# Noise-Presence-Probability-Based Noise PSD Estimation by Using DNNs

*Aleksej Chinaev, Jahn Heymann, Lukas Drude, Reinhold Haeb-Umbach*

Department of Communications Engineering, Paderborn University, 33100 Paderborn, Germany
Email: {chinaev,heymann,drude,haeb}@nt.upb.de
Web: nt.upb.de

## Abstract

A noise power spectral density (PSD) estimation is an indispensable component of speech spectral enhancement systems. In this paper we present a noise PSD tracking algorithm, which employs a noise presence probability estimate delivered by a deep neural network (DNN). The algorithm provides a causal noise PSD estimate and can thus be used in speech enhancement systems for communication purposes. An extensive performance comparison has been carried out with ten causal state-of-the-art noise tracking algorithms taken from the literature and categorized acc. to applied techniques. The experiments showed that the proposed DNN-based noise PSD tracker outperforms all competing methods with respect to all tested performance measures, which include the noise tracking performance and the performance of a speech enhancement system employing the noise tracking component.

## 1 Introduction

Noise power spectral density estimation is an essential component of any single-channel speech enhancement system using spectral subtractive or statistical-model-based algorithms [1]. The challenging task of noise PSD estimation in the presence of non-stationary noise has spurred the development of many sophisticated algorithms during the last years. A closer look into their functionality shows that they can be categorized along the following lines:

- Because of the sparseness of the clean speech PSD some noise trackers make use of a *minimum search* of the noisy PSD over a certain number of the previous frames, which are closely related to the desired noise PSD estimate [2–8].
- Other approaches employ a *voice activity detection (VAD)* or a *speech presence probability (SPP) estimation*, again exploiting the sparseness of speech, to find the noise-only time-frequency slots where the noise PSD estimate can be updated [3–6, 9–11].
- Due to the random nature of the signals, they are often modeled as realizations of random processes with given probability density function (PDF) enabling e.g. an analytical *bias compensation* of the noise PSD estimates [2, 4, 7, 8].
- Furthermore, the statistical modeling facilitates a *Bayesian inference* such as the minimum mean squared error (MMSE) estimators for noise PSD [7, 10].
- Since the short-time Fourier transform (STFT) coefficients of the noise signals are correlated in a certain neighbourhood (even for white noise), an *output smoothing*[1] becomes another very popular technique in the noise tracking [3–7, 9–11].

[1] In noise PSD tracking at least 3 types of smoothing can be distinguished: a smoothing for the *minimum search*, a smoothing as part of the *SPP estimation*, and a smoothing of PSD of a noisy signal resulting in estimates of a noise PSD tracker denoted here as the *output smoothing*.

A mandatory property of all approaches for noise PSD estimation, when used in communication scenarios, is its causality.

In recent years deep neural networks have made inroads in speech signal processing, and DNN-based approaches for speech enhancement have been developed [12]. The networks operate here usually as nonlinear filters mapping the noisy speech to clean speech as in [13]. Sometimes DNNs are combined with conventional speech enhancement techniques [14]. Recently, we effectively incorporated a DNN-based spectral mask estimation into a multi-channel speech enhancement system [15].

In this contribution we suggest to use a single-channel DNN-based noise presence probability (NPP) estimation for noise PSD tracking. To this extent we modify our mask estimation system of [15] to be causal and work on a single frame of a single-channel input signal. This asks for a replacement of the commonly used batch normalization of the input and/or hidden layers of the network with methods that do not compromise the latency of the system [16].

The remainder of this paper is structured as follows: In Section 2 we introduce mathematical notations and derive an NPP-based noise PSD estimator. Next, a causal DNN-based NPP estimation is introduced in Section 3. Further, in Section 4 we give an overview of ten state-of-the-art noise PSD estimators used in our experimental evaluation, and, after presentation of experimental results in Section 5, we draw some conclusions in Section 6.

## 2 NPP-based noise PSD estimation

We denote the periodogram of the noisy speech signal by $|Y(k,\ell)|^2$ and of the noise signal by $|D(k,\ell)|^2$ with a frequency bin index $k \in \{1,\cdots,K\}$ and a frame index $\ell \in \{1,\cdots,L\}$. The noise PSD is then defined as

$$\lambda_D(k,\ell) = E\left[|D(k,\ell)|^2\right], \tag{1}$$

where $E[\cdot]$ denotes the mathematical expectation operator. The main task of any noise PSD tracker is to estimate the noise PSD $\lambda_D(k,\ell)$ from the noisy PSD $|Y(k,\ell)|^2$ in a causal way, i.e., by using only past observations up to a current frame.

Assuming that a noise spectral mask $M_D(k,\ell)$ is given, and inspired by simplicity of the SPP-based noise PSD estimator proposed in [11] and summarized in [17], we propose a low-complexity NPP-based estimator using a recursive averaging

$$\hat{\lambda}_D(\ell) = (1 - M_D(\ell)) \cdot \hat{\lambda}_D(\ell-1) + M_D(\ell) \cdot |Y(\ell)|^2. \tag{2}$$

Since Eq. (2) is carried out for every frequency bin separately, we dropped the frequency index $k$ here. Note, that the noise spectral mask $M_D(\ell)$ in Eq. (2) plays the role of a time-varying smoothing parameter. Eq. (2) is similar to what is done in [11], where a speech presence probability estimate, so to say the complement to $M_D(\ell)$, is

used instead. But unlike [11, 17], $\hat{\lambda}_D(\ell)$ will not be further smoothed, as it will turn out that the NPP estimate delivered by a DNN is already very robust. This corresponds to the parameter choice $\alpha_{\text{pow}} = 0$ in [11] or $\beta = 0$ in [17].

The proposed noise PSD estimator uses the sparseness property of speech and applies a technique similar to *SPP estimation*. However we denoted our estimator as a NPP-based approach (and not as a SPP-based), since we make a distinction between NPP and speech absence probability (similar to [1], chapter 44.7.1).

# 3 Causal NPP estimation using DNN

As outlined in the previous section our proposed noise PSD estimator relies on a spectral mask $M_D(k,\ell)$ which indicates the probability of the presence of noise in the $k$-th frequency bin of frame number $\ell$. We propose to use a neural network to estimate this spectral mask.

In our previous work on a related task [15], we achieved the best results with a bi-directional Long Short-Term Memory network. However, this would limit our approach presented here to batch-processing as the whole utterance must be available to estimate the mask. To avoid this limitation and make the system causal, we omit the backward path (i.e. use a Long Short-Term Memory (LSTM) network) and double the number of units in this layer to allow the network to compensate for the missing backward units. Note, that due to the nature of LSTMs we are still able to exploit temporal dependencies through its internal state which is passed on to the next frame.

The scenario also prohibits us to use batch-normalization like in our previous works where we estimate the statistics over a whole utterance, even at test time. Instead, we normalize the input data using the statistics from the training data. Additionally we replace the Rectified Linear Unit (ReLU) activation function with the Exponential Linear Unit (ELU) activation function which has an effect similar to batch-normalization during training [18]. The resulting configuration of our LSTM network for the STFT window length of 1024 is summarized in Table 1.

The network input is a single frame of the magnitude spectrum from the noisy signal $|Y(k,\ell)|$. It then tries to estimate the NPP for every bin of this frame. To learn this relationship, we use ideal binary masks of noise as training targets which we calculate as

$$\text{IBM}_D(k,\ell) = \begin{cases} 1, & \frac{|S(k,\ell)|}{|D(k,\ell)|} < 10^{\text{th}_D} \\ 0, & \text{else} \end{cases} \tag{3}$$

where $S(k,\ell)$ are STFT coefficients of the clean speech signal. In this work, we empirically set the threshold $\text{th}_\mathbf{D}$ to $-1$. Thus we classify a time-frequency bin as noise-only, if it is significantly dominated by the noise signal. By doing so we preserve the time-frequency bins with weak energy of clean speech signal to be assigned to the noise signal. Using such binary masks during training leads to a conservative NPP estimate and a sparser mask with high contrasts as output.

Here, frequency bins with indices below 5 and above 500 (corresponding to frequencies below $\sim$ 78 Hz and above $\sim$ 7.8 kHz for a sample rate of 16 kHz) are always considered to contain noise. Further it should be mentioned, that while the training targets are either zero or one, the network output is continuous between zero and one.

**Table 1:** LSTM network configuration for NPP estimation

| Layer | Units | Type | Non-Linearity | $p_{\text{dropout}}$ |
|---|---|---|---|---|
| L1 | 512 | LSTM | Tanh | 0.5 |
| L2 | 1024 | FF | ELU | 0.5 |
| L3 | 1024 | FF | ELU | 0.5 |
| L4 | 513 | FF | Sigmoid | 0.0 |

The targets $\text{IBM}_D(k,\ell)$ from Eq. (3) are compared to the current output of the network $M_D(k,\ell)$ using a binary cross-entropy (BCE) cost

$$\begin{aligned}\text{BCE} = -\frac{1}{K}\frac{1}{L}\sum_{\ell=1}^{L}\sum_{k=1}^{K}\{&\text{IBM}_D(k,\ell)\cdot\log_2 M_D(k,\ell) \\ &+(1-\text{IBM}_D(k,\ell))\cdot\log_2(1-M_D(k,\ell))\}.\end{aligned} \tag{4}$$

We initialize the weights of all layers using a uniform distribution, i.e. $W \sim \mathscr{U}[-a,a]$. For the LSTM layer, the parameter $a$ is 0.04, while for the ELU layers and the last layer $a = \sqrt{6}/\sqrt{n_{\text{in}}+n_{\text{out}}}$, where $n_{\text{in}}$ and $n_{\text{out}}$ are the input and output size of each layer, respectively [19]. The biases are all initialized with zeros.

# 4 State-of-the-art noise PSD trackers

A very popular noise PSD tracker is the minimum statistics (MS) approach [2], whose first draft was published in [20]. As it is depicted in Table 2, the MS method implements a *minimum search* with previous averaging of the noisy PSD over time with a time-variant optimal smoothing constant and an elaborated *bias compensation*. Recently we proposed to use an alternative control function for calculation of the optimal smoothing constant resulting in the Bayesian-smoothed MS (BSMS) approach [8].

Another noise PSD estimator, presented in [9] and denoted further as a VAD recursive averaging (VAD-RA), applies an *output smoothing* of the noisy PSD controlled by a rough *VAD estimation* which indicates speech presence. Compared to [2] the noise PSD estimates of the VAD-RA approach is more smoothed. The same techniques are used by a SPP-based approach with fixed priors (SPP-FP) recently published in [11], where the authors propose to replace the hard decision of the VAD by a soft *SPP estimation* resulting in an unbiased MMSE-like estimator.

| | *Minimum search* | *VAD/SPP estimation* | *Bias compensation* | *Bayesian inference* | *Output smoothing* |
|---|---|---|---|---|---|
| MS-based [2, 8] | ✓ | | ✓ | | |
| VAD/SPP-based [9, 11] | | ✓ | | | ✓ |
| MCRA-based [3, 5, 6] | ✓ | ✓ | | | ✓ |
| IMCRA [4] | ✓ | ✓ | ✓ | | ✓ |
| MMSE-VAD [10] | | ✓ | | ✓ | ✓ |
| MMSE-BM [7] | ✓ | | ✓ | ✓ | ✓ |

**Table 2:** An overview of the techniques used in the ten state-of-the-art noise PSD estimators.

In contrast to [11], the *output smoothing* of the minima controlled recursive averaging (MCRA) algorithm is controlled by a *SPP estimation*, which is based on a previous *minimum search* technique [3]. Note, the MCRA approach employs all 3 types smoothing operations mentioned in Section 1. The MCRA method served as a corner stone for the development of a series of further noise PSD trackers. One of them, the enhanced MCRA (EMCRA) approach [5], aims to reduce the estimator's delayed response to an abrupt noise rise and to mitigate the speech leakage into the noise PSD estimates. For the SPP estimation to benefit from inter-frame correlations of the speech signal, [6] proposes to incorporate a first-order conditional maximum a posteriori (MAP) criterion into the MCRA noise tracker resulting in the MCRA-MAP approach.

Another well-known MCRA-based noise PSD tracker developed by the author of the MCRA method is an improved MCRA (IMCRA) approach [4], which upgrades the *minimum tracking* in speech activity and the *SPP estimation* of the MCRA noise tracker. Additionally IMCRA approach implements a sophisticated *bias compensation* not available in the MCRA method.

Using *Bayesian inference* for the estimation of the noise PSD estimate is a particular attribute of the two MMSE-based approaches [10] and [7], which also make use of the *output smoothing* technique. Although [10] and [7] use the same estimation rule, they embed it in the estimation procedure in different ways. While [10] named further as MMSE-VAD applies the MMSE estimator only for time-frequency bins without speech activity (as a VAD-like estimation), [7] called MMSE-BM implements a *bias compensation* and a *minimum search* techniques. The last technique serves in [7] to realize a so called safety-net method for overcoming a complete locking of the algorithm. Note, that we neglected a *bias compensation* of the MMSE-VAD approach as suggested by the author in [10].

Table 2 gives a summarizing overview over the various techniques used in the noise PSD trackers considered here. Note, that all noise PSD trackers mentioned above are causal and none of them needs any training phase.

# 5 Experimental evaluation

To evaluate the performance of the noise PSD trackers, we carried out a single-channel speech enhancement task on the development dataset of the third computational hearing in multisource environments (CHiME) challenge [21], where signals are sampled at 16 kHz. The simulated isolated data of the development dataset consist of 410 utterances in every of 4 different noise environments (on the bus, in a cafe, in a pedestrian area and on a street junction) containing around 2.88 hours of speech data overall. Note, that we used recordings of the 5$^{\text{th}}$ tablet microphone. The input global SNR of this data varies from $-3$ dB up to 33 dB, with an average of about 6 dB. For signal processing we transformed the data using a STFT size of 1024 with a shift of 256 and a Blackman window.

The proposed DNN for causal NPP estimation is trained on the training set of the third CHiME challenge [21]. It is well known that DNNs perform the better, the more data is available during training. We therefore used all six available channels during the training phase of the network. This also allows us to work with a mini-batch size of six without any need for masking or zero-padding. We employ ADAM [22] with a fixed $\alpha = 0.001$ and full backpropagation through time [23]. Additionally, if the norm of a gradient for this network was greater than one, we divided the gradient by its norm [24]. To achieve a better generalization, we used dropout for the input-to-hidden connection of the LSTM units [25] and for the input of the ELU layers [26], see Table 1. We never used dropout for the last layer. 8 epochs were sufficient to train the network.

To ensure the evaluation of the considered noise PSD trackers under the same conditions we assume for all approaches, that the first five frames in the beginning of every utterance are noise-only. The source code of the following noise PSD trackers was either provided by the original authors or taken from publicly available sources: MS [2], MCRA [3], IMCRA [4], MMSE-BM [7], BSMS [8] and SPP-FP [11]. The other noise trackers were implemented according to their published description.

Since the true noise PSD is not known, a noise periodogram $|D(k,\ell)|^2$ smoothed via recursive averaging with a constant smoothing factor $\alpha_{\text{ref}} \in (0;1)$ is often used as a reference noise PSD for the performance evaluation of the noise PSD estimators [27, 28]. The main disadvantage of this technique is the dependence of the optimal parameters of the noise PSD trackers on the choice of $\alpha_{\text{ref}}$. Observing that the knowledge of the true noise periodogram $|D(k,\ell)|^2$ delivers the best performance in spectral speech enhancement compared to use of a smoothed noise periodogram for different values of $\alpha_{\text{ref}}$, we suggest to choose $|D(k,\ell)|^2$ without any smoothing as the noise reference PSD similar to [11]. For performance evaluation of the noise PSD tracking we used the log-error mean (LEM) and a log-error variance (LEV) measures, which are defined in [29] and correspond to the noise PSD estimation error and the variance of the estimator, respectively.

To evaluate the impact of the noise PSD estimators on speech enhancement, we integrated the noise trackers in a single-channel speech enhancement system depicted in Fig. 1. Using a noise PSD estimate $\hat{\lambda}_D(k,\ell)$ an *a posteriori* SNR estimate is calculated

$$\hat{\gamma}(k,\ell) = \frac{|Y(k,\ell)|^2}{\hat{\lambda}_D(k,\ell)}, \tag{5}$$

which is used in the decision directed (DD) approach for the *a priori* SNR estimation [30]. For the DD approach we used a weighting factor 0.98, a minimum value of the *a priori* SNR of $-18$ dB and a real-valued log-spectral amplitude (LSA) gain function $G_{\text{LSA}}(k,\ell)$ [31, 32]. STFT coefficients of an enhanced signal $\hat{S}(k,\ell)$ are calculated by applying a gain function

$$G(k,\ell) = \max(G_{\text{LSA}}(k,\ell), G_{\min}) \tag{6}$$

with a gain floor $G_{\min} = -18$ dB to the noisy STFT coefficients $Y(k,\ell)$ [33]. As performance measures for speech quality of enhanced signals and noise reduction we chose the mean opinion score - listening quality objective (MOS-LQO) measure of enhanced signals [34] and the global output SNR denoted by $\text{SNR}_{\text{out}}$, respectively.

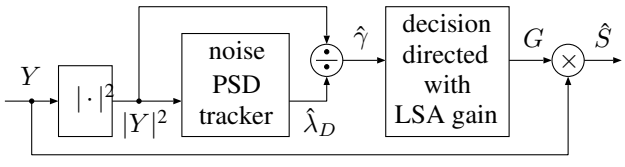


**Figure 1:** Single channel speech enhancement system.

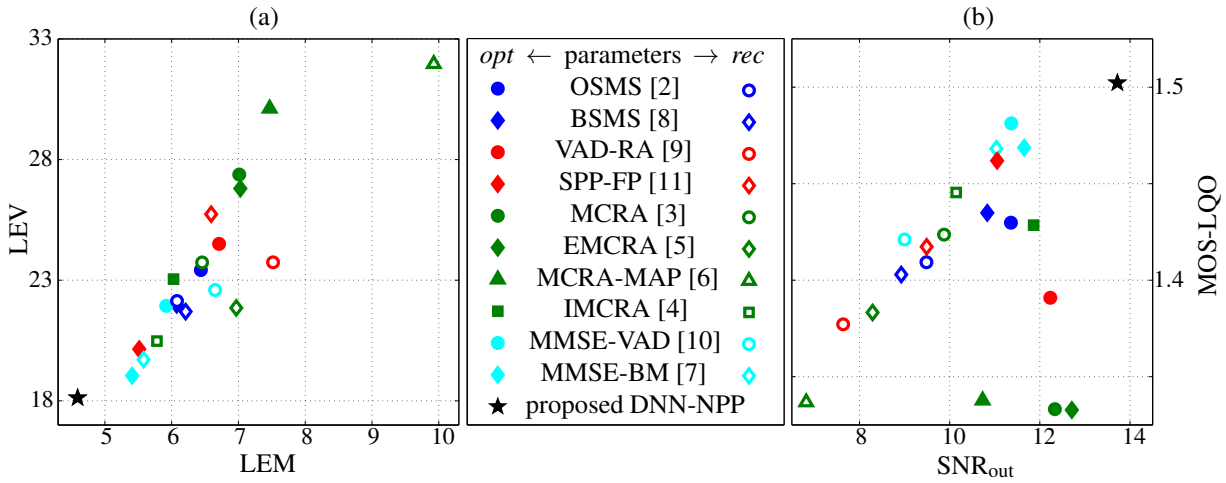


**Figure 2:** Experimental evaluation of the proposed DNN-NPP approach compared to the state-of-the-art noise trackers for the recommended (*rec*) and the optimized (*opt*) parameter sets: (a) noise PSD tracking performance in terms of LEM and LEV measures, (b) impact on the resulting speech enhancement in terms of the $\text{SNR}_{\text{out}}$ values and MOS-LQO scores.

Our experiments showed, that using the parameters of the considered noise trackers recommended by the authors did not lead to the best performance in terms of the used performance measures. Therefore we carried out a parameter optimization via a traditional grid search method on 25 % of the development set containing all 4 noise environments. As a performance metric for the parameter optimization we applied an average over all used performance measures scaled on the range $[0;1]$ on manually specified subset of parameters to be optimized.

The parameter optimization improved the noise tracker performance especially in terms of LEM and $\text{SNR}_{\text{out}}$ measures by 9.7 % and 23.5 %, respectively. A noteworthy outcome of our parameter optimization is the choice of the length of the window for the *minimum search*, which was set to 16 frames, corresponding to the time window of ca. 0.25 s. This value is relatively small compared to the window length in the range $[0.6\,\text{s}; 1.1\,\text{s}]$ recommended in the literature [2, 20]. Note, that a significant $\text{SNR}_{\text{out}}$ improvement of MCRA and EMCRA approaches achieved by optimization occurred on cost of speech quality loss of enhanced signals. These results confirm a trade-off between speech quality and noise suppression [35].

The remaining 75 % of the development set was used for the evaluation of the proposed approach denoted by DNN-NPP compared to the approaches from Table 2. The resulting performance measures, averaged over all utterances and noise environments, are depicted in the Fig. 2. Since our parameter optimization did not lead to a joint improvement in all performance measures, we decided to publish the resulting metrics for both the parameters recommended by their authors and the optimized parameters denoted as *rec* and *opt*, respectively.

It came as a surprise to us to see by how much the proposed DNN-NPP approach outperformed all state-of-the-art noise PSD trackers in all considered performance measures. Our evaluation results of the noise PSD tracking depicted in the Fig. 2(a) show that the noise trackers achieve quite different performance. Among the state-of-the-art approaches the best performance is achieved by the MMSE-BM and SPP-FP approaches. Compared to these two methods the proposed DNN-NPP noise tracker reduces strongly the LEM and slightly the LEV metrics by approximately 1 and 1.5 points, respectively. Furthermore the improved noise tracking of the proposed approach has a striking positive impact on the quality of the enhanced speech signals, as pictured in the Fig. 2(b). Among the state-of-the-art approaches the MMSE-VAD, MMSE-BM and SPP-FP noise trackers deliver the best signal quality. While the EMCRA, MCRA and VAD-RA approaches are particularly well at noise reduction, their estimates cause a poor quality of the enhanced signals. Due to a robust NPP estimation delivered by DNN, the proposed DNN-NPP method leads to the enhanced signals with the best noise reduction and the best signal quality among the state-of-the-art approaches. While the average improvement achieved by the proposed approach compared to the best state-of-the-art approaches in terms of $\text{SNR}_{\text{out}}$ comes to a significant value of 1.3 dB, the average improvement in MOS-LQO reaches small but consistent 0.03 score points.

# 6 Conclusions

In this paper we have presented a causal noise PSD tracking algorithm which employs a DNN-based noise presence probability estimation. The proposed system is a hybrid system, consisting of a DNN-based noise PSD tracker and a conventional speech spectral enhancement system. In an extensive experimental evaluation we observed that the proposed noise tracker outperforms the ten state-of-the-art noise tracking algorithms taken for comparison w.r.t. both the measures of noise PSD tracking performance and the measures of a speech enhancement system using the noise PSD tracker as one component. While the DNN-based noise tracker is computationally more demanding than the other approaches, it can be used in low-latency real-time applications and can cope with nonstationary noise. In future work, more components of the speech enhancement system will be replaced by neural processing.

# 7 Acknowledgements

The work was in part supported by Deutsche Forschungsgemeinschaft under contract no. Ha3455/11-1.

We would like to thank the developers of Chainer [36] for their neural network toolkit.

# References


[1] I. Cohen and S. Gannot, "Spectral Enhancement Methods," in *Springer Handbook of Speech Processing* (J. Benesty, M. M. Sondhi, and Y. A. Huang, eds.), pp. 873–902, Springer Berlin Heidelberg, 2008.

[2] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE Trans. on Speech and Audio Processing (SAP)*, vol. 9, pp. 504–512, July 2001.

[3] I. Cohen and B. Berdugo, "Noise Estimation by Minima Controlled Recursive Averaging for Robust Speech Enhancement," *IEEE Signal Processing Letters (SPL)*, vol. 9, pp. 12–15, Jan. 2002.

[4] I. Cohen, "Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging," *IEEE Trans. on SAP*, vol. 11, pp. 466–475, Sept. 2003.

[5] N. Fan, J. Rosca, and R. Balan, "Speech Noise Estimation using Enhanced Minima Controlled Recursive Averaging," *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. IV–581–IV–584, June 2007.

[6] J. M. Kum, Y. S. Park, and J. H. Chang, "Speech enhancement based on minima controlled recursive averaging incorporating conditional maximum a posteriori criterion," *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4417–4420, Apr. 2009.

[7] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4266–4269, Mar. 2010.

[8] A. Chinaev and R. Haeb-Umbach, "On Optimal Smoothing in Minimum Statistics Based Noise Tracking," in *Sixteenth Annual INTERSPEECH Conference of the International Speech Communication Association (ISCA)*, pp. 1785–1789, Sept. 2015.

[9] H. Hirsch and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," *In Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. 153–156, May 1995.

[10] R. Yu, "A low-complexity noise estimation algorithm based on smoothing of noise power estimation and estimation bias correction," *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4421–4424, Apr. 2009.

[11] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-Based Noise Power Estimation With Low Complexity and Low Tracking Delay," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, pp. 1383–1393, May 2012.

[12] A. Hussain, M. Chetouani, S. Squartini, A. Bastari, and F. Piazza, *Progress in Nonlinear Speech Processing*, ch. Nonlinear Speech Enhancement: An Overview, pp. 217–248. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007.

[13] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, "An Experimental Study on Speech Enhancement Based on Deep Neural Networks," *IEEE SPL*, vol. 21, pp. 65–68, Jan. 2014.

[14] X. L. Zhang and J. Wu, "Denoising deep neural networks based voice activity detection," in *2013 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 853–857, May 2013.

[15] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural Network Based Spectral Mask Estimation for Acoustic Beamforming," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016.

[16] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv e-prints*, 2015.

[17] M. Krawczyk-Becker, D. Fischer, and T. Gerkmann, "Utilizing spectro-temporal correlations for an improved speech presence probability based noise power estimation," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 365–369, Apr. 2015.

[18] D. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)," *CoRR*, vol. abs/1511.07289, 2015.

[19] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *In Proceedings of the Int. Conf. on Artificial Intelligence and Statistics (AISTATS)*, May 2010.

[20] R. Martin, "Spectral Subtraction Based on Minimum Statistics," *In Proc. of the European Signal Processing Conference (EUSIPCO)*, pp. 1182–1185, Sept. 1994.

[21] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third "CHiME" speech separation and recognition challenge: Dataset, task and baselines," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 504–511, Dec. 2015.

[22] D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *ArXiv e-prints*, Dec. 2014.

[23] P. J. Werbos, "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, vol. 78, pp. 1550–1560, Oct. 1990.

[24] R. Pascanu, T. Mikolov, and Y. Bengio, "Understanding the exploding gradient problem," *Computing Research Repository (CoRR)*, vol. abs/1211.5063, Nov. 2012.

[25] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent Neural Network Regularization," *Computing Research Repository (CoRR)*, vol. abs/1409.2329, Sept. 2014.

[26] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," vol. 15, pp. 1929–1958, June 2014.

[27] A. Chinaev, A. Krueger, D. H. Tran-Vu, and R. Haeb-Umbach, "Improved noise power spectral density tracking by a MAP-based postprocessor," in *37th Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4041–4044, Mar. 2012.

[28] A. Chinaev, R. Haeb-Umbach, J. Taghia, and R. Martin, "Improved single-channel nonstationary noise tracking by an optimized MAP-based postprocessor," in *38th Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7477–7481, May 2013.

[29] 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 *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4640–4643, May 2011.

[30] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-32, pp. 1109–1121, Dec. 1984.

[31] O. Cappe, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. on SAP*, vol. 2, pp. 345–349, Apr. 1994.

[32] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 33, pp. 443–445, Apr 1985.

[33] J. Yang, "Frequency domain noise suppression approaches in mobile telephone systems," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, pp. 363–366 vol.2, Apr. 1993.

[34] "Application guide for objective quality measurement based on Recommendations P.862, P.862.1 and P.862.2." ITU-T Recommendation P.862.3, Nov. 2007.

[35] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 28, pp. 137–145, Apr. 1980.

[36] S. Tokui, K. Oono, S. Hido, and J. Clayton, "Chainer: a Next-Generation Open Source Framework for Deep Learning," in *Proc. of Workshop on Machine Learning Systems (LearningSys) in The Twenty-ninth Annual Conf. on Neural Information Processing Systems (NIPS)*, Dec. 2015.