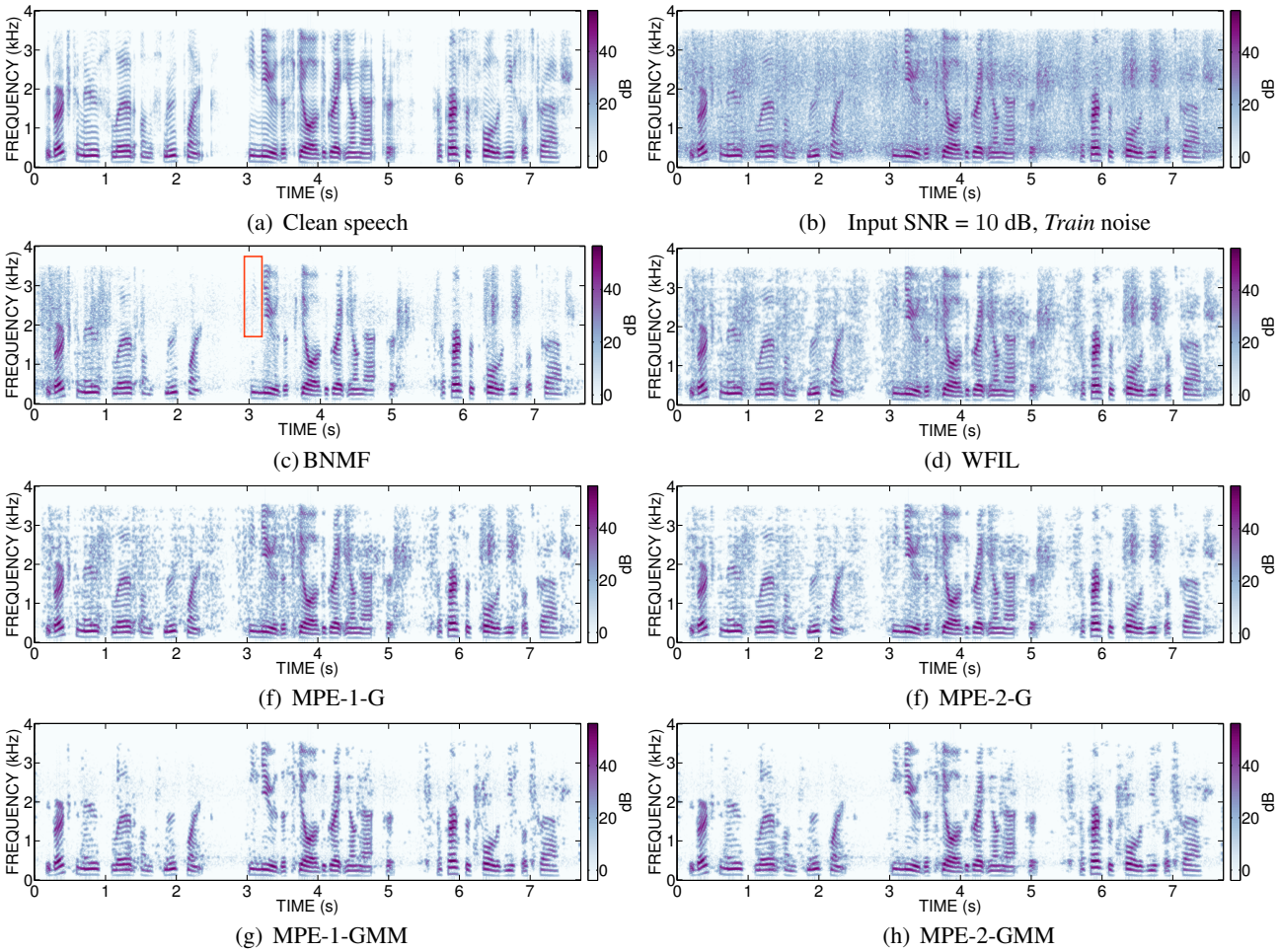
(g) SSNR (*Street* noise) (h) PESQ (*Street* noise) (i) STOI (*Street* noise)

Figure 1: *Performance comparison of various algorithms for different noise types in terms of SSNR, PESQ, and STOI scores.*

factorization method (BNMF), wherein one optimizes the MSE of the clean speech spectral amplitude with the help of a dictionary trained offline on clean speech [15]. Matlab implementations of WFIL and LSA are available in [27]. The implementations use the VAD proposed in [26]. For MPE-1-G/MPE-2-G, we use the same VAD. For the GMM approach, a VAD is not needed, since it is a pre-trained model. For the BNMF implementation, we use the Matlab code provided online by the authors [15]. The choice of WFIL and LSA for performance benchmarking is motivated by the extensive comparison reported in [27], which established conclusively that these result in higher speech quality and intelligibility than the competing techniques. The BNMF technique has been shown to be the best among NMF based speech denoising approaches.

Three objective scores are computed for performance evaluation: (i) Segmental signal-to-noise-ratio (SSNR), calculated by averaging the SNRs over short speech segments; (ii) Perceptual evaluation of speech quality (PESQ) [28], which is widely used to measure the perceptual speech quality in narrowband telephone networks, speech codecs, and denoised speech; and (iii) Short-time objective intelligibility score (STOI), which has been shown to be highly correlated with the intelligibility of the denoised speech [29]. The scores are averaged over 30 different speech files corresponding to 10 independent and randomly selected noise realizations for each input SNR.

Figure 1 shows the performance comparison of the techniques for *F16*, *train*, and *street* noise. We observe that for all the noise types under consideration, MPE-1-GMM and MPE-2-GMM exhibit a higher SSNR gain compared with the competing algorithms (cf. Figures 1(a), 1(d), and 1(g)). Further, in the case of *F16* noise, and for other noise types with input SNR greater than 5 dB, MPE-1-G exhibits a better denoising performance in terms of SSNR. Among the proposed MPE estimators, the SSNR gain obtained using MPE-2-G turns out to be the least. In terms of PESQ scores (cf. Figures 1(b), 1(e), and 1(h)), we observe that, for all the noise types considered, MPE-2-GMM leads to the best performance. For *F16* noise, MPE-1-G and MPE-2-G also result in fairly high PESQ scores.



(a) Clean speech (b) Input SNR = 10 dB, *Train* noise

(c) BNMF (d) WFIL

(f) MPE-1-G (f) MPE-2-G

(g) MPE-1-GMM (h) MPE-2-GMM

Figure 2: *Spectrograms of the denoised speech obtained using different algorithms.*

For input SNR exceeding 5 dB, MPE-1-G, MPE-1-GMM, and MPE-2-GMM exhibit a denoising performance superior to their competitors in terms of STOI (cf. Figures 1(c), 1(f), and 1(i)).

To summarize, MPE-2-GMM exhibits a better performance compared with all the other techniques. In the case of *street* and *train* noise, GMM-based MPE estimators show a superior denoising performance than their Gaussian counterparts. In the case of *F16* noise, the Gaussian model led to a better denoising. This is probably because the *F16* noise is relatively stationary compared with *street* and *train* noise, and the adaptive variance estimation using a VAD is reasonably accurate.

To demonstrate the time-frequency structure, distribution of residual noise, and speech distortion, we show the spectrograms of the denoised, noisy, and clean speech signals in Figure 2 corresponding to the *train* noise. We observe that WFIL has a higher residual noise than all the other algorithms. BNMF suppresses noise, especially in the silence regions, but it introduces speech distortions in some regions (cf. Figure 2(c), high frequency region (2.5 to 3.5 kHz) just after 3 s, highlighted using a red rectangle). MPE-1-GMM and MPE-2-GMM yield superior noise suppression and less speech distortion. In the case of MPE-1-G/MPE-1-GMM, a small amount of musical noise is present, which is suppressed to some extent in MPE-2-G/MPE-2-GMM, since by construction, MPE-2 incorporates temporal smoothing while computing the point-wise shrinkage estimator.

4. Conclusions

We proposed a novel criterion for speech denoising based on the probability of error. Our formalism does not place any statistical assumptions on the clean speech signal. Notwithstanding its simplicity, the performance of the proposed denoiser turned out to be competitive with the state-of-the-art techniques under real-world noise conditions. Further, an implicit assumption of the proposed framework is that the clean signal admits a parsimonious representation in a chosen basis, which is true of the speech signal in the DCT domain, and that the noise does not, which makes the point-wise shrinkage a natural choice for denoising. The proposed framework relies on modeling the noise p.d.f., for which we develop Gaussian and GMM-based approximations. The standard deviation of the Gaussian model for noise is updated recursively using a VAD, whereas the parameters for the GMM are pre-trained. Updating the GMM parameters adaptively might lead to an improvement in the denoising performance under real-world noise conditions. Two versions of point-wise shrinkage were considered, one instantaneous and the other involving a certain degree of temporal smoothing, with the latter leading to a superior performance. All the same, excessive smoothing might deteriorate the performance and the optimal degree of smoothing to be incorporated in the MPE framework must be ascertained.

5. References

[1] P. Lockwood and J. Boudy, "Experiments with a non-linear spectral subtractor (NSS), hidden Markov models and the projections, for robust recognition in cars," *Speech Commun.*, vol. 11, issue 2, pp. 215–228, Jun. 1992

[2] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," *Proc. IEEE Intl. Conf. Acoust., Speech, Signal Process.*, vol. 4, pp. 4164–4167, May 2002.

[3] Y. Hu and P. Loizou, "A perceptually motivated approach for speech enhancement," *IEEE Trans. Speech, Audio Process.*, vol. 11, no. 5, pp. 457–465, Sep. 2003.

[4] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insight into noise reduction Wiener filter," *IEEE Trans. Speech, Audio Process.*, vol. 14, no. 4, pp. 1218–1234, Jul. 2006.

[5] S. Srinivasan, J. Samuelsson, and W. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. Speech, Audio Process.*, vol. 14, no.1, pp. 163–176, Jan. 2006.

[6] P. Scalart, and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," *in Proc. IEEE Intl. Conf. Acoust., Speech, Signal Process.*, vol. 2, pp. 629–632, May. 1996.

[7] F. Jabloun and B. Champagne, "Incorporating the human hearing properties in the signal subspace approach for speech enhancement," *IEEE Trans. Speech, Audio Process.*, vol. 11, no. 6, Nov. 2003.

[8] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-squared error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[9] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-squared error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.

[10] P. C. Loizou, "Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum," *IEEE Trans. Speech, Audio Process.*, vol. 13, no. 5, pp. 857–869, Sep. 2005.

[11] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized gamma priors," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, no. 6, pp. 1741–1752, Aug. 2007.

[12] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden Markov models," *IEEE Trans. Signal Process.*, vol. 40, no. 4, pp. 725–735, Apr. 1992.

[13] T. Lotter and P. Vary, "Speech enhancement by maximum a posteriori estimation using super-Gaussian speech model," *EURASIP J. Appl. Signal Process.*, vol. 7, pp. 1110–1126, 2005.

[14] P. Mowlaee and R. Saeidi, "Iterative closed-loop phase-aware single-channel speech enhancement," *IEEE Signal Process. Lett.*, vol. 20, no. 12, pp. 1235–1239, Dec. 2013.

[15] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no.10, pp. 2140–2151, Oct. 2013.

[16] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech, Audio Process.*, vol. 13, no. 5, pp. 845–856, Sep. 2005.

[17] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, Jan. 2014.

[18] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 1, pp. 7–19, 2015.

[19] J. Sadasivan, S. Mukherjee, and C. S. Seelamantula, "An optimum shrinkage estimator based on minimum-probability-of-error criterion and application to signal denoising," in *Proc. IEEE Intl. Conf. on Acoust. Speech and Signal Process.*, pp. 4249–4253, 2014.

[20] I. Y. Soon, S. N. Koh, and C. K. Yeo, "Noisy speech enhancement using discrete cosine transform," *Speech Commun.*, vol. 24, pp. 249–257, Jun. 1998.

[21] H. W. Sorenson and D. L. Alspach, "Recursive Bayesian estimation using Gaussian sums," *Automatica*, Vol. 7, pp. 465–479, 1971.

[22] R. Redner and H. Walker. "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Review*, Vol. 26, no. 2, pp. 195–239, Apr. 1984.

[23] C. Fraley and A. Raftery, "How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis," Technical Report 329, Dept. Statistics, Univ. Washington, Seattle, WA, 1998.

[24] Y. Hu and P. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Commn.*, vol. 49, pp. 588–601, Jul. 2007.

[25] A. Varga and H. J. M. Steeneken, "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 Commn.*, vol. 12, no. 3, pp. 247–251, 1993.

[26] J. Sohn and N. S. Kim, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[27] P. Loizou, *Speech Enhancement — Theory and Practice*. CRC Press, 2007.

[28] ITU-T Rec. P.862, "Perceptual evaluation of speech quality (PESQ) – An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," International Telecommunication Union, Feb. 2001.

[29] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.* vol. 19, pp. 2125–2136, Sep. 2011.