A STUDY OF NOISE PSD ESTIMATORS FOR SINGLE CHANNEL SPEECH ENHANCEMENT

Mathew Shaji Kavalekalam[1], Jesper Kjær Nielsen[1], Mads Græsbøll Christensen[1] and Jesper B. Boldt[2]

[1]Audio Analysis Lab, CREATE, Aalborg University, Denmark {msk, jkn, mgc}@create.aau.dk
[2]GN Hearing A/S, DK 2750, Ballerup, Denmark jboldt@gnresound.com

ABSTRACT

The estimation of the noise power spectral density (PSD) forms a critical component of several existing single channel speech enhancement systems. In this paper, we evaluate one new and some of the existing and commonly used noise PSD estimation algorithms in terms of the spectral estimation accuracy and the enhancement performance for different commonly encountered background noises, which are stationary and non-stationary in nature. The evaluated algorithms include the Minimum Statistics, MMSE, IMCRA methods and a new model-based method.

Index Terms— speech enhancement, noise PSD estimation, autoregressive models

1. INTRODUCTION

Speech enhancement algorithms have a wide range of applications such as in digital hearing aids, speech recognition systems, mobile communications, etc [1], where the desired speech is degraded by acoustic background noise. These algorithms can be broadly categorised into single and multi channel algorithms. In this paper, we are only concerned with the former class of algorithms. The single channel speech enhancement algorithms must generally incorporate some assumptions to remove the background noise from the desired signal. For example, the Wiener filter assumes the second-order statistics of the speech/noise signal to be known. In practical scenarios, these statistics must be estimated from noisy observations. Thus, a very critical part present in most of the single channel speech enhancement methods is the estimation of the noise PSD [2, 3]. A significant amount of work has been done in the past decades to solve this problem.

In this paper, we evaluate some of the well known noise PSD estimation algorithms along with a new model-based approach [4]. Previously, an evaluation of noise PSD estimators was carried out in [5]. This study compared some of the existing noise PSD estimators in terms of the spectral estimation accuracy. In this study, we also evaluate the noise PSD estimators in terms of its enhancement capabilities in some of the typically encountered background noises. The estimation of noise PSD is not a trivial task especially in the case of non-stationary noises. In such scenarios, the noise PSD estimate has to be updated as rapidly as possible. An under-estimation or over-estimation of the noise PSD can lead to residual noise or speech distortion. In the current study, we evaluate different noise PSD estimation algorithms for different types of commonly encountered background noise, which are stationary and non-stationary in nature. The well-known algorithms that we have evaluated in this paper are Minimum Statistics (MS) method [6], Improved minima controlled recursive averaging (IMCRA) [7] method and minimum mean squared error (MMSE) based estimation [8]. In addition to these algorithms, we also evaluate a new model-based approach for estimating the noise PSD. A detailed description regarding this method can be found in [4]. Here we focus on evaluating its performance. This method uses a priori information regarding the speech and noise spectral shapes in the form of autoregressive (AR) parameters stored in trained speech and noise codebooks.

The remainder of this paper will be organised as follows. Section 2 gives a brief introduction to the noise PSD estimation problem and an overview of the model-based method for estimating the noise PSD. A brief overview of the compared algorithms is given in Section 3. The experiments used in the evaluation of the algorithms will be explained in section 4 followed by the results and conclusion in Sections 5 and 6 respectively.

This work was supported by Innovations fund Denmark (Grant no: 99-2014-1)

2. MODEL BASED APPROACH FOR ESTIMATING THE NOISE PSD

This section formulates the noise PSD estimation problem and gives a brief overview of the model-based approach for estimating the noise PSD. We refer the interested readers to a companion paper [4] (for further details). It is assumed here that N samples of noisy signal are observed as

$$\boldsymbol{y} = \boldsymbol{s} + \boldsymbol{e}, \tag{1}$$

where $\boldsymbol{y} \in \mathbb{R}^N$, $\boldsymbol{s} \in \mathbb{R}^N$, and $\boldsymbol{e} \in \mathbb{R}^N$ are the noisy speech, the clean speech, and the noise, respectively. The basic task here is to estimate the noise PSD which is typically defined as [9]

$$\phi_e(\omega) = \lim_{N\to\infty} \frac{1}{N}\mathbb{E}[|E(\omega)|^2|\boldsymbol{y}] \tag{2}$$

where $\mathbb{E}$ is the expectation operator and $E(\omega) = \boldsymbol{f}^H(\omega)\boldsymbol{e}$ is the DFT of the noise with $\boldsymbol{f}(\omega) = [1 \quad \exp(j\omega) \quad \cdots \quad \exp(j\omega(N-1))]^T$. The conditional expectation in (2) is the second moment of the density $p(E(\omega)|\boldsymbol{y})$ which leads to (2) be rewritten in terms of $p(\boldsymbol{e}|\boldsymbol{y})$ as

$$\phi_{\mathrm{e}}(\omega) = \lim_{N\to\infty} \frac{1}{N}\boldsymbol{f}^H(\omega)\left[\int_{\mathbb{R}^{N\times 1}} \boldsymbol{e}\boldsymbol{e}^T p(\boldsymbol{e}|\boldsymbol{y})d\boldsymbol{e}\right]\boldsymbol{f}(\omega)\,. \tag{3}$$

To compute the posterior $p(\boldsymbol{e}|\boldsymbol{y})$, statistical models denoted as $\{\mathcal{M}_k\}_{k=1}^K$, are used for explaining the generation of data. These models can be incorporated into (3) as,

$$\begin{aligned}\phi_{\mathrm{e}}(\omega) &\approx \frac{1}{N}\sum_{k=1}^{K} p(\mathcal{M}_k|\boldsymbol{y}) \\ &\quad\times \boldsymbol{f}^H(\omega)\left[\int_{\mathbb{R}^{N\times 1}} \boldsymbol{e}\boldsymbol{e}^T p(\boldsymbol{e}|\boldsymbol{y},\mathcal{M}_k)d\boldsymbol{e}\right]\boldsymbol{f}(\omega) \qquad (4)\\ &= \sum_{k=1}^{K} p(\mathcal{M}_k|\boldsymbol{y})\phi_{\mathrm{e}}(\omega|\mathcal{M}_k). \qquad (5)\end{aligned}$$

where $\{p(\mathcal{M}_k|\boldsymbol{y})\}_{k=1}^{K}$ denote the model probabilities, which ensure that models explaining the data well are given more weight in comparison to other models. The models that have been used are autoregressive (AR) models for speech and noise denoted by [10, 11]

$$p(\boldsymbol{s}|\sigma_{\mathrm{s},k}^2, \mathcal{M}_k) = \mathcal{N}(\boldsymbol{0}, \sigma_{\mathrm{s},k}^2 \boldsymbol{R}_{\mathrm{s}}(\boldsymbol{a}_k)) \quad (6)$$

$$p(\boldsymbol{e}|\sigma_{\mathrm{e},k}^2, \mathcal{M}_k) = \mathcal{N}(\boldsymbol{0}, \sigma_{\mathrm{e},k}^2 \boldsymbol{R}_{\mathrm{e}}(\boldsymbol{b}_k)) \quad (7)$$

where $\sigma_{\mathrm{s},k}^2$, $\sigma_{\mathrm{e},k}^2$, $\boldsymbol{R}_{\mathrm{s}}(\boldsymbol{a}_k)$, $\boldsymbol{R}_{\mathrm{e}}(\boldsymbol{b}_k)$, $\boldsymbol{a}_k$, and $\boldsymbol{b}_k$ are the excitation noise variance, the normalised covariance matrices, and the AR-parameters of the speech and the noise, respectively. It can be shown under certain assumptions that the normalised covariance matrix corresponding to speech and noise can be diagonalised by the DFT matrix [12, 11]. The excitation variances are treated as unknown random variables with the priors,

$$p(\sigma_{\mathrm{s},k}^2|\mathcal{M}_k) = \text{Inv-}\mathcal{G}(\alpha_{\mathrm{s},k}, \beta_{\mathrm{s},k}) \quad (8)$$

$$p(\sigma_{\mathrm{e},k}^2|\mathcal{M}_k) = \text{Inv-}\mathcal{G}(\alpha_{\mathrm{e},k}, \beta_{\mathrm{e},k}). \quad (9)$$

As seen from (4) and (5), the posteriori model probabilities and the second moment of the posterior needs to be computed to get the final noise PSD estimate. As there is no closed form solution to obtain this, a variational Bayesian framework [13, 14] is used to produce an analytical approximation of the full joint posterior used in (4) as

$$p(\boldsymbol{e}, \sigma_{\mathrm{s},k}^2, \sigma_{\mathrm{e},k}^2|\boldsymbol{y}, \mathcal{M}_k)p(\mathcal{M}_k|\boldsymbol{y}) \approx q(\boldsymbol{e}|\boldsymbol{y}, \mathcal{M}_k)q(\sigma_{\mathrm{s},k}^2, \sigma_{\mathrm{e},k}^2|\boldsymbol{y}, \mathcal{M}_k)q(\mathcal{M}_k|\boldsymbol{y})\,. \quad (10)$$

Since the posterior factor $q(\boldsymbol{e}|\boldsymbol{y}, \mathcal{M}_k)$ is a normal distribution, its second moment and the posterior model probabilities $q(\mathcal{M}_k|\boldsymbol{y})$ is substituted in (5) to get the final noise PSD estimate. More details regarding the derivation of this method can be found in `http://tinyurl.com/jknvbn`.

3. OVERVIEW OF THE EXISTING ALGORITHMS

In this section, we will give a brief overview of the existing noise PSD estimation algorithms that have been evaluated in this paper.

3.1. Minimum Statistics

This method [6] tracks the minima of the smoothed noisy spectrum for each frequency component. The method is based on the observation that the speech and noise component are statistically independent and that the power of the noisy signal often goes down to the power of the noise signal. The smoothed noisy spectrum is calculated using a recursive smoothing equation. Since this method is based on computing the minimum of the smoothed noisy spectrum over a moving window, the noise PSD estimate is necessarily biased. This is overcome in [6] to some extent by using a bias compensation factor in time and frequency.

3.2. IMCRA

In this method [7], the noise PSD estimate is obtained by a recursive averaging of the noisy spectral values using a time varying frequency dependent smoothing parameter, that is adjusted according to the speech presence probability (SPP) for each frequency component. The a priori SPP are calculated in this method after two iterations of smoothing and minima tracking. The final SPP (used for the recursive averaging) is then computed using the a priori SPP and the estimated a priori SNR.

3.3. MMSE

This method [8] derives an MMSE estimator of the noise PSD coefficients. Here, the speech and noise spectral coefficients are modelled as normally distributed random variables that are independent with each other. The first step involves the computation of the conditional expectation of the noise periodogram given the noisy signal which involves a weighted combination of noise PSD estimate from the previous frame and the noisy periodogram from the current frame. The final noise PSD estimate is then obtained by a recursive averaging of the estimated noise periodogram.

4. EXPERIMENTS

We will now describe the experiments that have been carried out to evaluate the four noise PSD estimation algorithms. Section 4.1 describes the parameters that have been used for implementing the different noise PSD estimation algorithms. Sections 4.2 and 4.3 explains the experiments done to evaluate the estimation accuracy and the enhancement capabilities of the noise PSD estimation algorithms, respectively.

4.1. Implementation Details

We have evaluated a total of four algorithms: MS, IMCRA, MMSE and the new model based approach. The test signals used for evaluation were taken from the EUROM database [15]. The clean speech signals were then degraded by 5 types of typically encountered background noise: babble, street, station, exhibition and restaurant from the NOIZEUS database [16]. The model based approach for estimating the noise PSD explained in Section 2 requires the speech and noise codebooks to be trained offline. For the experiments we have trained a speech codebook of 64 entries and a noise codebook of 12 entries. The codebooks were trained using a variation of the LBG algorithm [17]. The training data used for creating the speech codebook consisted of audio samples from the EUROM database. It should be noted that we have trained a codebook that is independent of the speaker. The data used for generating the noise codebook consisted of noise files from the NOIZEUS database. Different codebooks were trained for different types of noise, which were then appended together to form a larger codebook. The noise codebook had a size of 16 entries, which consisted of 4 entries each for babble, restaurant and exhibition and 2 entries each for street and station. It should be noted that, while testing for a particular noise scenario, the noise codebook entries corresponding to that scenario is **NOT** used for the estimation of noise PSD. The codebooks as well as MATLAB code for generating the codebooks will be available at `http://tinyurl.com/jknvbn`. The AR order for the speech and noise models were chosen to be 14. All the noise PSD estimation algorithms evaluated here work on a frame size of 32 ms with 50% overlap.

4.2. Estimation Accuracy

We have used the log spectral distortion between the estimated noise PSD and the reference noise PSD to measure the spectral estimation accuracy of the algorithms. The reference PSD in this case is computed by taking the periodogram of the noise only signal. The mean log spectral distortion across the whole signal is given by

$$\text{LogErr} = \frac{10}{NL} \sum_{l=0}^{L-1} \sum_{k=0}^{N-1} \left| \log_{10}, \frac{\phi_e(k,l)}{\hat{\phi}_e(k,l)} \right| \quad (11)$$

where $\phi_e(k,l)$ is the true noise PSD and $\hat{\phi}_e(k,l)$ is the estimated noise PSD at the k^{th} frequency index of the l^{th} frame. This term can be separated into distortion due to over-estimation and under-estimation of the noise PSD, which can be written as LogErr = $\text{LogErr}_{\text{ov}} + \text{LogErr}_{\text{un}}$, where $\text{LogErr}_{\text{ov}}$ and $\text{LogErr}_{\text{un}}$ are defined as [8]

$$\text{LogErr}_{\text{ov}} = \frac{10}{NL}\sum_{l=0}^{L-1}\sum_{k=0}^{N-1}\left|\min\left(0, \log_{10}\frac{\phi_e(k,l)}{\hat{\phi}_e(k,l)}\right)\right| \tag{12}$$

$$\text{LogErr}_{\text{un}} = \frac{10}{NL}\sum_{l=0}^{L-1}\sum_{k=0}^{N-1}\max\left(0, \log_{10}\frac{\phi_e(k,l)}{\hat{\phi}_e(k,l)}\right). \tag{13}$$

Overestimation of the noise PSD measured by $\text{LogErr}_{\text{ov}}$ is likely to cause speech distortion during the enhancement stage, whereas $\text{LogErr}_{\text{un}}$ gives a measure of the residual noise present in the enhanced signal. A plot of these measures for different acoustic background noises is shown in Section 5.

4.3. Enhancement performance

The estimated noise PSD is then incorporated in a speech enhancement framework. For this, we first estimate the a priori SNR using the decision directed approach [2]. The estimated a priori SNR is then incorporated in a Wiener filter for speech enhancement. In this work, we have used the Segmental SNR (segSNR), Segmental speech SNR (spSNR) and segmental noise reduction (segNR) which has also been used in [18, 8] to evaluate the enhancement performance. segSNR, spSNR and segNR are denoted as

$$\text{segSNR} = \frac{10}{L}\sum_{l=0}^{L-1}\log_{10}\frac{\sum_{m=1}^{M}s^2(lM+m)}{\sum_{m=1}^{M}(s(lM+m)-\hat{s}(lM+m))^2} \tag{14}$$

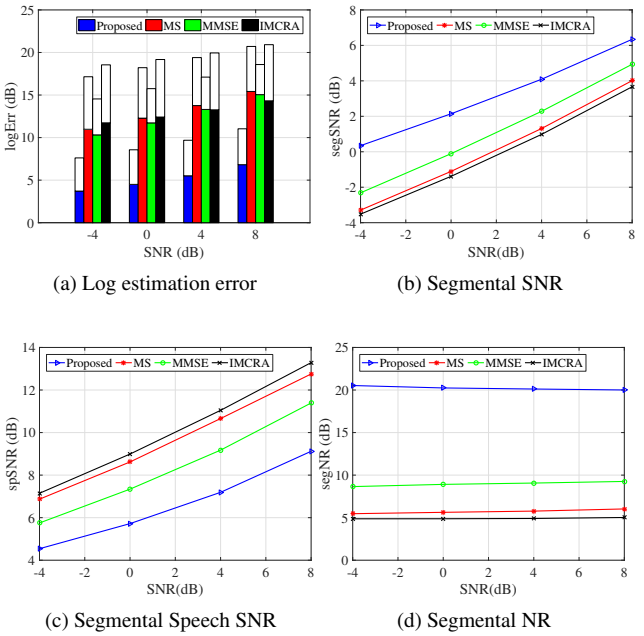
$$\text{spSNR} = \frac{10}{L}\sum_{l=0}^{L-1}\log_{10}\frac{\sum_{m=1}^{M}s^2(lM+m)}{\sum_{m=1}^{M}(s(lM+m)-\tilde{s}(lM+m))^2} \tag{15}$$

$$\text{segNR} = \frac{10}{L}\sum_{l=0}^{L-1}\log_{10}\frac{\sum_{m=1}^{M}e(lM+m)^2}{\sum_{m=1}^{M}\tilde{e}(lM+m)^2} \tag{16}$$

where $s(n)$ denotes the clean signal, $e(n)$ denotes the noise signal, $\hat{s}(n)$ denotes the enhanced signal and M denotes the number of samples in a segment. The term $\tilde{s}(n)$ and $\tilde{e}(n)$ are obtained by the applying the estimated Wiener filter onto $s(n)$ and $e(n)$ respectively. The spSNR measures the speech distortion, where an increase in speech distortion is indicated by a decrease in spSNR. segNR gives a measure of the residual noise present in the signal after enhancement. segSNR improvement takes into account both the speech distortion and noise reduction. A plot of these measures for different acoustic background noise is shown in section 5.

5. RESULTS

In this section we plot the performance metrics introduced in Sections 4.2 and 4.3 for different background noises. Figure 1 shows the results obtained for babble noise. Figure 1a corresponds the log error distortion for different methods as a function of the input SNR. The lower shaded area of the bar plot corresponds to $\text{LogErr}_{\text{ov}}$ caused due to over estimation of the noise PSD and upper part corresponds to $\text{LogErr}_{\text{un}}$ caused due to under estimation of the noise PSD. It can be seen that the model based approach performs the best in terms of log distortion measure followed by MMSE, MS and IMCRA. Figure 1b shows the segmental SNR for the different methods as a function of the input SNR. Figures 1c and 1d show the segmental speech SNR and noise reduction, respectively. It can be seen that even though the model based approach performs the best in terms of segSNR and segNR, it also has the lowest spSNR. This indicates a high noise reduction at the cost of speech distortion. IMCRA which performs the worst in terms of noise reduction performs the best in terms of speech distortion. This is a common trade-off observed in speech enhancement [19]. Figures 2, 3, 4 and 5 show the obtained results for restaurant, exhibition, street and station noise respectively. These figures also show a similar trend as observed for the babble noise. It should be noted that the benefit of using the model based approach over the other methods is more pronounced in relatively non-stationary noises such as babble and the restaurant noise. This can be explained by the zero tracking delay of the model based approach in comparison to other nethods which atleast have a few hundred milliseconds of tracking delay [4, 8].



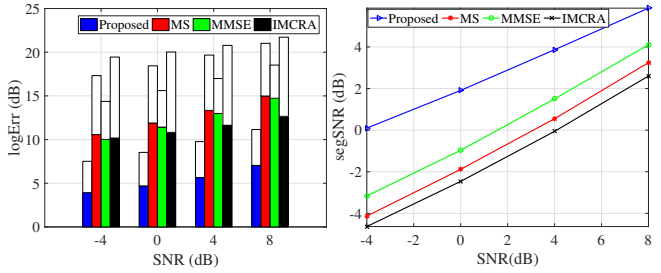
(a) Log estimation error (b) Segmental SNR

(c) Segmental Speech SNR (d) Segmental NR

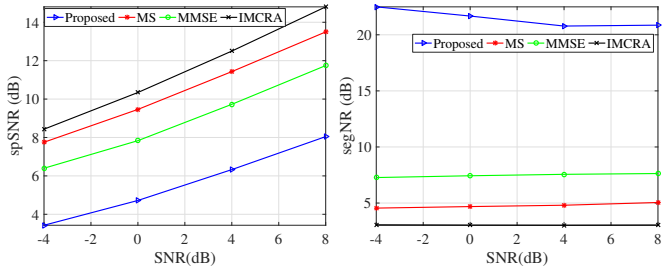
Fig. 1: Performace measures of the algorithms for babble noise. The lower part of the subfigure 1a represents the $\text{LogErr}_{\text{ov}}$ and the upper part in white represents $\text{LogErr}_{\text{un}}$ error due to the underestimation of noise PSD. Subfigures 1b, 1c and 1d represent the segmental SNR, segmental speech SNR and segmental NR respectively

6. DISCUSSION AND CONCLUSION

The estimation of noise PSD is a very critical component of a speech enhancement system. Thus, in this paper, we have evaluated four noise PSD estimators for single channel speech enhancement in some of the typically encountered background noises. The evaluated algorithms consisted of MS, MMSE, IMCRA and a new model based method. It was observed that the model-based method outperformed other algorithms in terms of the spectral estimation accuracy for all the noise types. In terms of the enhancement performance, the model-based approach outperformed the other algorithms for relatively non-stationary noises such as babble and restaurant noise

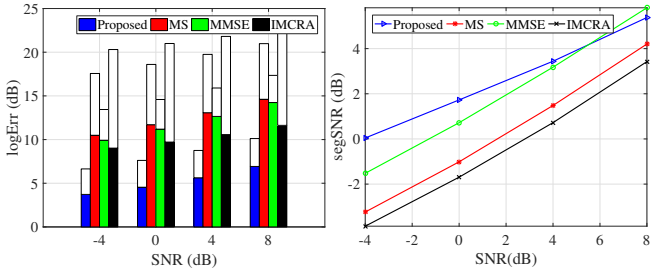


(a) Log estimation error (b) Segmental SNR

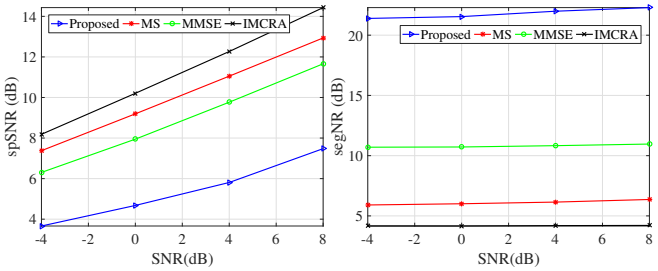


(c) Segmental Speech SNR (d) Segmental NR

Fig. 2: Performance measures for different algorithms for restaurant noise

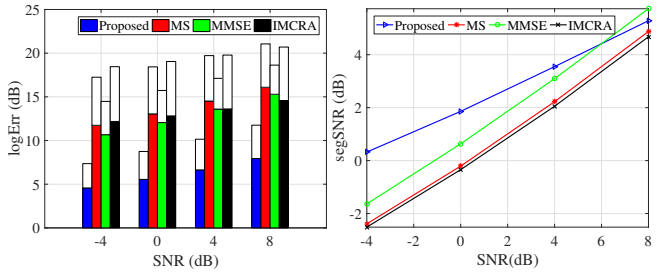


(a) Log estimation error (b) Segmental SNR

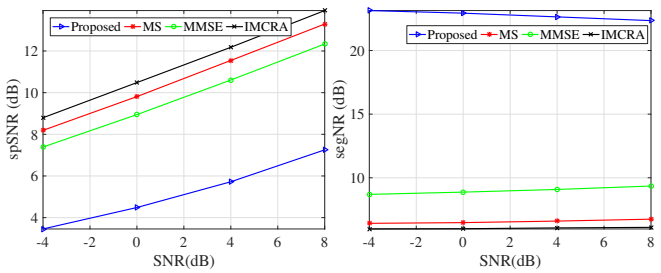


(c) Segmental Speech SNR (d) Segmental NR

Fig. 4: Performance measures for different algorithms for street noise

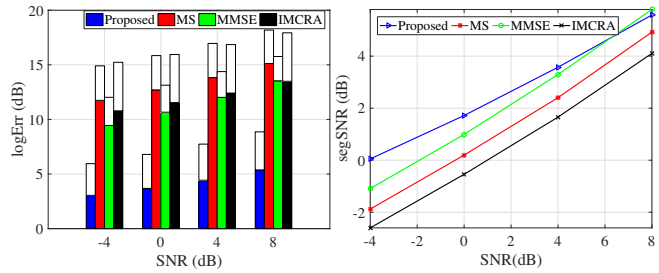


(a) Log estimation error (b) Segmental SNR

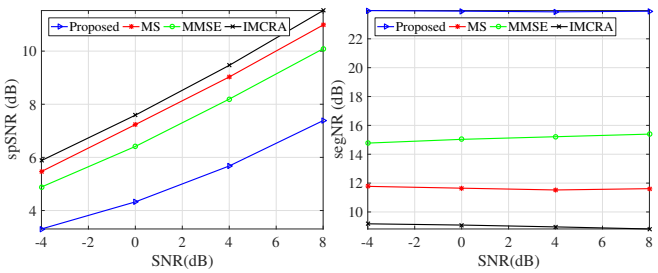


(c) Segmental Speech SNR (d) Segmental NR

Fig. 3: Performance measures for different algorithms for exhibition noise



(a) Log estimation error (b) Segmental SNR



(c) Segmental Speech SNR (d) Segmental NR

Fig. 5: Performace measures of the different algorithms for station noise.

irrespective of the SNR. In the case of more stationary noise types such as station and street noise, the benefit of using the model-based approach is observed only in lower SNRs.

7. REFERENCES

[1] J. Benesty, J. R. Jensen, M. G. Christensen, and J. Chen, *Speech enhancement: A signal subspace perspective*, Elsevier, 2014.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a

minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[3] P. C. Loizou, *Speech enhancement: theory and practice*, CRC press, 2013.

[4] J. K. Nielsen, M. S. Kavalekalam, M. G. Christensen, and J. B. Boldt, "Model-based noise psd estimation from speech in non-stationary noise," in *IEEE International Conference on Acoustics, Speech and Signal Processing. Proceedings*, 2018.

[5] J. Taghia, J. Taghia, N. Mohammadiha, J. Sang, V. Bouse, and R. Martin, "An evaluation of noise power spectral density estimation algorithms in adverse acoustic environments," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 4640–4643.

[6] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on speech and audio processing*, vol. 9, no. 5, pp. 504–512, 2001.

[7] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Transactions on speech and audio processing*, vol. 11, no. 5, pp. 466–475, 2003.

[8] T. Gerkmann and R. C. Hendriks, "Unbiased mmse-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, 2012.

[9] P. Stoica, R. L. Moses, et al., *Spectral analysis of signals*, vol. 452, Pearson Prentice Hall Upper Saddle River, NJ, 2005.

[10] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 163–176, 2006.

[11] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-based bayesian speech enhancement for nonstationary environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 441–452, 2007.

[12] R. M. Gray et al., "Toeplitz and circulant matrices: A review," *Foundations and Trends® in Communications and Information Theory*, vol. 2, no. 3, pp. 155–239, 2006.

[13] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Machine learning*, vol. 37, no. 2, pp. 183–233, 1999.

[14] C. M. Bishop, *Pattern recognition and machine learning*, springer, 2006.

[15] D. Chan, A. Fourcin, D. Gibbon, B Granstrom, et al., "Eurom-a spoken language resource for the eu," in *Proceedings of the 4th European Conference on Speech Communication and Speech Tecnology, Eurospeech'95*, 1995, pp. 867–880.

[16] Y. Hu and P. C. Loizou, "Subjective comparison of speech enhancement algorithms," in *Acoustics, Speech and Signal Processing (ICASSP), 2006 IEEE International Conference on*. IEEE, 2006.

[17] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Communications*, vol. 28, no. 1, pp. 84–95, 1980.

[18] T. Lotter and P. Vary, "Speech enhancement by map spectral amplitude estimation using a super-gaussian speech model," *EURASIP journal on applied signal processing*, vol. 2005, pp. 1110–1126, 2005.

[19] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction wiener filter," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1218–1234, 2006.