Model-based Noise PSD Estimation from Speech in Non-stationary Noise

Nielsen, Jesper Kjær; Kavalekalam, Mathew Shaji; Christensen, Mads Græsbøll; Boldt, Jesper Bünsow

Published in:
IEEE International Conference on Acoustics, Speech, and Signal Processing

DOI (link to publication from Publisher):
10.1109/ICASSP.2018.8461683

Creative Commons License
Unspecified

Publication date:
2018

Document Version
Accepted author manuscript, peer reviewed version

MODEL-BASED NOISE PSD ESTIMATION FROM SPEECH IN NON-STATIONARY NOISE

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

[1]Audio Analysis Lab, CREATE
Aalborg University, Denmark
{jkn,msk,mgc}@create.aau.dk

[3]GN ReSound
Denmark
jboldt@gnresound.com

ABSTRACT

Most speech enhancement algorithms need an estimate of the noise power spectral density (PSD) to work. In this paper, we introduce a model-based framework for doing noise PSD estimation. The proposed framework allows us to include prior spectral information about the speech and noise sources, can be configured to have zero tracking delay, and does not depend on estimated speech presence probabilities. This is in contrast to other noise PSD estimators which often have a too large tracking delay to give good results in non-stationary situations and offer no consistent way of including prior information about the speech or the noise type. The results show that the proposed method outperforms state-of-the-art noise PSD estimators in terms of tracking speed and estimation accuracy.

Index Terms— Noise PSD estimation, speech enhancement, noise statistics.

1. INTRODUCTION

The healthy human auditory system has a remarkable ability to extract the desirable information from a noisy speech signal. Even in situations such as a cocktail party where the background noise is non-stationary and the signal-to-noise ratio (SNR) is very low, normal hearing people are not only able to cope with the situation, but able to enjoy it. For people with a hearing defect, however, noisy situations such as a cocktail party are often mentally fatiguing and very challenging to deal with. These hearing impaired people often rely on a hearing aid for the speech enhancement, but the performance of the current hearing aid technology is far from enabling its users to thrive in difficult situations such as a cocktail party. Speech enhancement is not only important to the hearing impaired person in a cocktail party situation, but in any situation where the desired speech is observed in noise. Moreover, not only humans benefit from speech enhancement since, e.g., speaker identification and speech recognition algorithms are often designed for a clean speech signal [1].

Any speech enhancement algorithm must incorporate some prior knowledge in order to successfully separate the desired speech from the unwanted background noise. For example, the popular Wiener filter and many other speech enhancement algorithms such as maximum SNR, MVDR, and LCMV [1] (see also [2] for a comparison) assume that the second-order statistics of the speech and/or noise are known somehow. In practice, however, the statistics is often unknown and time-varying. Therefore, the prior knowledge must be represented in an alternative way so that the statistics can be estimated directly from the noisy speech. In this paper, we make contributions to the solution of exactly this problem.

The work was partly sponsored by Innovation Fund Denmark (Grant No. 99-2014-1).

Many people have been analysing the problem of estimating the noise power spectral density (PSD) or, equivalently, the second-order noise statistics, from a noisy speech signal. The most basic approach to estimating the noise PSD has been to use a voice activity detector (VAD) to inform the estimation algorithm about when speech is absent so that the noise PSD can be estimated. Unfortunately, such VADs are often difficult to tune in low SNRs, and they do not work well when the noise is non-stationary [3, 4]. Moreover, they are inefficient since they typically disable the noise PSD estimator across the entire frequency range, even if speech is only present in a few frequency bands. This has motivated the use of a soft VAD in each frequency band. A prominent example of this is the minimum statistics (MS) method [3, 5]. The algorithm is built on the assumption that the noise PSD is slowly varying with time and that the power of the noisy signal frequently goes down to the noise power level. Although the MS principle is simple, a lot of heuristics go into estimating a very important smoothing parameter and to correct the negative bias of the estimator. In fact, a full journal paper has been published on the latter issue [6]. Other problems with the principle are that the variance of the estimated noise PSD is bigger than for other methods [3, 4] and that very long tracking delays can occur, in particular when the noise power is increasing. Precisely these two issues were addresses in the MCRA [7–9] and later in the improved MCRA (IMCRA) [4] methods. Unfortunately, however, there might still be a considerable tracking delay in IMCRA if the noise power is increasing [10] and a lot of hand-crafting is still involved in tuning the algorithm and in doing bias correction. In [10, 11], the MS principle was abandoned in favour of MMSE estimators. These MMSE estimators were demonstrated to have a much better tracking speed than the MS and IMCRA methods and can be considered to be the best noise tracker currently [12]. One of the disadvantages of the MMSE estimators is that the first five time frames are assumed to be noise only to initialise the tracker. Another disadvantage is that it is not clear what prior information is actually built into the MMSE estimators about the speech and the noise, besides that the speech and noise spectral coefficients are modelled as independent and normally distributed random variables. This model assumption is very common in noise PSD estimation, but does not by itself enable us to separate a mixture into its components. Additional prior information is, therefore, necessary to find a unique solution to the problem, but the current noise trackers often rely on heuristic tricks for making the problem solvable rather than explicitly stating the model assumptions. Approaches based on, e.g., vector Taylor series [13] or nonnegative matrix factorisations (NMF) [14] give such model based estimates of the noise statistics via a separate training step. The clear advantage of these approaches is that it is much easier to understand the applicability and limitations of the model and, consequently, the noise PSD estimator. Moreover, we do not have to compensate for artefacts such as an unwanted bias, and we can change the built-in prior information via the model. For example, a

hearing aid user often communicate with the same people, but such information cannot be built into current noise PSD trackers.

In this paper, we propose a new noise PSD estimator which has some resemblance to both the NMF approach and the MMSE estimators. However, we derive our estimator directly in a flexible statistical framework which can be used in situations where we have specific prior information, but also in situations where we do not. By virtue of being model-based, we can in principle also use the proposed framework for noise PSD estimation with no tracking delay, even if speech is continuously present.

2. THE ESTIMATION PROBLEM AND THE MODEL

We assume that we observe N samples from the noisy speech signal

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

where $\boldsymbol{y} \in \mathbb{R}^{N\times 1}$, $\boldsymbol{s} \in \mathbb{R}^{N\times 1}$, and $\boldsymbol{e} \in \mathbb{R}^{N\times 1}$ are the noisy speech, the clean speech, and the noise, respectively. Given $\boldsymbol{y}$, we seek to estimate the noise PSD which is typically defined as [15, p. 7]

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

where $\mathbb{E}$ is the expectation operator and $E(\omega) = \boldsymbol{f}^H(\omega)\boldsymbol{e}$ is the DFT of the noise with $\boldsymbol{f}(\omega) = \{\exp(j\omega n)\}_{n=0,\ldots,N-1}$. The conditional expectation in (2) is the second moment of the density $p(E(\omega)|\boldsymbol{y})$. However, it can also be written in terms of the density $p(\boldsymbol{e}|\boldsymbol{y})$ as

$$\mathbb{E}\left[|E(\omega)|^2|\boldsymbol{y}\right] = \boldsymbol{f}^H(\omega)\left[\int_{\mathbb{R}^{N\times 1}} \boldsymbol{e}\boldsymbol{e}^T p(\boldsymbol{e}|\boldsymbol{y}) d\boldsymbol{e}\right]\boldsymbol{f}(\omega)\ . \tag{3}$$

The problem of estimating the noise PSD is, therefore, essentially that of computing the second moment of the posterior $p(\boldsymbol{e}|\boldsymbol{y})$.

To compute the posterior $p(\boldsymbol{e}|\boldsymbol{y})$, we elicit several statistical models $\{\mathcal{M}_k\}_{k=1}^K$ for how the data vector $\boldsymbol{y}$ was generated. Such models can easily be included in (3) as

$$\mathbb{E}\left[|E(\omega)|^2\big|\boldsymbol{y}\right] = \sum_{k=1}^{K} p(\mathcal{M}_k|\boldsymbol{y})\,\mathbb{E}\left[|E(\omega)|^2\big|\boldsymbol{y},\mathcal{M}_k\right]\ . \tag{4}$$

Thus, we obtain a model averaged noise PSD estimator if we insert (4) in (2). The model probabilities $\{p(\mathcal{M}_k|\boldsymbol{y})\}_{k=1}^K$ ensure that those models which explain the data well will contribute with a larger weight than those models which do not explain the data well. In principle, there are no limits on which models can be used. From a practical perspective, however, it is advantageous to use models that lead to tractable algorithms while still being a sufficiently accurate representation of how the speech and the noise were generated. In this paper, we will use autoregressive processes to model the speech and the noise, i.e.,

$$p(\boldsymbol{s}|\sigma^2_{\mathrm{s},k},\mathcal{M}_k) = \mathcal{N}(\boldsymbol{0},\sigma^2_{\mathrm{s},k}\boldsymbol{R}_{\mathrm{s}}(\boldsymbol{a}_k)) \tag{5}$$

$$p(\boldsymbol{e}|\sigma^2_{\mathrm{e},k},\mathcal{M}_k) = \mathcal{N}(\boldsymbol{0},\sigma^2_{\mathrm{e},k}\boldsymbol{R}_{\mathrm{e}}(\boldsymbol{b}_k)) \tag{6}$$

where $\sigma^2_{\mathrm{s},k}$, $\sigma^2_{\mathrm{e},k}$, $\boldsymbol{R}_{\mathrm{s}}(\boldsymbol{a}_k)$, $\boldsymbol{R}_{\mathrm{e}}(\boldsymbol{b}_k)$, $\boldsymbol{a}_k$, and $\boldsymbol{b}_k$ are the excitation noise variances, the normalised covariance matrices, and the AR-parameters of the speech and the noise, respectively. We assume that the AR-processes are periodic in N since the normalised covariance matrices then are diagonalised by the DFT matrix $\boldsymbol{F}$. That is,

$$\boldsymbol{R}_{\mathrm{s}}(\boldsymbol{a}_k) = N^{-1}\boldsymbol{F}\boldsymbol{D}_{\mathrm{s}}(\boldsymbol{a}_k)\boldsymbol{F}^H \tag{7}$$

$$[\boldsymbol{F}]_{nl} = \exp(j2\pi(n-1)(l-1)/N)\ , \quad n,l = 1,\ldots,N \tag{8}$$

$$\boldsymbol{D}_{\mathrm{s}}(\boldsymbol{a}_k) = \left(\boldsymbol{\Lambda}^H_{\mathrm{s}}(\boldsymbol{a}_k)\boldsymbol{\Lambda}_{\mathrm{s}}(\boldsymbol{a}_k)\right)^{-1} \tag{9}$$

$$\boldsymbol{\Lambda}_{\mathrm{s}}(\boldsymbol{a}_k) = \mathrm{diag}(\boldsymbol{F}^H\left[\boldsymbol{a}_k^T \quad \boldsymbol{0}\right]^T) \tag{10}$$

with similar definitions for $\boldsymbol{R}_{\mathrm{e}}(\boldsymbol{b}_k)$. Although it might seem unfounded to assume periodicity in N, this assumption is actually implicitly made when using the asymptotic covariance matrix of an AR-process for finite length signals as in [16] or when interpreting the Itakura-Saito (IS) distortion measure [17, 18] as the maximum likelihood estimator of short-time speech spectra. Precisely the IS distortion measure has been very popular in the speech community for decades, partly due to it also being a perceptually meaningful distortion measure [19], and has lately also been used successfully as a distortion measure for nonnegative matrix factorisation (NMF) [20]. Moreover, the above model actually has the signal model used in [10, 11] as a special case. Specifically, if we select $K = 1$ and set the AR-orders to $N-1$, then the speech and noise spectral coefficients are modelled as independent and normally distributed random variables and the noise PSD estimator in (2) is the foundation of the MMSE-estimators in [10, 11]. As discussed in the introduction, however, this frequency domain model does not by itself allow us to separate the noisy mixture into its components.

2.1. Prior Information

Inspired by the work in [16, 21], the AR-parameters are here assumed known for a given model. Thus, a model in our framework corresponds to one combination of so-called codebook entries in the framework of [16, 21]. That is, if we have a speech and a noise codebook consisting of K_{s} and K_{e} trained AR-vectors, respectively, we have a total of $K = K_{\mathrm{s}}K_{\mathrm{e}}$ models[1]. At first glance, it might seem a disadvantage that these codebooks have to be trained, but they actually offer an excellent way of including prior spectral information. For example, if the noise PSD estimator has to operate in a particular noise environment such as a car cabin or mainly process speech from a single person such as in mobile telephony, we can use codebooks with typical normalised AR-spectra for these sources. Conversely, in the absence of any specific information about the speaker(s) and the noise environment(s), we can use classified codebooks [16] where we first classify the speaker/noise type and then use the corresponding codebooks which have been trained on different speakers and noise types. Moreover, the noise PSD estimate from any noise tracker can also be included as a noise codebook vector. This also means that the proposed framework can be used to combine existing noise trackers in a consistent fashion. A potential problem of the model-based approach is that the number of models grows with the product of the codebook sizes, and this might lead to an intractable computational complexity. This is also one of the reasons why we use models whose covariance matrices can be diagonalised by the DFT matrix.

The excitation noise variances are not pre-trained, but are treated as unknown random variables with the prior

$$p(\sigma^2_{\mathrm{s},k}|\mathcal{M}_k) = \text{Inv-}\mathcal{G}(\alpha_{\mathrm{s},k},\beta_{\mathrm{s},k}) \tag{11}$$

$$p(\sigma^2_{\mathrm{e},k}|\mathcal{M}_k) = \text{Inv-}\mathcal{G}(\alpha_{\mathrm{e},k},\beta_{\mathrm{e},k}) \tag{12}$$

where Inv-$\mathcal{G}(\cdot,\cdot)$ denotes the inverse Gamma density. Similarly, we also have a prior mass function $p(\mathcal{M}_k)$ for the models. Speech is normally processed frame-by-frame, often with some overlap. Consequently, values for the excitation noise variances and models that work well in one frame, should also work reasonably well in the

[1] Note that a codebook is not restricted to only include AR-spectra, but can in principle include any type of spectrum as in NMF. We here focus on a parametric representation of the spectra in terms of AR-parameters since this leads to codebooks with a small memory footprint, can be used for short segment sizes, and allows us to train the codebooks using standard vector quantisation techniques developed for speech coding [22].

next frame, and the priors are an excellent tool for using previous information in the current frame. In a completely stationary environment, for example, the posterior distribution of one frame should be the prior distribution in the next frame. The more non-stationary the signals are, the broader the prior of the current frame should be compared to the posterior of the previous frame. In the limit, no information is carried over from one frame to the next, and we use uninformative priors with $\alpha_{\cdot,k} = \beta_{\cdot,k} \to 0$ and $p(\mathcal{M}_k) = K^{-1}$. In this paper, we focus on exactly this limiting case in the simulations. Besides not having enough space here to give a complete description of a general frame transition model, this choice is motivated by that 1) babble noise is typically very non-stationary, and 2) we wish to demonstrate that the proposed model-based approach works well, even without any smoothing between frames. This is in contrast to current state-of-the-art noise trackers which at best have a tracking delay of a few hundred milliseconds [10]. Before going to the simulations, however, we first describe how the noise PSD is estimated from the model and the data.

3. INFERENCE

To estimate the noise PSD, we have to compute the posterior model probabilities $p(\mathcal{M}_k|\boldsymbol{y})$ as well as the second moment of the posterior $p(\boldsymbol{e}|\boldsymbol{y}, \mathcal{M}_k)$ (see (4)) by combining the information in the data with the prior information. Unfortunately, neither of these posteriors exist in closed-form, and we, therefore, have to content ourselves with either analytical or stochastic approximations. For our inference problem, the variational Bayesian (BS) framework [23,24] produces a simple analytical approximation if we assume that the full joint posterior factorises as

$$p(\boldsymbol{e}, \sigma^2_{\mathrm{s},k}, \sigma^2_{\mathrm{e},k}|\boldsymbol{y}, \mathcal{M}_k)p(\mathcal{M}_k|\boldsymbol{y}) \approx q(\boldsymbol{e}|\boldsymbol{y}, \mathcal{M}_k)q(\sigma^2_{\mathrm{s},k}, \sigma^2_{\mathrm{e},k}|\boldsymbol{y}, \mathcal{M}_k)q(\mathcal{M}_k|\boldsymbol{y}) \; . \quad (13)$$

Unfortunately, the derivation of the three factors in the approximation is lengthy so we only state the results here and refer the interested reader to a supplementary document for a detailed derivation (available at `http://tinyurl.com/jknvbn`). From the derivation, we obtain that the posterior factor $q(\boldsymbol{e}|\boldsymbol{y}, \mathcal{M}_k)$ is given by

$$q(\boldsymbol{e}|\boldsymbol{y}, \mathcal{M}_k) = \mathcal{N}(\hat{\boldsymbol{e}}_k, \hat{\boldsymbol{\Sigma}}_k) \quad (14)$$

where

$$\hat{\boldsymbol{\Sigma}}_k = \left[\frac{a_{\mathrm{s},k}}{b_{\mathrm{s},k}}\boldsymbol{R}_{\mathrm{s}}^{-1}(\boldsymbol{a}_k) + \frac{a_{\mathrm{e},k}}{b_{\mathrm{e},k}}\boldsymbol{R}_{\mathrm{e}}^{-1}(\boldsymbol{b}_k)\right]^{-1} \quad (15)$$

$$\hat{\boldsymbol{e}}_k = \frac{a_{\mathrm{s},k}}{b_{\mathrm{s},k}}\hat{\boldsymbol{\Sigma}}_k\boldsymbol{R}_{\mathrm{s}}^{-1}(\boldsymbol{a}_k)\boldsymbol{y} \; . \quad (16)$$

The scalars $a_{\mathrm{s},k}$, $b_{\mathrm{s},k}$, $a_{\mathrm{e},k}$, and $b_{\mathrm{e},k}$ are obtained from the factor $q(\sigma^2_{\mathrm{s},k}, \sigma^2_{\mathrm{e},k}|\boldsymbol{y}, \mathcal{M}_k)$ which is given by

$$q(\sigma^2_{\mathrm{s},k}, \sigma^2_{\mathrm{e},k}|\boldsymbol{y}, \mathcal{M}_k) = \text{Inv-}\mathcal{G}(a_{\mathrm{s},k}, b_{\mathrm{s},k})\text{Inv-}\mathcal{G}(a_{\mathrm{e},k}, b_{\mathrm{e},k}) \quad (17)$$

where

$$a_{\mathrm{s},k} = \alpha_{\mathrm{s},k} + N/2 \quad (18)$$

$$b_{\mathrm{s},k} = \beta_{\mathrm{s},k} + \left[\hat{\boldsymbol{s}}_k^T\boldsymbol{R}_{\mathrm{s}}^{-1}(\boldsymbol{a}_k)\hat{\boldsymbol{s}}_k + \mathrm{tr}\left(\boldsymbol{R}_{\mathrm{s}}^{-1}(\boldsymbol{a}_k)\hat{\boldsymbol{\Sigma}}_k\right)\right]/2 \quad (19)$$

$$a_{\mathrm{e},k} = \alpha_{\mathrm{e},k} + N/2 \quad (20)$$

$$b_{\mathrm{e},k} = \beta_{\mathrm{e},k} + \left[\hat{\boldsymbol{e}}_k^T\boldsymbol{R}_{\mathrm{e}}^{-1}(\boldsymbol{b}_k)\hat{\boldsymbol{e}}_k + \mathrm{tr}\left(\boldsymbol{R}_{\mathrm{e}}^{-1}(\boldsymbol{b}_k)\hat{\boldsymbol{\Sigma}}_k\right)\right]/2 \quad (21)$$

$$\hat{\boldsymbol{s}}_k = \boldsymbol{y} - \hat{\boldsymbol{e}}_k \; . \quad (22)$$

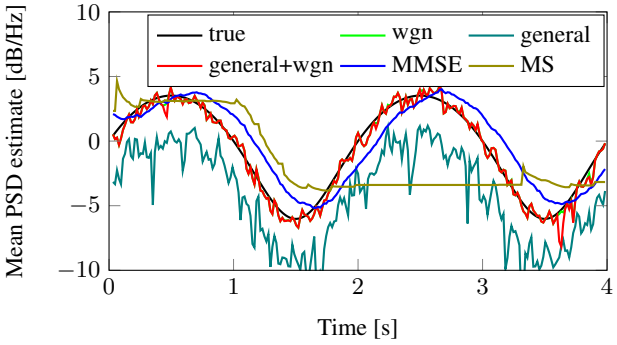

Fig. 1. Estimates of the noise variance for modulated white Gaussian noise. The displayed results are averaged over frequency.

The above solution is not a closed-form solution for the parameters of the posterior factors. Instead, these are computed iteratively, and the VB framework guarantees that the algorithm converges to a mode. Since the normalised covariance matrices are diagonalised with the DFT matrix, we can easily evaluate the matrix inverses and the traces above. An interesting observation is that the VB algorithm essentially performs Wiener filtering in (16). Convergence of the VB algorithm can be monitored via the variational lower bound $\mathcal{L}_k$ which is related to the posterior model factor as

$$q(\mathcal{M}_k|\boldsymbol{y}) \propto \exp(\mathcal{L}_k)p(\mathcal{M}_k) \; . \quad (23)$$

Unfortunately, the variational lower bound consists of many terms so we refer the interested reader to the supplementary document for the full expression.

Since the posterior factor $q(\boldsymbol{e}|\boldsymbol{y}, \mathcal{M}_k)$ is a normal distribution, its second moment is

$$\mathbb{E}[\boldsymbol{e}\boldsymbol{e}^T|\boldsymbol{y}, \mathcal{M}_k] = \hat{\boldsymbol{e}}_k\hat{\boldsymbol{e}}_k^T + \hat{\boldsymbol{\Sigma}}_k \; . \quad (24)$$

Inserting this and the posterior model factor in (4) and (2) gives

$$\phi_{\mathrm{e}}(\omega) \approx \frac{1}{N}\sum_{k=1}^{K} q(\mathcal{M}_k|\boldsymbol{y})\left[|\boldsymbol{f}^H(\omega)\hat{\boldsymbol{e}}_k|^2 + \boldsymbol{f}^H(\omega)\hat{\boldsymbol{\Sigma}}_k\boldsymbol{f}(\omega)\right]$$

where we have ignored the limit operator. This PSD estimator is essentially a model-averaged version of the MMSE estimators in [10, 11]. However, the proposed estimator does not depend on threshold parameters to avoid stagnation, on bias compensation, or on unknown parameters which have to be estimated by computing speech presence probabilities. Moreover, the proposed estimator has a consistent way of including prior spectral information in the form of codebooks, and it works for a single data frame, even for uninformative prior distributions on all the excitation noise variances.

4. EVALUTATION

This paper has focused on motivating and deriving the proposed noise PSD estimator. Therefore, there is only a limited space left to provide evidence for that the fundamental principle works, but we have a more thorough evaluation in [25]. Here, we consider two different experiments. First, we demonstrate that the proposed noise PSD estimator works with zero tracking delay. Second, we apply the proposed noise PSD estimator to the difficult problem of estimating the PSD of babble noise from a noisy mixture. In both experiments, the speech codebook consisted of 64 AR vectors of order 14. It was trained using a variation of the LBG-algorithm

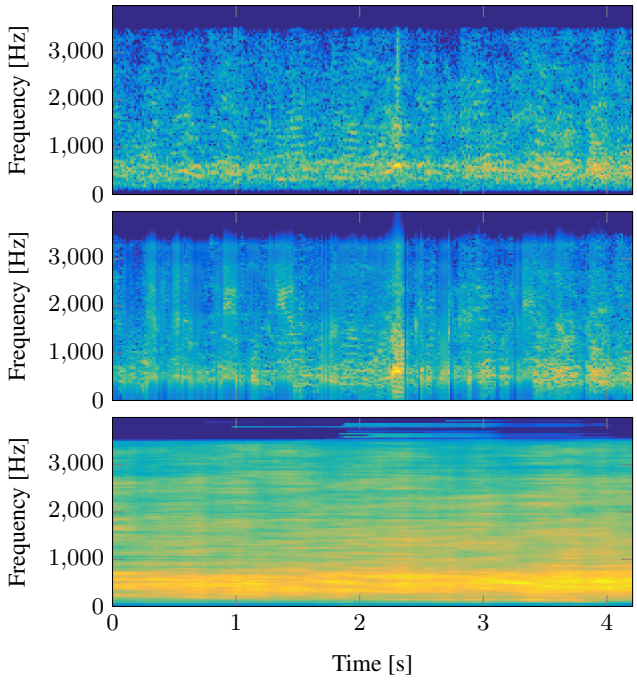

Fig. 2. The spectrogram of the babble noise PSD (top) compared to the noise PSD estimates of the proposed (middle) and MMSE methods (bottom).

method [26] on both male and female speech from the EUROM English database [27]. A noise codebook consisting of 12 AR vector of order 14 was trained on different noise types from the NOIZEUS database [28]. These noise types included restaurant, exhibition, street, and station noise. Thus, we did not train the codebook on babble noise which we are using for testing in the second experiment. As alluded to in Sec. 2.1, we used non-informative priors corresponding to no smoothing between frames. The codebooks as well as the MATLAB code for generating the presented results are available at `http://tinyurl.com/jknvbn`.

4.1. Tracking speed

The first experiment assessed the tracking speed of the estimator and is very similar to the first experiment in [10]. Thus, we estimated the noise power of modulated white Gaussian noise where the noise variance was time-varying with a frequency of 2 Hz. We compared the proposed method for three different noise codebooks to the MMSE method [10] and the MS method [3]. For the proposed method, the three different noise codebooks were a) a codebook consisting of only one entry modelling a flat spectrum; b) the noise codebook described above; and c) a combination of a) and b). Fig. 1 shows the results for the various noise PSD estimates averaged over frequency. As in [10], it is observed that MS tracked the noise variance poorly and that the MMSE method tracked much better, but with a delay of a few hundred milliseconds. On the other hand, the proposed method with noise codebook a) and c) had no tracking delay and produced visually identical results. The latter observation suggests that the algorithm assigned all weight to the true model and used that for estimating the noise PSD. Finally, the proposed method with noise codebook b) underestimated the noise variance and had a much larger variance. This illustrates that we get a degraded performance if we use incorrect prior information in the codebook.

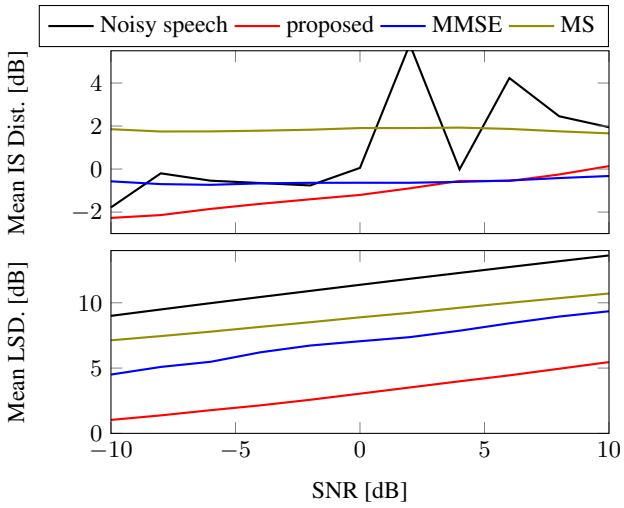

Fig. 3. The IS distance and LSD between the babble noise spectrogram and estimated noise PSDs for various methods.

4.2. Babble noise PSD estimation

In the second experiment, we estimated the babble noise PSD from a mixture of speech and babble noise at different SNRs in steps of 2 dB from -10 dB to 10 dB. The babble noise was taken from the NOIZEUS database [28] and the speech signal was taken from the CHiME database [29]. Thus, neither of these signals were used for training the codebooks. For every SNR, we measured the average Itakura-Saito (IS) distance and the average log-spectral distortion (LSD) between the babble noise spectrogram and the estimated noise PSD for four different methods using the default MMSE method settings of 32 ms windows with a 50 % overlap. Aside from the proposed, the MMSE, and the MS methods, we also used the spectrogram of the observed mixture as a reference method. The results are shown in Fig. 2 and Fig. 3. In Fig. 2, we have plotted the babble noise spectrogram (top), the proposed noise PSD estimate (middle), and the MMSE PSD estimate (bottom) for an SNR of 0 dB. Clearly, the proposed PSD estimate contains many more details than the MMSE PSD estimate. For example, there is a short burst in the babble noise at around 2.3 s which was captured by the proposed method, but smoothed out by the MMSE method. In Fig. 3, the performance of the different estimators are quantified in terms of the IS distance and the LSD. The proposed method outperformed the other methods, except for the IS distance for an SNR above 3 dB where the proposed method and the MMSE method have similar performance.

5. CONCLUSION

In this paper, we have developed a framework for doing noise PSD estimation using parametric models. These models offer a way of including prior information into the estimator to obtain a better estimation accuracy. More concretely, we proposed a class of models based on pre-training codebooks. These codebooks contained typical spectra for the speech and the noise, but could in principle also include the PSD estimates from other estimators. The developed framework also contained model comparison to ensure that models which explain the data well have a larger weight in the model averaged noise PSD estimate. Via two experiments, we demonstrated the potential applicability and improvements in the tracking speed and estimation accuracy over two state-of-the-art methods.

6. REFERENCES

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

[2] K. B. Christensen, M. G. Christensen, J. B. Boldt, and F. Gran, "Experimental study of generalized subspace filters for the cocktail party situation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2016, pp. 420–424.

[3] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 9, no. 5, pp. 504–512, 2001.

[4] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, 2003.

[5] R. Martin, "Spectral subtraction based on minimum statistics," in *Proc. European Signal Processing Conf.*, 1994, vol. 6, pp. 1182–1185.

[6] R. Martin, "Bias compensation methods for minimum statistics noise power spectral density estimation," *Elsevier Signal Processing*, vol. 86, no. 6, pp. 1215–1229, 2006.

[7] I. Cohen and B. Berdugo, "Spectral enhancement by tracking speech presence probability in subbands," in *International Workshop on Hands-Free Speech Communication*, 2001, pp. 95–98.

[8] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Elsevier Signal Processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

[9] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Process. Lett.*, vol. 9, no. 1, pp. 12–15, 2002.

[10] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, 2012.

[11] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2010, pp. 4266–4269.

[12] 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 *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2011, pp. 4640–4643.

[13] P. J. Moreno, B. Raj, and R. M. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 1996, vol. 2, pp. 733–736.

[14] 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, 2013.

[15] P. Stoica and R. L. Moses, *Spectral Analysis of Signals*, Englewood Cliffs, NJ, USA: Prentice Hall, May 2005.

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

[17] F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelihood method," in *Proc. Int. Congr. Acoust.*, 1968, pp. 17–20.

[18] F. Itakura and S. Saito, "A statistical method for estimation of speech spectral density and formant frequency," *Trans. Inst. Electron. Commun. Eng. (Japan)*, vol. 53, no. 1, pp. 36–43, 1970.

[19] R. Gray, A. Buzo, A. Gray, and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 367–376, 1980.

[20] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[21] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-based Bayesian speech enhancement for nonstationary environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 441–452, 2007.

[22] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Dordrecht, The Nederlands: Kluwer Academic Publishers, 1992.

[23] 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.

[24] C. M. Bishop, *Pattern Recognition and Machine Learning*, New York, NY, USA: Springer, Aug. 2006.

[25] M. S. Kavalekalam, J. K Nielsen, M. G. Christensen, and J. B. Boldt, "A study of noise PSD estimators for single channel speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018.

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

[27] D. Chan, A. Fourcin, D. Gibbon, B. Granstrom, M. Huckvale, G. Kokkinakis, K. Kvale, L. Lamel, B. Lindberg, A. Moreno, J. Mouropoulos, F. Senia, I. Trancoso, C. Veld, and J. Zeiliger, "EUROM-A spoken language resource for the EU," in *Proc. European Conf. Speech Commun. and Speech Technol.*, 1995, pp. 867–880.

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

[29] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE Workshop on Autom. Speech Recog. and Underst.*, 2015, pp. 504–511.