# Robust DNN-Based Speech Enhancement with Limited Training Data

*Robert Rehr and Timo Gerkmann*

Signal Processing (SP), Department of Informatics, Universität Hamburg, Germany
Email: {robert.rehr,timo.gerkmannn}@uni-hamburg.de

## Abstract

In conventional speech enhancement, statistical models for speech and noise are used to derive clean speech estimators. The parameters of the models are estimated blindly from the noisy observation using carefully designed algorithms. These algorithms generalize well to unseen acoustic conditions, but are unable to reduce highly non-stationary noise types. This shortcoming motivated the usage of machine-learning-based (ML-based) algorithms, in particular deep neural networks (DNNs). But if only limited training data are available, the noise reduction performance in unseen acoustic conditions suffers. In this paper, motivated by conventional speech enhancement, we propose to use the *a priori* and *a posteriori* signal-to-noise ratios (SNRs) for DNN-based speech enhancement systems. Instrumental measures show that the proposed features increase the robustness in unknown noise types even if only limited training data are available.

## 1 Introduction

Speech plays a central role in the applications of many personal electronic devices, e.g., in hearing aids, mobile phones and voice-controlled personal assistants. In noisy environments, the speech signal captured by the device's microphones may be corrupted by undesired background noise. Noise degrades the quality and potentially also the intelligibility of speech. Further, noise deteriorates the performance of automatic speech recognition systems. To satisfy the demand for high quality speech communication, enhancement algorithms are utilized to reduce the detrimental effects of noise. In this paper, single-channel speech enhancement algorithms are considered. Such algorithms can be used to enhance noisy speech signals captured by a single microphone and can also be used to improve the output of spatial filtering approaches.

Single-channel speech enhancement has been a research topic for several decades [1]–[5]. Many algorithms leverage the short-time Fourier transform (STFT) where the time-frequency coefficients that are dominated by noise are attenuated. Conventional approaches assume the complex coefficients of speech and noise to follow a known distribution which is used to analytically derive statistically optimal estimators [1], [6], [7]. Such estimators depend on the parameters of the employed distributions which include the speech power spectral density (PSD) and the noise PSD. The PSDs are estimated blindly from the noisy observation using specifically designed algorithms [1], [8]–[10]. In this paper, we refer to these conventional enhancement algorithms as non-machine-learning-based (non-ML-based) enhancement schemes.

The shortcomings of non-ML-based algorithms, namely the inability to suppress highly non-stationary noise types, such as transients, and the speech distortions caused by these algorithms, have motivated the usage of machine-learning-based (ML-based) methods. Instead of estimating properties of speech and noise blindly from the noisy observations, machine-learning (ML) algorithms leverage training examples to learn these properties prior to the processing. For this, various ML algorithms have been employed, e.g., Gaussian mixture models (GMMs) and hidden Markov models (HMMs) [3], non-negative matrix factorization (NMF) [4] and deep neural networks (DNNs) [5], [11]. Especially deep learning techniques show potential to improve speech enhancement in highly non-stationary noises. However, the robustness in unseen acoustic conditions is still discussed [12]–[14].

The generalization of DNN-based algorithms improves generally with increasing number and diversity of the training examples. However, for specific acoustic conditions, only limited training data may be available or obtaining additional training data may be expensive, e.g., in robotics. In this paper, we propose a novel method to improve the generalization of DNN-based speech enhancement algorithms for unseen acoustic conditions if only limited training data are available. The proposed approach combines ML-based methods with non-ML-based noise and speech PSD estimators. Despite the shortcomings of non-ML approaches in highly non-stationary noise, non-ML-based algorithms have been proven to be robust against many different acoustic environments. Further, these algorithms are invariant to changes of the input level. Hence, we propose to use estimates of the *a priori* signal-to-noise ratio (SNR), i.e., the ratio of speech and noise PSD, and the *a posteriori* SNR, i.e., the ratio of noisy input periodogram and noise PSD, as input features. These features are motivated by conventional non-ML clean speech estimators, which are often functions of these two quantities. Further, the features have been previously used to train data-driven gain functions [15], [16], but in contrast to recent DNN-based approaches neighbouring frequency bands have been assumed to be independent and no context has been considered. In contrast to the previously proposed noise aware training (NAT) [5], [17] and its dynamic variants [18], [19], where the estimated noise PSD is appended to the input features, the proposed features are normalized by the noise PSD estimate. This is somewhat related to [20], where an ideal ratio mask (IRM) [21] is predicted by a DNN, which is used as an input for a following enhancement network. In our work, we exploit the generalization of non-ML-based approaches such that training another DNN to predict the IRM is avoided. A similar enhancement structure has been used in [22], but the noise and the speech PSD have been estimated using ML-based algorithms. We compare the proposed features to NAT-based features using instrumental measures. In case of limited training data, Perceptual Evaluation of Speech Quality (PESQ) [23] indicates that the signal quality of the enhanced signals is higher for the proposed features than for NAT-based features in unseen noise conditions.

In Section 2, we recapitulate non-ML-based speech enhancement. Section 3 introduces the ML-based enhancement scheme, recapitulates the previously used noise-aware features and presents the proposed features. The evaluation and the results are shown in Section 4.

## 2 Conventional Speech Enhancement

In this section, conventional non-ML-based speech enhancement is considered and a brief overview of the used speech

ISBN 978-3-8007-4767-2

and noise PSD estimators is given.

The non-ML-based enhancement scheme used in this paper leverages the STFT representation of the noisy input signal for the enhancement. This representation is obtained by splitting the input signal into overlapping segments, which are transformed to the Fourier domain after a spectral analysis window has been applied. This yields the complex spectra of the noisy segments $Y_{k,\ell}$, where $k$ is the frequency index and $\ell$ is the segment index. The interaction between speech and noise is modeled by an additive relationship, which reflects the physical properties of sound. Correspondingly, $Y_{k,\ell}$ can be written as

$$Y_{k,\ell} = S_{k,\ell} + N_{k,\ell}, \tag{1}$$

where $S_{k,\ell}$ and $N_{k,\ell}$ are the speech and noise spectra, respectively.

For many non-ML-based speech enhancement algorithms, the estimation of the clean speech coefficients can be expressed as a multiplication of the noisy input $Y_{k,\ell}$ and a gain function $G_{k,\ell}$, i.e.,

$$\hat{S}_{k,\ell} = \max(G_{k,\ell}, G_{\min}) Y_{k,\ell}. \tag{2}$$

Here, $G_{\min}$ is a lower limit of the gain function which is often used to reduce artifacts and fluctuations in the enhanced signal [24]. In this work, the Wiener filter is used. Its gain function is given by

$$G_{k,\ell} = \frac{\Lambda^{\mathrm{s}}_{k,\ell}}{\Lambda^{\mathrm{s}}_{k,\ell} + \Lambda^{\mathrm{n}}_{k,\ell}} = \frac{\xi_{k,\ell}}{1+\xi_{k,\ell}}, \tag{3}$$

where $\Lambda^{\mathrm{s}}_{k,\ell}$ and $\Lambda^{\mathrm{n}}_{k,\ell}$ are the speech and the noise PSD, respectively. Further, $\xi_{k,\ell}$ denotes the *a priori* SNR which is given by the ratio $\xi_{k,\ell} = \Lambda^{\mathrm{s}}_{k,\ell} / \Lambda^{\mathrm{n}}_{k,\ell}$. The time-domain representation of the enhanced speech is obtained by transforming $\hat{S}_{k,\ell}$ back to the domain. The resulting segments of the estimated clean speech signal are merged using an overlap-add procedure after applying a synthesis window.

The noise PSD is estimated using the algorithm described in [10], [25]. The method allows to track background noises that change moderately fast, e.g., passing cars on a busy street. Quickly changing background noises such as transient sounds of cutlery or in factories cannot be tracked. The *a priori* SNR $\xi_{k,\ell}$ is estimated using the cepstral smoothing approach presented in [9], [26]. This approach causes less undesired artifacts in the enhanced signal than the commonly used decision-directed approach [1].

# 3 ML-Based Speech Enhancement

In this section, the DNN-based enhancement scheme used in this paper is described. In the second part, previously proposed input features are described that do not use a noise PSD estimate for normalization. After that, the proposed normalized input features are presented.

## 3.1 Algorithm

Similar to the non-ML enhancement algorithms, the ML-based approach also utilizes the STFT for the enhancement. Instead of using the Wiener filter, a feed-forward DNN is used to map features extracted from the noisy input spectra to an IRM. The IRM has been proposed in [21] and is defined as

$$\mathrm{IRM}(k,\ell) = \frac{|S_{k,\ell}|^2}{|S_{k,\ell}|^2 + |N_{k,\ell}|^2}. \tag{4}$$

The IRM has similarities to the Wiener filter, but uses the speech periodogram $|S_{k,\ell}|^2$ and the noise periodogram $|N_{k,\ell}|^2$ instead of the respective PSDs.

During processing, the predicted ratio mask $\widehat{\mathrm{IRM}}(k,\ell)$ obtained from a trained DNN is used as a multiplicative factor to estimate the clean speech coefficients. Correspondingly, the estimated clean speech spectrum for the ML-based approach is given by

$$\hat{S}_{k,\ell} = \max(\widehat{\mathrm{IRM}}(k,\ell), G_{\min}) Y_{k,\ell}. \tag{5}$$

Similar to (2), a lower limit is introduced again and time-domain signal is obtained using an overlap-add procedure.

## 3.2 Non-Normalized Input Features

Various non-noise-aware features have been considered for DNN-based speech enhancement, e.g., Gammatone filterbank features, log-mel spectra, mel-frequency cepstral coefficients or amplitude modulation spectra [27]. Such features are generally directly based on the noisy input spectrum $Y_{k,\ell}$. In this paper, we include the logarithmized noisy spectra, i.e., $y^{(\log)}_{k,\ell} = \log(|Y_{k,\ell}|^2)$ as example of non-noise-aware input features, which have also been used in [5]. Here, $\log(\cdot)$ denotes the natural logarithm.

To improve the robustness of DNN-based speech enhancement, NAT-based features have been proposed. NAT has been initially used in [17] to improve the robustness of automatic speech recognition algorithms in unseen noise conditions. For this, the noisy input features have been augmented by a static estimate of the noise PSD. This estimate has been obtained by averaging the noisy input periodogram over the first segments. The idea has been adapted for speech enhancement in [18]–[20] where the static noise PSD has been replaced by a dynamic estimate. In [18], [19], conventional non-ML noise PSD estimators have been considered whereas ML-based noise PSD estimation has been used in [18], [20]. Here, we focus on the former and employ the concatenation of noisy log-spectra $y^{(\log)}_{k,\ell}$ and the logarithmized noise PSD $\Lambda^{\mathrm{n},(\log)}_{k,\ell} = \log(\Lambda^{\mathrm{n}}_{k,\ell})$ as the input feature vector. The noise PSD is estimated using the speech presence probability based noise PSD estimator presented in [10], [25].

## 3.3 Proposed Input Features

In case of limited training data, the robustness of DNN-based enhancement schemes against unseen acoustic conditions can be improved. For this, we propose to use normalized features. Here, in contrast to NAT and its dynamic variants, the noise PSD is not appended to the noisy input features but used for normalization. More specifically, we propose to use the logarithmized *a priori* SNR $\xi^{(\log)}_{k,\ell} = \log(\xi_{k,\ell})$ and the logarithmized *a posteriori* SNR $\gamma^{(\log)}_{k,\ell} = \log(\gamma_{k,\ell})$, where $\gamma_{k,\ell}$ is defined as

$$\gamma_{k,\ell} = |Y_{k,\ell}|^2 / \Lambda^{\mathrm{n}}_{k,\ell}. \tag{6}$$

The *a priori* SNR and *a posteriori* SNR can be used separately or can be concatenated such that both SNRs are used as inputs. The speech and the noise PSD are estimated using the non-ML approaches used for the enhancement approach in Section 2.

The proposed normalized features are inspired by conventional non-ML-based clean speech estimators, e.g., [1], [6], [7] and have been previously used to estimate data-driven gain functions [15], [16]. However, as contextual information can be easily included during training, the DNN can potentially

exploit correlations along time and frequency, which cannot be easily included in conventional non-ML estimators. Such dependencies have also not been included in [15], [16]. Thus, if carefully trained, the DNN-based approach has the potential to yield more powerful estimators as the correlations over time and frequency are included. Further, similar to NAT-based features, an estimate of the noise PSD is included to allow for a robust enhancement in unseen acoustic conditions. But due to the normalization, the proposed feature are scale-invariant, i.e., they do not depend on the overall level of the input signal.

# 4 Evaluation

In this section, the non-ML algorithm from Section 2 and the ML-based algorithm from Section 3 are evaluated. First, we explain the experimental setup and describe the used audio material, the parameters of the algorithms and the training of the ML-based approaches. After this, we show results where the quality of the speech signals is predicted by PESQ [23]. We demonstrate that the proposed features are invariant to level changes of the input features, whereas the non-normalized features are not, and compare the performance of the algorithms.

## 4.1 Audio Material, Parameters and Training

In our experiments, the sampling rate of the signals is set to 16 kHz. For the STFT, a segment length of 32 ms and an overlap of 50 % is used. The analysis and synthesis window is given by a square-root Hann window. For all features, contextual information is included. For this, the features extracted from three previous frames are appended to the feature vector of the current frame. We do not use features from future frames to allow the DNN-based algorithms to yield the same latency as the non-ML estimator. For the employed parameters of the STFT, this results in a feature dimensionality of $(3+1)\cdot 257 = 1028$, where 257 coefficients result from removing the mirror spectrum. For concatenated features, i.e., the NAT-based features and the combination of *a priori* and *a posteriori* SNR, this dimensionality doubles to 2056.

For the prediction of the IRM, a feed-forward DNN with three hidden layers is used. Each hidden layer comprises 1024 rectifying linear units (ReLUs) [28]. The non-linearities in the output layer are sigmoid non-linearities to match the limited range of the IRM. The size of the output layer is given by the STFT parameters which results in 257 units. The parameters of the DNN are optimized by minimizing the squared error between the predicted IRM and the true IRM as

$$J = \sum_k \sum_\ell \left| \widehat{\mathrm{IRM}}(k,\ell) - \mathrm{IRM}(k,\ell) \right|^2 . \tag{7}$$

The weights and biases are initialized using the method proposed in [29]. After that the weights and biases are optimized by stochastic gradient descent where the learning rate is reduced from 0.4 to 0.1 using $\mathrm{LR} = \max(0.4 \cdot 0.95^{E-1}, 0.1)$. Here, LR is the learning rate and $E$ is the number of the current epoch. All models are trained for 100 epochs using a batch size of 128 samples. For each epoch the error is computed on a validation set, which is constructed by selecting 15 % of the training data. In the experiments, the model with the lowest error on the validation set is employed.

In this work, we use a set of nine different noise types. It contains the factory 1 and the babble noise taken from the NOISEX-92 database [30]. From the same database, an amplitude modulated version of the pink and white noise are included. Further background noises have been obtained from the freesound database. Here, the sounds of an overpassing propeller plane (https://freesound.org/s/115387/), the interior of a passenger jet during flight (https://freesound.org/s/188810/), a vacuum cleaner (https://freesound.org/s/67421/) and a traffic noise (https://freesound.org/s/75375/) are employed. Further, a two-talker babble noise is used which is generated by mixing two read out stories taken from https://www.vorleser.net. The stories are read by a male and a female speaker, respectively, and are mixed such that the SNR between the two signals is 0 dB. For generating the noise file, the speech pauses have been removed. The noise types are used to conduct cross-validation experiments where all noise types except one are included in the training set. The training data of each cross-validation set are augmented by additionally including a highly non-stationary noise type which is generated from the noise snippets collected by [31]. The noise excerpts in this database are generally short and are, hence, concatenated multiple times in various orders to give a continuous noise signal. Long noise excerpts are split into roughly 2 second long snippets during this generation. This noise type is referred to as concatenated short noise excerpts (CSNE). The remaining unseen noise type is used for testing in the evaluations.

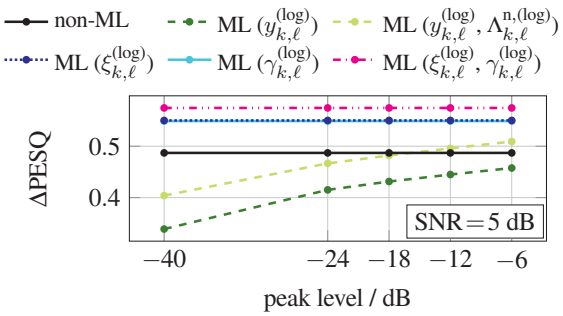


**Figure 1:** PESQ improvements of ML-based speech enhancement algorithms using different input features in comparison to non-ML-based enhancement schemes for different peak levels of the speech input.

For training of the ML-based enhancement schemes, 3992 sentences of the TIMIT training set [32] are used. It is ensured that the number of sentences spoken by male and female speakers is the same. These sentences are artificially corrupted by the background noises above, where each sentence is embedded once in the respective noise types used for training. This results in about 20 hours of training material for each cross-validation set. The position of each sentence is chosen randomly and to allow the non-ML noise PSD estimator to adapt to the background noise, each sentence is prepended by a two second noise only segment. This segment is removed from the data finally used for training. A portion of 10 % of training data contains only noise to allow the DNN to learn how to remove such parts. To allow the ML-based enhancement schemes to see the effects of different input SNRs and changes in the overall level of the input signals, such variations are included in the training data. The peak level of each training sentence is varied between $-26$ dB and $-3$ dB. Similarly, also the SNR of each sentence is varied between $-10$ dB and 15 dB.

## 4.2 Results

For the instrumental evaluation, 64 sentences spoken by male and 64 sentences spoken by female speakers from the TIMIT

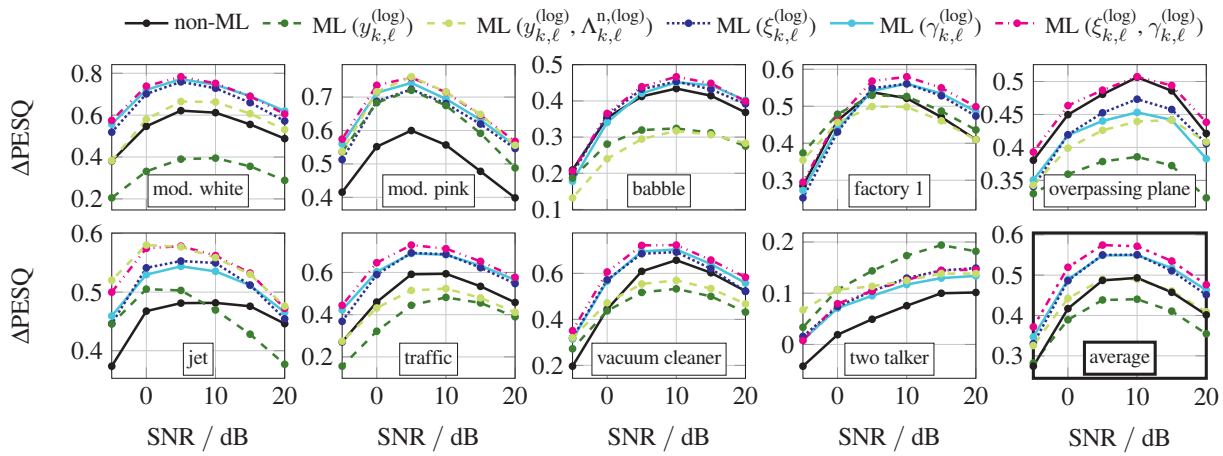


**Figure 2:** PESQ improvements of ML-based speech enhancement algorithms using different input features in comparison to non-ML enhancement schemes.

testing set are used. Again, the sentences are embedded at random positions in the background noise. Also for the testing case, a two-second initialization period is used for the non-ML noise PSD estimator to avoid initialization artifacts. The initialization period is excluded from the evaluation of the instrumental measures. The minimum gain $G_{\min}$ is set to $-20$ dB in all evaluations.

First, we show that the considered non-normalized features are scale dependent. For this, we vary the peak level of the speech signal and set it to $-40$ dB, $-24$ dB, $-18$ dB, $-12$ dB and $-6$ dB. The SNR of the input signal is set to 5 dB. For each noise type the DNN is trained on the remaining noise types of the employed noise pool described in Section 4.1, i.e., all noise types are unseen. Figure 1 shows the PESQ results averaged over all noise types of the employed noise pool except CSNE. The results show that the performance of the DNN-based enhancement schemes varies with the level of the input signal when non-normalized features are used. Figure 1 shows that the performance decreases with decreasing level. The non-ML approach and the DNN-based enhancement schemes based on the proposed normalized features do not exhibit this behavior because of the scale-invariant normalized input features. Further, this result gives a preview on the performance of the different features. In general, the proposed features yield the highest scores and outperform the non-normalized features. As more accurate statistics and correlations along time and frequency can be included, the DNN-based approach outperforms the conventional non-ML enhancement.

Furthermore, we evaluated PESQ depending on the noise type and the input SNR which is varied between $-5$ dB and 20 dB in 5 dB steps. In this experiment, the peak level of clean speech sentence is varied between $-26$ dB and $-3$ dB. Figure 2 shows the results for all noise types used in the cross-validation except the CSNE. Again, the evaluated noise type is always unseen. For a low amount of training data as used in this experiment, the proposed normalized features generally yield higher scores than the non-normalized features. An exception is the two talker background where, however, all approaches yield relatively low scores compared to other noise types. Further, NAT, i.e., the logarithmized periodogram $y_{k,\ell}^{(\log)}$ combined with the noise PSD $\Lambda_{k,\ell}^{\mathrm{n},(\log)}$, yields similar or higher scores than using only the logarithmized noisy spectra $y_{k,\ell}^{(\log)}$. Comparing NAT to the *a priori* SNR $\xi_{k,\ell}^{(\log)}$ or *a posteriori* SNR $\gamma_{k,\ell}^{(\log)}$ for the overpassing plane or the jet noise shows that the features have a similar performance. For most of the remaining noise types, e.g., babble noise, the quality predicted by PESQ is higher for the *a posteriori* and *a priori* SNR. If only one of the SNRs is used as feature, the input layer is only half as wide as for NAT. As a consequence, the computational complexity resulting network is lower such that these features can be considered an attractive alternative to NAT. The combination of both SNRs, i.e., the *a posteriori* SNR and the *a priori* SNR, generally yields the highest scores.

An interesting example which shows the advantages of the proposed features is the amplitude modulated white noise. In this case, the remaining noise types in the training data do not reflect the spectral envelope of white noise. Analyzing the signals processed using this model shows that the noise is hardly reduced which explains the lower PESQ scores. If the normalized features are employed, the ML-based speech enhancement do not suffer from this problem such that the unknown noise can be removed. Consequently, the performance remains high which demonstrates the benefit of the proposed normalized features.

## 5 Conclusions

This paper addresses the generalization of DNN-based speech enhancement schemes when only limited training data is available. For this case, we propose to combine ML-based and conventional non-ML-based approaches. More specifically, we propose to use the *a priori* and *a posteriori* SNR as input features. The features are motivated by conventional non-ML-based speech enhancement algorithms. In contrast to the previously proposed NAT, where the noise PSD is appended to the feature vector, the estimated noise PSD is used as a normalization term. Using a cross-validation experiment, we show that the proposed features are scale-invariant. Further, the features are less prone to variations of the background noise if only limited training data are available. The SNR-based features generally yield the highest scores from which we conclude that the generalization of ML-based speech enhancement is improved by the proposed features and that the advantages of DNN-based enhancement schemes can be maintained.

# References


[1] 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, Dec. 1984.

[2] Y. Ephraim, "A bayesian estimation approach for speech enhancement using hidden markov models", *IEEE Transactions on Signal Processing*, vol. 40, no. 4, pp. 725–735, Apr. 1992.

[3] D. Y. Zhao and W. B. Kleijn, "HMM-based gain modeling for enhancement of speech in noise", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 882–892, Mar. 2007.

[4] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2140–2151, Oct. 2013.

[5] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, "A regression approach to speech enhancement based on deep neural networks", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, Jan. 2015.

[6] C. Breithaupt, M. Krawczyk, and R. Martin, "Parameterized MMSE spectral magnitude estimation for the enhancement of noisy speech", in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, NV, USA, Apr. 4, 2008, pp. 4037–4040.

[7] R. C. Hendriks, R. Heusdens, and J. Jensen, "Log-spectral magnitude MMSE estimators under super-gaussian densities", in *Interspeech*, Brighton, United Kingdom, 2009, pp. 1319–1322.

[8] 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, Jul. 2001.

[9] C. Breithaupt, T. Gerkmann, and R. Martin, "A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing", in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, NV, USA, Apr. 4, 2008, pp. 4897–4900.

[10] 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, May 2012.

[11] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder", in *Interspeech*, Lyon, France, Aug. 2013.

[12] T. May and T. Gerkmann, "Generalization of supervised learning for binary mask estimation", in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, Juan-les-Pins, France, Sep. 2014, pp. 154–158.

[13] S. E. Chazan, J. Goldberger, and S. Gannot, "A hybrid approach for speech enhancement using MoG model and neural network phoneme classifier", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2516–2530, Dec. 2016.

[14] M. Kolbæk, Z. H. Tan, and J. Jensen, "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 149–163, Jan. 2017.

[15] T. Fingscheidt and S. Suhadi, "Data-driven speech enhancement", in *ITG Conference on Speech Communication*, Kiel, Germany, Apr. 2006.

[16] J. Erkelens, J. Jensen, and R. Heusdens, "A data-driven approach to optimizing spectral speech enhancement methods for various error criteria", *Speech Communication*, Speech Enhancement, vol. 49, no. 7, pp. 530–541, Jul. 1, 2007.

[17] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition", in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, BC, Canada, May 2013, pp. 7398–7402.

[18] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "Dynamic noise aware training for speech enhancement based on deep neural networks", in *Interspeech*, Singapore, Singapore, Sep. 2014.

[19] A. Kumar and D. Florencio, "Speech enhancement in multiple-noise conditions using deep neural networks", in *Interspeech*, San Francisco, CA, USA, Sep. 2016, pp. 3738–3742.

[20] Q. Wang, J. Du, L. R. Dai, and C. H. Lee, "Joint noise and mask aware training for DNN-based speech enhancement with sub-band features", in *Hands-free Speech Communications and Microphone Arrays (HSCMA)*, San Francisco, CA, USA, Mar. 2017, pp. 101–105.

[21] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.

[22] S. Mirsamadi and I. Tashev, "Causal speech enhancement combining data-driven learning and suppression rule estimation", in *Interspeech*, San Francisco, CA, USA, Sep. 2016, pp. 2870–2874.

[23] "P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs", International Telecommunication Union, ITU-T recommendation, Jan. 2001.

[24] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise", in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Washington, D.C., USA, Apr. 1979, pp. 208–211.

[25] T. Gerkmann and R. C. Hendriks, "Noise power estimation based on the probability of speech presence", in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2011, pp. 145–148.

[26] T. Gerkmann and R. Martin, "On the statistics of spectral amplitudes after variance reduction by temporal cepstrum smoothing and cepstral nulling", *IEEE Transactions on Signal Processing*, vol. 57, no. 11, pp. 4165–4174, 2009.

[27] M. Delfarah and D. Wang, "Features for masking-based monaural speech separation in reverberant conditions", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1085–1094, May 2017.

[28] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks", in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, Jun. 14, 2011, pp. 315–323.

[29] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks", in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, Chia Laguna Resort, Sardinia, Italy, May 2010, pp. 249–256.

[30] H. J. M. Steeneken and F. W. M. Geurtsen, "Description of the RSG.10 noise database", TNO Institute for perception, Technical Report IZF 1988-3, 1988.

[31] G. Hu. (2005). A corpus of nonspeech sounds, [Online]. Available: http://web.cse.ohio-state.edu/pnl/corpus/HuNonspeech/HuCorpus.html (visited on 01/09/2018).

[32] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, *TIMIT acoustic-phonetic continuous speech corpus*, 1993.