Improving the Generalizability of Deep Neural Network Based Speech Enhancement

Robert Rehr, *Student Member, IEEE,* Timo Gerkmann, *Senior Member, IEEE*

***Abstract*—Enhancing noisy speech is an important task to restore its quality and to improve its intelligibility. In traditional non-machine-learning (ML) based approaches the parameters required for noise reduction are estimated blindly from the noisy observation while the actual filter functions are derived analytically based on statistical assumptions. Even though such approaches generalize well to many different acoustic conditions, the noise suppression capability in transient noises is low. More recently, ML and especially deep learning has been employed for speech enhancement and studies show promising results in noise types where non-ML based approaches fail. However, due to their data-driven nature, the generalizability of ML based approaches to unknown noise types is still discussed. To improve the generalizability of ML based algorithms and to enhance the noise suppression of non-ML based methods, we propose a combination of both approaches. For this, we employ the *a priori* signal-to-noise ratio (SNR) and the *a posteriori* SNR estimated by non-ML based algorithms as input features in a deep neural network (DNN) based enhancement scheme. We show that this approach allows ML based speech estimators to generalize quickly to unknown noise types even if only few noise conditions have been seen during training. Instrumental measures such as Perceptual Evaluation of Speech Quality (PESQ) and the segmental SNR indicate strong improvements in unseen conditions when using the proposed features. Listening experiments clearly confirm the improved generalization of our proposed combination.**

***Index Terms*—Deep neural networks, machine learning, generalizability, limited training data, speech enhancement.**

I. Introduction

THE most common way of communicating for human beings is to use vocalized speech. The presence of background noise, however, may deteriorate the speech intelligibility [1], [2], as well as the quality of the speech signal. Besides human perception, background noise also affects automatic speech recognition algorithms for human-machine interfaces and results in lower recognition rates [3], [4]. Speech enhancement algorithms therefore play an important role for noise robust speech recognition and for improving the speech quality in hearing aid and telecommunication applications. In this paper, single-channel speech enhancement algorithms are considered that either assume that the noisy signal has been captured by a single microphone or process the output of a beamformer [5].

Single-channel speech enhancement has been a research topic for many decades and has led to many different methods, e.g., [6]–[16]. Here, we distinguish between two broad categories of single-channel speech enhancement schemes, namely machine-learning (ML) based and non-ML based approaches. ML based enhancement schemes generally follow a two-step approach to enhance a noisy speech signal. First, the model parameters of an ML algorithm are tuned on training data. After that, the obtained models are used to separate the speech component from the background noise. On the contrary, non-ML based approaches do not learn any models from training data prior to processing. Instead the parameters required for the enhancement are estimated on-line and blindly from the noisy observation.

Non-ML based enhancement schemes such as [6], [7], [12], [13], [17]–[19], commonly operate in the short-time Fourier transform (STFT) domain where a filter function is applied to suppress the coefficients that mainly contain noise. The employed filter functions are often derived as statistically optimal estimators leading to various solutions depending on the distributional assumptions used to model the speech and noise coefficients [7], [17], [18], [20]–[22]. The resulting estimators are functions of the distributions' parameters such as the speech power spectral density (PSD) and the noise PSD, which are estimated blindly from the noisy observation. Most of the existing noise PSD estimators exploit the assumption that the background noise varies more slowly than the speech signal. Following this assumption, approaches that track the minima of the noisy input periodograms [19] have been proposed to estimate the noise PSD. Other approaches identify time-frequency points having considerably larger magnitudes than the previously estimated noise PSD as speech and exclude them from the noise PSD updates [9], [13]. From the noise PSD estimate and the noisy observation, the speech PSD is obtained, e.g., [7], [12], which is required to design the noise reduction filters [7], [18], [20]–[22]. Such approaches have been proven to generalize well to many different acoustical environments and provide good results in moderately varying noise types. However, they lack the ability to track fast changes of the background noise. Due to their underlying assumption that noise varies more slowly than speech, highly non-stationary noises, e.g., transient noises like the sound of forks and knives in a restaurant situation, or sudden changes of the acoustical environment are often considered as speech by non-ML based approaches. As a consequence, such transient sounds are generally not suppressed by non-ML based enhancement schemes.

The shortcomings of non-ML based enhancement schemes especially with respect to the limited noise tracking capabilities have motivated the usage of ML algorithms for speech enhancement, e.g., [8], [11], [14]–[16], [23]–[31]. In [8], [11], [25] generative models such as hidden Markov models and Gaussian mixture models have been employed to learn typical

R. Rehr and T. Gerkmann are with the Signal Processing Group, Department of Informatics, Universität Hamburg, Germany (e-mail: robert.rehr@uni-hamburg.de; timo.gerkmann@uni-hamburg.de)

Manuscript received September 6, 2017.

shapes of the speech and noise PSDs or related quantities. The shapes that most likely explain the noisy observation are employed to design filter functions, similar to the non-ML approaches. Other ML approaches train two separate dictionaries for speech and noise, e.g., using nonnegative matrix factorization (NMF), to untangle the speech component from the noisy mixture, [14], [23], [27]. Also here, the separated speech and noise components are often used to design a filter function which is applied to the noisy STFT spectra. Recently, deep neural networks (DNNs) are more intensively investigated for speech enhancement applications [15], [16], [26], [28], [29], [31]. In contrast to other learning approaches, DNNs are used to learn a non-linear function, e.g., a mapping from the noisy observation or features extracted from it to a filter or masking function [26], [29], [31]. DNNs can as well be utilized to learn a mapping where the target is directly given by the clean speech coefficients as in [15] such that no filter function is required. The most recent studies on ML based speech enhancement approaches, e.g., [15], [29], [31], show that DNN based approaches in principal have the ability to reduce transient noises. Even though theses types of ML based enhancement schemes show promising results to improve the performance in such noise types, one of the major concerns with ML based approaches is the generalizability to noise types that have not been seen during training. For DNN based approaches, this issue is encountered with large and diverse training data [15], [29]. Here, hundreds or even thousands of different noise types are included during training to enable the DNN based enhancement schemes to generalize to unseen noise conditions. Even though large training data sets increase the generalizability, using thousands of noise types may still be inappropriate as in real-world scenarios virtually infinitely many noise types can possibly occur as argued in [16], [32].

In this paper, we combine the generalization strengths of non-ML based and the improved performance in transient noise of ML based speech enhancement schemes. For this, we propose to employ pre-processed features taken from non-ML single-channel speech enhancement algorithms in a DNN based speech enhancement framework. Specifically, we propose to employ the *a priori* signal-to-noise ratio (SNR), i.e., the ratio of the clean speech PSD and the noise PSD, and the *a posteriori* SNR, i.e., the ratio of the noisy periodogram and the noise PSD, as features. The usage of the *a priori* SNR and the *a posteriori* SNR is motivated by non-linear clean speech estimators, e.g., [7], [18] where these quantities result from the derivation of minimum-mean squared optimal estimators. From this choice we expect the advantages that the additional time-varying information provided by means of the noise and speech PSD estimates improves the generalizability as well as the performance of ML based enhancement schemes. The proposed approach has similarities to the noise-aware training used in [15], [33] where a fixed noise PSD estimate is obtained by averaging the first and last segments of the input features. In [15], [33], this estimate is then appended to the feature vector to allow the DNN to adapt to different noise types. Here, we use time-varying estimates as features for the DNN which provide the ML based approach additional information, e.g., changes in the spectral shape of the background noise. Time-varying noise PSD estimates have also been considered in [32] where the fixed noise estimate appended to the feature vector has been replaced by the noise PSD estimate obtained from [13]. The study in [32], however, mainly considers office environments and focuses on the generalizability of ML based methods to unseen mixtures of sounds taken from the same environment. In our work, there is no restriction to specific acoustic environments. Further, our study focuses on the generalizability of ML based approaches to arbitrary noise types if only a limited amount of noise types is available for training. Our results confirm that the proposed features allow to train DNN based enhancement schemes that generalize to unseen noise types even if only few noise types have been seen during training. This effect is validated using cross-validation experiments where different sets of noise types for training and testing are used. Additionally, we support our findings using subjective evaluations.

First, we describe the employed algorithms in Section II and the used parameters and training methods in Section III. The results of the instrumental measures is shown in Section IV while the listening experiment is described in Section V.

II. Enhancement Algorithms

In this section, an overview of the enhancement algorithms considered in this paper is given. First, the non-ML based speech enhancement scheme based on [12], [13] and the Wiener filter is introduced. After that, an ML based enhancement scheme similar to [15] is presented which serves as the basis for the proposed combination described afterwards.

A. Non-ML Based Enhancement Scheme

In this part, the non-ML based enhancement scheme is introduced. This algorithm operates in the STFT domain which is obtained by splitting the noisy input signal into overlapping segments and taking the Fourier transform of each segment after a tapered spectral analysis window has been applied. We employ the physically plausible assumption that the speech signal and the noise signal mix additively, i.e.,

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

The symbols $S_{k,\ell}$ and $N_{k,\ell}$ denote the complex clean speech spectrum and the complex noise spectrum, respectively, while $Y_{k,\ell}$ is the resulting spectrum of the noisy signal. Furthermore, k is the frequency index and ℓ is the segment index. The speech coefficients are estimated using the Wiener filter gain function $G_{k,\ell}$ as

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

where G_{min} acts as a lower limit on the Wiener filter gain $G_{k,\ell}$. The minimum gain G_{min} is an important parameter to limit artifacts in the enhanced signal such as fluctuations in the reduced background noise or musical tones [34]. The Wiener gain $G_{k,\ell}$ is given by

$$G_{k,\ell} = \frac{\Lambda^s_{k,\ell}}{\Lambda^s_{k,\ell} + \Lambda^n_{k,\ell}} = \frac{\xi_{k,\ell}}{\xi_{k,\ell} + 1}, \tag{3}$$

where $\Lambda^s_{k,\ell}$ and $\Lambda^n_{k,\ell}$ denote the speech PSD and the noise PSD, respectively. Further,

$$\xi_{k,\ell} = \frac{\Lambda^s_{k,\ell}}{\Lambda^n_{k,\ell}} \tag{4}$$

is referred to as *a priori* SNR. The *a priori* SNR $\xi_{k,\ell}$ and the noise PSD $\Lambda^n_{k,\ell}$ are estimated blindly from the noisy observation using [12], [13]. The clean speech estimates $\hat{S}_{k,\ell}$ are transformed back to the time-domain for each segment ℓ. Each enhanced time-domain segment is weighted by a tapered synthesis window and by using an overlap-add method the time-domain signal is reconstructed.

The noise PSD $\Lambda^n_{k,\ell}$ and the *a priori* SNR $\xi_{k,\ell}$ are estimated on-line and blindly from the noisy observation. For this, the speech presence probability based noise PSD estimator proposed in [13] is used in this work. This estimator allows to track moderate changes in the background noise such as passing cars. However, it cannot track transient disturbances. For estimating the *a priori* SNR $\xi_{k,\ell}$, the temporal cepstrum smoothing approach described in [12] is employed. In contrast to the commonly used decision-directed approach [7], this approach causes less musical tones.

B. ML Based Enhancement Scheme

In this section, the employed ML based speech enhancement scheme is described. The architecture resembles approaches that have been proposed in [15], [31].

Similar to the non-ML based enhancement scheme, also the ML based approach operates in the STFT domain. Here, a feed-forward DNN is used to predict an ideal ratio mask (IRM) from input features extracted from the noisy input signal. The IRM has been proposed in [26] and is similar to the Wiener filter gain function shown in (3) with the difference that the speech periodogram $|S_{k,\ell}|^2$ and the noise periodogram $|N_{k,\ell}|^2$ are employed instead of the respective PSDs as

$$\mathrm{IRM}(k,\ell) = \frac{|S_{k,\ell}|^2}{|S_{k,\ell}|^2 + |N_{k,\ell}|^2}. \tag{5}$$

Similar to the Wiener filter, the predicted IRM obtained from the DNN is used to estimate the clean speech coefficients $\hat{S}_{k,\ell}$ as

$$\hat{S}_{k,\ell} = \max\left(\widehat{\mathrm{IRM}}(k,\ell), G_{\min}\right) Y_{k,\ell}, \tag{6}$$

where $\widehat{\mathrm{IRM}}(\cdot)$ denotes the IRM estimated by the DNN. As in (2), also here, we enforce a lower limit $G_{\min}$. As for the non-ML based enhancement scheme, the time-domain signal is reconstructed using an overlap-add method.

The DNN's architecture comprises three hidden layers with rectified linear units (ReLUs) [35] as non-linearities and an output layer with sigmoidal activation functions. The input features of this reference DNN approach are based on the logarithmized noisy periodograms, i.e.,

$$y^{(\log)}_{k,\ell} = \log\left(|Y_{k,\ell}|^2\right). \tag{7}$$

All spectral coefficients of a segment ℓ, i.e., $y^{(\log)}_{k,\ell}$ for all frequency bins k for a given frame ℓ, are stacked in a feature vector. Additionally, a context of three past segments is added to this vector by appending the respective log-spectra to the end of the vector. We do not add context from future segments to keep the algorithmic latency as low as for the non-ML based enhancement scheme.

C. Proposed Combination

In this part, the combination of the non-ML based enhancement scheme in Section II-A and the ML based scheme in Section II-B is described.

The goal of our combination is to exploit the advantages of ML based and non-ML based enhancement strategies. Accordingly, we want to preserve the ability to reduce non-stationary noises of the ML based approaches while the robustness towards unseen noise types of non-ML based approaches is maintained at the same time. For retaining the noise reduction capabilities, the combination also incorporates an ML part where, again, a DNN is employed. Similar to the ML based approach described in Section II-B, the DNN's task is to map features extracted from the noisy input signal to an IRM. For this, we employ the same architecture as in Section II-B. To incorporate the generalizability from non-ML based enhancement schemes, we employ the logarithmized *a priori* SNR $\xi^{(\log)}_{k,\ell} = \log(\xi_{k,\ell})$ based on (4) and the *a posteriori* SNR $\gamma^{(\log)}_{k,\ell} = \log(\gamma_{k,\ell})$ as input features, where the *a posteriori* SNR is defined as

$$\gamma_{k,\ell} = \frac{|Y_{k,\ell}|^2}{\Lambda^n_{k,\ell}}. \tag{8}$$

Again, the *a priori* SNR $\xi_{k,\ell}$ and the noise PSD $\Lambda^n_{k,\ell}$ are estimated blindly from the noisy observation using [12], [13]. For each segment, the features, i.e., $\xi^{(\log)}_{k,\ell}$ and $\gamma^{(\log)}_{k,\ell}$, are stacked in a vector and, again, a temporal context of three previous segments is included.

From these features we expect the advantage that the employed DNN can more easily adapt to unseen noise types as the noise PSD $\Lambda^n_{k,\ell}$ normalizes both, the *a priori* and the *a posteriori* SNR. Further, the features also include an estimate of the speech PSD $\Lambda^s_{k,\ell}$ in $\xi_{k,\ell} = \Lambda^s_{k,\ell}/\Lambda^n_{k,\ell}$ which may additionally be exploited by the DNN in the process of predicting the clean speech coefficients. The employed features are motivated by various non-ML based clean speech estimators described in the literature, e.g., [7], [18], [20]–[22] which are also non-linear functions of the *a priori* SNR $\xi_{k,\ell}$ and the *a posteriori* SNR $\gamma_{k,\ell}$ but result by deriving the analytical expression of the minimum mean-squared error optimal estimator from a given statistical model. Training a DNN on the logarithmized versions of the *a priori* SNR $\xi^{(\log)}_{k,\ell}$ and the *a posteriori* SNR $\gamma^{(\log)}_{k,\ell}$ also results in such a non-linear function. Correspondingly, this allows the DNN to be interpreted as a non-linear clean speech estimator similar to non-ML based enhancement schemes. However, non-ML based clean speech estimators often neglect temporal and spectral correlations. These can be easily included in DNN based approaches by extending the feature vector over multiple frequency bins k and time segments ℓ as done here and in many other approaches, e.g., [15], [29], [31]. Hence, training a DNN on such data may result in a more advanced estimator compared to most non-ML based clean speech estimators.

III. Audio Material and Parameters

For all algorithms, the STFT uses 32 ms segments which overlap by 50 %. For the analysis step as well as the synthesis step a square-root Hann window is employed. The mirror spectrum is omitted in the extracted features. We employ a sampling rate of 16 kHz such that the resulting dimension of the input features for the DNN based enhancement scheme described in Section II-B is $257 \times (3+1) = 1028$ including the context. For the proposed combination, the dimensionality of the input features doubles to 2056 due to the concatenation of the *a priori* SNR and the *a posteriori* SNR. The output layer of all considered DNN based enhancement schemes has the dimension 257. The number of units in each hidden layers amounts to 1024 for both DNN based approach. For all employed enhancement schemes, the minimum gain is set to $G_{\text{min}} = -15$ dB.

The employed background noises are taken from a fixed pool of eight noise types. It includes the babble noise and the factory 1 noise taken from the NOISEX-92 database [36]. Further, a modulated version of NOISEX-92's pink and white noise are included similar to [13]. The remaining noise types are taken from the freesound database http://www.freesound.org. Among them are the sounds of an overpassing propeller plane[1], the interior of a passenger jet during flight[2], a vacuum cleaner[3], and a traffic noise[4]. In the evaluation, different subsets of the these noise types are taken for training while the remaining noise types are used for testing. The specific choices of the test and training sets are described in the respective parts of Section IV.

The speech material is taken from the TIMIT database [37]. For the training of the DNN based enhancement schemes, a gender balanced set of 2392 sentences taken from the TIMIT training set is employed. All sentences are embedded once in each noise type used for training at a random temporal position. For the employed noise PSD estimator, a two second initialization period is added at the beginning of each sentence to avoid initialization artifacts during feature extraction. This period is removed from the final features used for training. To allow the DNN to learn the effect of different SNRs, the sentences are embedded in the background noise at SNRs ranging from -5 dB to 15 dB. The SNR is randomly chosen for each sentence and also the scaling is randomly varied for each sentence by adjusting the peak level of the speech signal from -26 dB and -6 dB. These variations are included in the training data, to allow the DNN to learn a scale-independent function of the IRM.

The parameters of the DNN are adapted by minimizing the following optimization criterion

$$J = \sum_{\ell} \sum_{k} \left| \log\left(\widehat{\text{IRM}}(k,\ell) + \varepsilon\right) - \log\left(\text{IRM}(k,\ell) + \varepsilon\right) \right|^2. \tag{9}$$

Here, the squared error of the logarithmized quantities is minimized which is motivated by the human loudness perception which approximately follows a logarithmic law. Further, ε is a bias term which is used to avoid that extremely low gains of the target IRM are overly penalized by the cost function. Here, $\varepsilon = 0.1$ is employed such that differences between the target IRM and the DNN output are treated as irrelevant if the target IRM is below -20 dB. The weights and biases of the layers are initialized using the Glorot method [38]. After the initialization, the weights are optimized using the AdaGrad approach [39] where the learning rate has been set to 0.005 while a batch size of 8192 samples has been used. The order of the training observations is randomized. To avoid overfitting of the network, an early stopping scheme is employed where the training procedure is stopped if the error J is not reduced by more than $1\,\%$ over 20 iterations on a validation set. The validation set is constructed by randomly selecting 15 % of the training set.

[1]https://freesound.org/s/115387/
[2]https://freesound.org/s/188810/
[3]https://freesound.org/s/271460/
[4]https://freesound.org/s/75375/

IV. Instrumental Evaluation

In this section, the performance of the DNN based and the non-DNN based enhancement scheme, as well as, the combination of the two is evaluated using instrumental measures. For this, 128 sentences from the TIMIT test set are taken where it is ensured that the audio material is gender balanced. Similar to the training, the clean speech sentences are embedded at random positions in the background. All sentences are mixed at SNR ranging from -5 dB to 20 dB in 5 dB steps. Furthermore, also here, an initialization period of two seconds is added to avoid initialization artifacts of the employed noise PSD estimator [13]. This period is omitted during the evaluation, i.e., the instrumental measures are only evaluated on the part that contains the embedded sentence. Also during the evaluation, the scaling is varied between -26 dB and -6 dB peak level.

The algorithms are evaluated using the Perceptual Evaluation of Speech Quality (PESQ) [40] measure which is used as measure of the quality of the enhanced signals. Additionally, the segmental SNR (SegSNR), segmental noise reduction (SegNR), and segmental speech SNR (SegSSNR) are employed [41], [42]. The measures are defined as

$$\text{SegSSNR} = \frac{10}{\mathbb{V}} \sum_{\ell \in \mathbb{V}} \log_{10} \frac{\sum_{m=0}^{M-1} s^2[\ell M + m]}{\sum_{m=0}^{M-1} \left(s[\ell M + m] - \tilde{s}[\ell M + m]\right)^2} \tag{10}$$

$$\text{SegNR} = \frac{10}{\mathbb{V}} \sum_{\ell \in \mathbb{V}} \log_{10} \frac{\sum_{m=0}^{M-1} n^2[\ell M + m]}{\sum_{m=0}^{M-1} \tilde{n}^2[\ell M + m]} \tag{11}$$

$$\text{SegSNR} = \frac{10}{\mathbb{V}} \sum_{\ell \in \mathbb{V}} \log_{10} \frac{\sum_{m=0}^{M-1} s^2[\ell M + m]}{\sum_{m=0}^{M-1} \left(s[\ell M + m] - \hat{s}[\ell M + m]\right)^2}. \tag{12}$$

For (10)–(12), the segments have a length of $M = 160$ samples which corresponds to 10 ms and the set $\mathbb{V}$ denotes all voice active segments. Here, $s[\cdot]$ and $n[\cdot]$ denote the time-domain signals of the unprocessed clean speech signal and noise signal, respectively. Further, $\hat{s}[\cdot]$ is the time-domain representation of

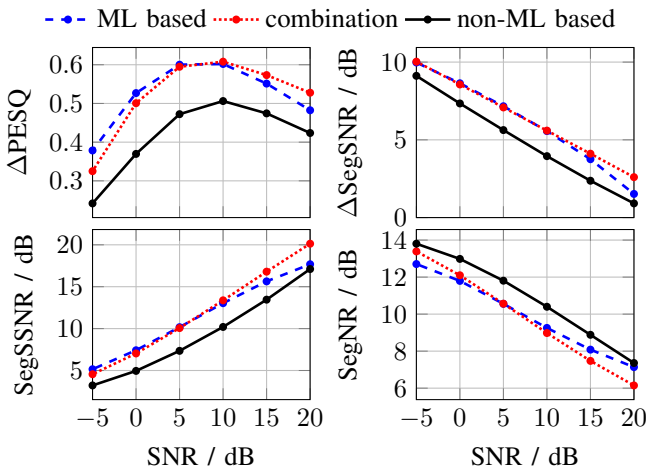

Fig. 1. PESQ improvements, segmental SNR improvements, segmental speech SNR, and segmental noise reduction averaged over the noise types seen during training, i.e., the training set in Table I (mod. pink noise, mod. white noise, traffic noise, factory 1 noise). Only unseen realizations of the known noise types are employed in the evaluation.

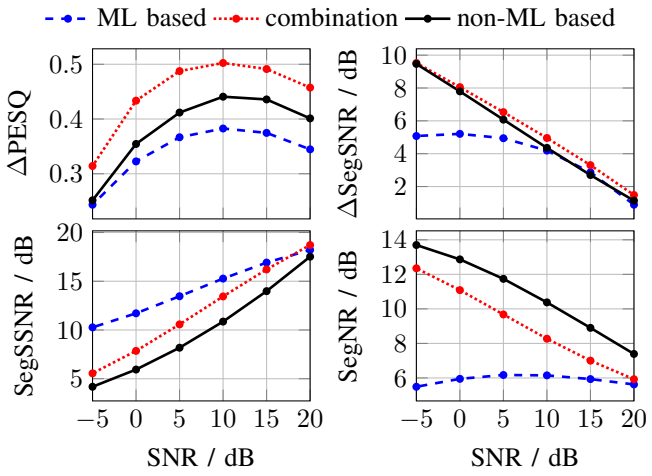

Fig. 2. The same as Fig. 1 but for the unseen noise types, i.e., the testing set in Table I (babble noise, overpassing plane, aircraft interior, vacuum cleaner).

the enhanced signal, while $\tilde{s}[\cdot]$ and $\tilde{n}[\cdot]$ result from applying the filter functions used to obtain $\hat{S}_{k,\ell}$ to the speech only and noise only signals [41]. Similar to PESQ, the segmental SNR is a measure of the enhanced signal's quality. The segmental speech SNR is a measure of the distortion of the speech signal, while the segmental noise reduction shows the amount of background noise reduction. For all employed measures, larger values indicate improvements in performance.

A. Performance in Matching and Non-Matching Conditions

The first experiment focuses on how the performance of the ML based algorithms changes if the background noise has been seen during training or not. For this, we make the selection for the training and testing noise types as given in Table I. As we are also interested in the performance for seen noise types in this experiment, only the first 120 seconds of the noise types in the training set are actually used for training. This approximately corresponds to half of the audio material

TABLE I
SELECTION OF TRAINING AND TESTING NOISE TYPES USED FOR THE EXPERIMENTS IN SECTION IV-A. ONLY THE FIRST 120 S OF THE TRAINING NOISE TYPES ARE EMPLOYED FOR TRAINING.

set	noise types
training	mod. pink; mod. white; factory 1 noise; traffic noise
testing	aircraft interior; babble noise; overpassing plane; vacuum cleaner

available for training. During testing, only the last part is used. As a result, there is no case where the specific realization of background noise used for testing has been seen during training, even if the noise type is known to the ML based enhancement schemes. Additionally, the algorithms are also evaluated on the unseen noise types in the test set in Table I.

Fig. 1 shows the results of the instrumental measures averaged over all seen noise types. Similarly, Fig. 2 shows the results for the unseen noises types. Fig. 1 confirms that the ML based speech enhancement schemes may outperform non-ML based enhancement schemes. Over a large range of SNRs, the PESQ improvements of both ML based enhancement schemes are 0.1 points larger compared to the non-ML based enhancement scheme. Furthermore, using the proposed combination, similar results can be achieved as using the state-of-the-art DNN based enhancement scheme denoted by "ML based". Similarly also the segmental SNR measure indicates improvements of both ML based enhancement schemes over the non-ML based enhancement scheme. Considering the segmental speech SNR and the segmental noise reduction, the higher PESQ and segmental SNR scores appear to be explained by the lower speech distortion which comes at the cost of a moderately reduced noise reduction.

Considering the results obtained for the unseen noise types shown in Fig. 2, the results show that the performance of the ML based approach that uses only the noisy log-spectral coefficients drops if the noise types used for training and testing differ. However, the proposed combination still shows improvements over the non-ML based enhancement approach which confirms the advantage of the proposed combination. Considering the segmental SNR, the performance of the ML based enhancement scheme which only employs noisy log-spectra is clearly reduced. The noise reduction is much lower compared to the seen noise types, which also explains the lower speech distortions indicated by the segmental speech SNR. From this it can be followed that using these features, the ML based enhancement scheme does not generalize well to unseen noise types. In contrast, if the proposed combination is employed, the segmental speech SNR and noise reduction remain comparable to the seen noise types. This indicates that these features make it easier for the DNN to learn a mapping that generalizes to unseen noise types as it can rely on the noise-robust single-channel estimation algorithms, while still benefiting from the advantages of modern ML approaches.

B. Cross-Validation Experiments

To ensure that the results shown in Fig. 1 and Fig. 2 are not restricted to the specific choice of noise types used for training

TABLE II
SEEN AND UNSEEN NOISE TYPES DURING TRAINING FOR THE CROSS-VALIDATION SETUP WITH FOUR SEEN NOISES AND EIGHT VARIATIONS. ●: SEEN NOISE TYPE, ○: UNSEEN NOISE TYPE.

	1	2	3	4	5	6	7	8
mod. pink	●	●	●	●	○	○	○	○
mod. white	●	○	○	○	●	●	○	○
factory 1 noise	○	●	●	○	●	○	●	○
traffic noise	●	●	○	○	○	○	●	○
babble noise	○	●	○	●	○	●	○	●
overpassing plane	●	○	○	○	●	●	○	●
aircraft interior	○	○	●	●	●	●	●	●
vacuum cleaner	○	○	●	●	○	○	●	●

TABLE III
CROSS-VALIDATION SETUP AS IN TABLE II BUT WITH SEVEN SEEN NOISE TYPES FOR EACH GROUP. ●: SEEN NOISE TYPE, ○: UNSEEN NOISE TYPE.

	1	2	3	4	5	6	7	8
mod. pink	●	●	●	●	●	●	●	○
mod. white	●	●	●	●	●	●	○	●
factory 1 noise	●	●	●	●	●	○	●	●
traffic noise	●	●	●	●	○	●	●	●
babble noise	●	●	●	○	●	●	●	●
overpassing plane	●	●	○	●	●	●	●	●
aircraft interior	●	○	●	●	●	●	●	●
vacuum cleaner	○	●	●	●	●	●	●	●

and testing in Section IV-A, we perform a cross-validation. Here, random combinations of the noise types taken from the pool described in Section III are used to generate different training and testing environments. First, similar to the previous experiment four noise types are used for training while the remaining noise types are used for testing. Table II shows the eight different combinations of training and testing sets used in this experiment. In contrast to the previous experiment, the performance is only evaluated on the unseen noise types. Thus, all available audio material for the training noise types is employed for training, i.e., no audio material from the training noise types is used for testing. Fig. 3 shows the average of the instrumental measures over all noise types and all training and testing combinations given in Table II.

The results show a similar picture as Fig. 2. Using the DNN only with the noisy log-spectra generally results in lower PESQ improvements and segmental SNRs improvements compared to the non-ML based approach. Again, the very low noise reduction appears to explain this behavior. Only if the proposed combination of features is employed, improvements in PESQ and the segmental SNR can be achieved over the non-ML based enhancement scheme. This, again, shows the advantage of the proposed combination over the noisy log-spectra used for the ML based approach. Also this verifies that the increased robustness towards unseen noise types observed in Section IV-A is not a result of the specific choice of training and testing noise types.

This cross-validation experiment is repeated using 7 noise types for training while the remaining noise type from the pool described in Section III is used for testing. The corresponding configurations are shown in Table III and the results are depicted in Fig. 4. Again, the results are averaged over all

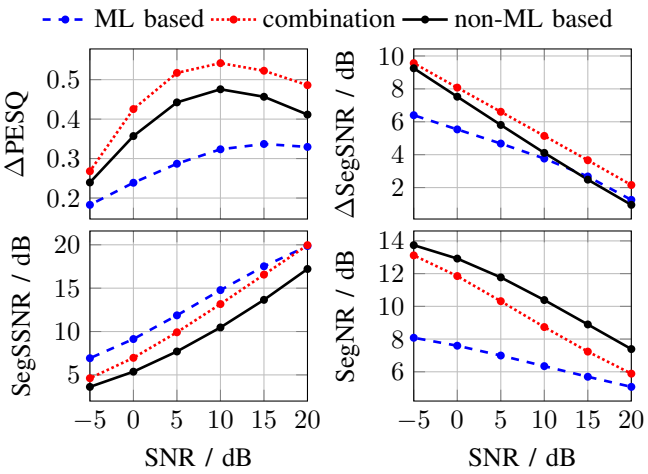

Fig. 3. PESQ improvements, segmental SNR improvements, segmental speech SNR, and segmental noise reduction averaged over all training and testing set combinations given in Table II. For each evaluation on a given testing set in Table II, the corresponding training set was used to train the ML based enhancement schemes.

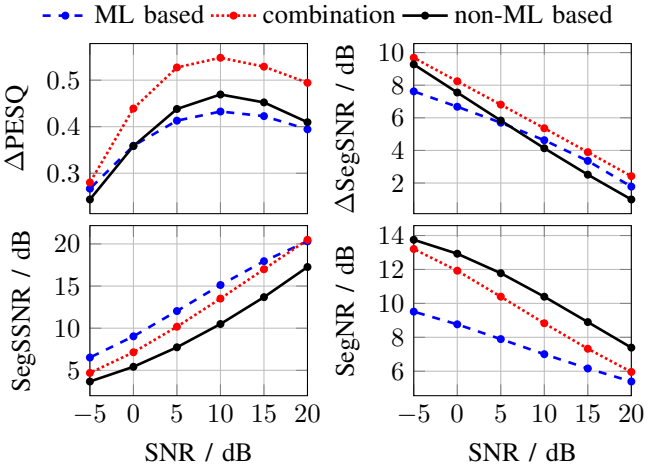

Fig. 4. Same as in Fig. 3 but using the training and testing set combinations in Table III.

unseen noise types and combinations.

Fig. 4 shows that including more noise types for training improves the generalizability if only the noisy log-spectra are used as features in the ML based estimator. In comparison to the experiments where only four noise types have been included, the performance is now comparable to the non-ML based enhancement scheme. Still, there is a gap in performance in comparison to the evaluation on seen noise types given in Fig. 1. The performance of the proposed combination features is similar to the previous experiments, i.e., improvements over the non-ML based enhancement scheme can be observed. This shows that including non-ML based features in the input of the DNN also increases the generalizability towards unseen noise conditions after adding three additional noise types to the training set.

V. Subjective Evaluation

Instrumental measures such as PESQ and the segmental SNR measures give an indication on how the quality of the processed signals would be judged by humans. Still, as such measures cannot perfectly model human perception, we verify the instrumental results in Section IV using subjective evaluation tests. Here, a multi-stimulus test with hidden reference and anchor (MUSHRA) [43] is employed to compare the algorithms described in Section II. For this experiment, a sentence of a male and a female speaker is embedded in factory 1 noise and traffic noise at an SNR of 5 dB. The noisy signals are processed by the speech enhancement schemes described in Section II. For the ML based algorithms, the same training strategies used for the instrumental evaluation in Section IV are employed again. Correspondingly, the ML based enhancement schemes are first trained on four noise types as in Table I where factory 1 and traffic noise are included in the training set. Secondly, the noisy signal is processed using the ML based enhancement schemes where the respective noise types have not been seen during training. For the enhancement of the factory 1 noise, once the training/testing set no. 6 in Table II and once the set no. 6 in Table III is used. For the traffic noise, the set no. 3 in Table II and the set no. 5 in Table III are employed.

In each trial of the experiment, the participants compared six stimuli. In addition to the processed signals, the noisy signal is included and a reference signal is presented where the speech signal and the background noise are mixed at an SNR of 20 dB. Lastly, a low quality anchor is added where the speech signal is low pass filtered at 2 kHz and mixed at an SNR of -5 dB. This signal is enhanced using a non-ML based enhancement algorithm where where the noise PSD is estimated using [13] while the speech PSD is obtained using the decision-directed approach [7]. The smoothing constant is set to 0.9 and the signal is enhanced using the Wiener filter where a more aggressive lower limit of -20 dB is employed. This results in an enhanced signal with very poor quality due to many musical tone artifacts and strong speech distortions for the anchor. The audio examples used for the listening experiment are available under https://www.inf.uni-hamburg.de/en/inst/ab/sp/publications/tasl2017-dnn-rr.

A total of 11 subjects with age in the range of 24 to 38 years who are not familiar with single-channel signal processing have participated in the MUSHRA. The experiment took place in a quiet office. The diotic signals were presented via Beyerdynamic DT-770 Pro 250 Ohm headphones attached to an RME Fireface UFX+ sound card. All signals were normalized in amplitude. The test consisted of two phases. First, the participants were asked to complete a training phase to familiarize with the presented sounds and to adjust the volume to a comfortable level. For this, a subset of the processed signals was presented. In the second part of the experiment, the participants were asked to rate the signals according to their overall preference on a scale from 0 to 100, where 0 was labeled with "bad" and 100 with "excellent". The order of the presentation of algorithms and conditions were randomized between all subjects.

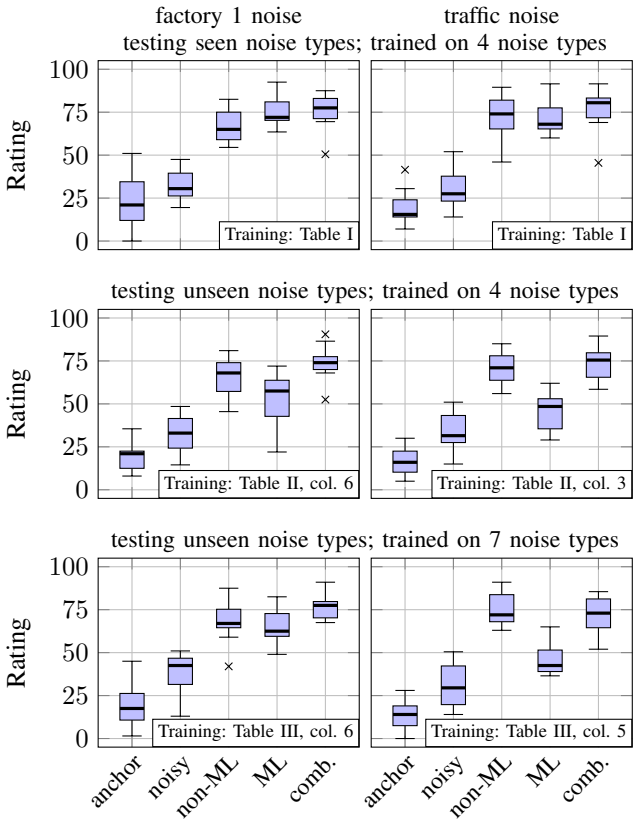

Fig. 5. Box plots for the subjective rating of different enhancement schemes. The left column shows the results for factory 1 noise and the right column for traffic noise as test signals. The text boxes in the plots refer to the training sets in the respective tables that were used to train the ML based enhancement schemes. The rows show different training strategies. For a statistical analysis, see Table IV and Table V.

For the evaluation, we average the ratings over the two speakers for each tested scenario. The results are shown in Fig. 5 in dependence of the employed noise type and the employed training/testing set for ML based algorithms. All listeners were able to correctly identify the hidden reference and assigned the highest score to it. The anchor signal and the noisy signal were assigned the lowest scores in most of the cases, i.e., the enhanced signals were generally preferred over these two signals. Comparing the ratings for the enhancement schemes, the first line of Fig. 5 indicates that, in traffic noise, all enhancement schemes have been rated similarly if the noise type has been seen during training. In factory noise, however, both ML based speech enhancement schemes yield better sounding results than the non-ML based algorithm for the same training strategy. This confirms the prediction by the instrumental measures in Fig. 1 and demonstrates again the potential of DNN based enhancement over non-ML based enhancement in the context of speech enhancement in the presence of transient noise. It also verifies that the proposed combination results in a similar performance as the state-of-the-art DNN based speech enhancement schemes denoted by "ML". Considering the case where the processed noise types have not been included in the training set, i.e., the lower two lines of Fig. 5, the ratings for the ML based approach only using the noisy log-spectra as features drop while the

TABLE IV
RESULTS OF THE REPEATED MEASURES ANOVA FOR ALL EVALUATED ACOUSTIC SCENARIOS TO TEST IF THE FACTOR "ENHANCEMENT ALGORITHM" HAS AN EFFECT ON THE RATINGS. THE NUMBERS GIVEN FOR $F(\cdot,\cdot)$ ARE THE DEGREES OF FREEDOM IN THE NUMERATOR AND DENOMINATOR OF THE F STATISTIC, RESPECTIVELY [44]. HERE, • DENOTES STATISTICAL SIGNIFICANCE, I.E., $p < 0.05$, WHILE ○ INDICATES NO STATISTICAL SIGNIFICANCE. THE ✓ INDICATES THAT THE TESTED NOISE TYPE IS INCLUDED IN THE TRAINING SET. THE GRAPHICAL REPRESENTATION OF THE DATA ANALYZED IN THIS TABLE IS GIVEN IN FIG. 5 AND PAIRWISE SIGNIFICANCE TESTS ARE GIVEN IN TABLE V.

testing noise	training set	seen	F statistic	$p < 0.05$
factory 1	Tab. I	✓	$F(4, 40) = 66.83$	•
	Tab. II, col. 6		$F(4, 40) = 42.88$	•
	Tab. III, col. 6		$F(4, 40) = 62.80$	•
traffic	Tab. I	✓	$F(4, 40) = 74.19$	•
	Tab. II, col. 3		$F(4, 40) = 69.25$	•
	Tab. III, col. 5		$F^*(1.81, 18.08) = 84.47$	•

*: Greenhouse-Geisser correction [45] with $\epsilon = 0.452$ employed.

TABLE V
RESULT OF PAIRWISE t-TEST FOR ALL ACOUSTIC SCENARIOS AND STIMULUS COMBINATIONS. THE SYMBOL • INDICATES STATISTICAL SIGNIFICANCE, I.E., $p < 0.05$, WHILE ○ INDICATES NO STATISTICAL SIGNIFICANCE. THE GRAPHICAL REPRESENTATION OF THE DATA ANALYZED IN THIS TABLE IS GIVEN IN FIG. 5 WHILE THE OVERALL SIGNIFICANCE OF THE FACTOR "ENHANCEMENT ALGORITHM" IS TESTED IN TABLE IV.

	anchor	non-ML	ML	comb.
non-ML	•			
ML	•	•		
comb.	•	•	○	
noisy	○	•	•	•

(a) factory 1 noise (seen), training set: Table I

	anchor	non-ML	ML	comb.
non-ML	•			
ML	•	○		
comb.	•	○	○	
noisy	○	•	•	•

(b) traffic noise (seen), training set: Table I

	anchor	non-ML	ML	comb.
non-ML	•			
ML	•	•		
comb.	•	•	•	
noisy	•	•	•	•

(c) factory 1 noise (unseen), training set: Table II, col. 6

	anchor	non-ML	ML	comb.
non-ML	•			
ML	•	•		
comb.	•	○	•	
noisy	•	•	○	•

(d) traffic noise (unseen), training set: Table II, col. 3

	anchor	non-ML	ML	comb.
non-ML	•			
ML	•	○		
comb.	•	•	•	
noisy	•	•	•	•

(e) factory 1 noise (unseen), training set: Table III, col. 6

	anchor	non-ML	ML	comb.
non-ML	•			
ML	•	•		
comb.	•	○	•	
noisy	•	•	•	•

(f) traffic noise (unseen), training set: Table III, col. 5

ratings for the proposed combination remain high. These results demonstrate once more that the proposed combination increases the generalizability of the ML based enhancement scheme. Additionally, the proposed combination shows higher ratings as the non-ML enhancement scheme in factory noise. As this acoustic environment contains many transient sounds, this indicates that the proposed features still allow to exploit the advantages of DNN based enhancement schemes. Contrarily, the traffic noise contains only moderately varying noise in the form of passing cars. These changes in the background noise can still be tracked reasonably well using the non-ML based approach [13] which results in only small differences between the proposed ML and the non-ML based enhancement scheme.

In the remainder of this section, we conduct a statistical analysis for the results shown in Fig. 5. For each acoustic scenario, a repeated measures analysis of variance (ANOVA) [44] is performed to test if the factor "enhancement algorithm" has a significant effect on the participants' rating. For this, we employ a significance level of 5 % for all statistical tests. For each acoustic scenario, we validated that the residuals of the general linear model fitted during the process of the repeated measures ANOVA are normally distributed using the Shapiro-Wilk test [46]. According to Mauchly's test [47], the sphericity assumption is violated for the traffic noise scenario where the ML model in Table III, col. 5 is used. For this case, we apply a Greenhouse-Geisser correction [45] with $\epsilon = 0.452$ while for the remaining cases sphericity is assumed. Table IV summarizes the results of the repeated measures ANOVA and shows that the factor "enhancement algorithm" has a statistically significant influence on the rating for all considered scenarios.

Therefore, we continue to employ post-hoc tests to identify the sources of significance. For this, pairwise comparisons using matched pair t-tests are employed where we employ a Bonferroni-Holm [48] correction to account for the error inflation of pairwise testing. The results are shown in Table V. The significance tests confirm the results discussed for Fig. 5. For most of the cases, the participant's ratings for the enhanced signals are significantly different compared to the anchor and the noisy signal. One exception is the unseen traffic noise scenario for the case where the training set in Table III, col. 3 has been employed. For this case, results indicate that using the ML based enhancement scheme using only noisy log-noisy spectra as features and the noisy signal does not yield significant improvements over simply using the noisy signal. For the scenarios where the noise has been seen during training, i.e., the first row in Table V, no significant difference in the ratings of the evaluated enhancement schemes is found for the traffic noise scenario. For factory noise, however, a significant difference in the ratings for the proposed combination and the non-ML based enhancement scheme is obtained. Additionally, the difference in ratings between the ML based enhancement scheme using only noisy log-spectra and the ML based enhancement scheme is significant. This demonstrates, again, the DNN's power in single-channel speech enhancement to outperform state-of-the-art non-ML based enhancement schemes in transient noise types. However, considering the practically more relevant cases where the background noise is unknown to the ML based enhancement schemes, the means of the ML based enhancement scheme using log-noisy features and the non-ML based scheme are significantly different for all scenarios. This means that the performance of the ML based enhancement was rated significantly *worse* compared to the non-ML based enhancement scheme in unseen noise. Comparing the proposed combination and the non-ML based enhancement scheme, the means are only significantly different for the factory 1 noise while no significant difference was found for the traffic noise. Following that, the proposed combination was rated *better* than the non-ML based enhancement scheme in factory 1

noise. Lastly, the means of the proposed combination and the compared ML based enhancement scheme are significantly different for all unseen conditions. From this it is concluded that the enhancement based on the proposed combination is rated better than the enhancement based on the noisy log-spectra only.

This strong evidence leads us to the conclusion that the proposed features significantly improve the generalization of DNN based enhancement schemes especially in cases where training material is limited. Further, while there is no significant difference between the proposed combination and the non-ML based enhancement scheme in traffic noise, the proposed DNN based algorithm is better in factory 1 noise. From this we follow that due to the fast tracking of the non-ML noise PSD estimator in [13], no difference between two enhancement approaches is observed in traffic noise. However, as the DNN is more capable of suppressing the transient noises in the factory environment, the proposed combination is rated higher there. This shows that the advantages of the ML based and non-ML based enhancement schemes are efficiently combined with the proposed approach.

VI. Conclusions

In this paper, we combined ML based and non-ML based enhancement schemes to take advantage of both approaches. For this, we proposed to use the *a priori* SNR and the *a posteriori* SNR estimated by non-ML based approaches as features in a DNN based enhancement scheme. If the noise type is known, the proposed combination performs similar to a comparable state-of-the-art ML based algorithm that uses noisy log-spectra as features. In unseen noise conditions, however, the performance drops if only the log-spectra of the noisy observations are employed while the performance of the proposed combination remains high. From this, it is concluded that the proposed combination retains the advantages of ML based approaches, e.g., the suppression capability in transient noise types, and considerably increases the generalizability to unseen noise types. This is supported by the MUSHRA based listening experiments where, in unseen noise conditions, the proposed combination was significantly preferred over the ML based enhancement scheme using log-spectra of the noisy observations. The audio examples used for the listening experiment are available under https://www.inf.uni-hamburg.de/en/inst/ab/sp/publications/tasl2017-dnn-rr.

References

[1] N. R. French and J. C. Steinberg, "Factors Governing the Intelligibility of Speech Sounds," *The Journal of the Acoustical Society of America*, vol. 19, no. 1, pp. 90–119, 1947.

[2] J. M. Festen and R. Plomp, "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," *The Journal of the Acoustical Society of America*, vol. 88, no. 4, pp. 1725–1736, Oct. 1990.

[3] D. Kolossa and R. Haeb-Umbach, *Robust Speech Recognition of Uncertain or Missing Data*, 1st ed. Springer-Verlag Berlin Heidelberg, 2011.

[4] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An Overview of Noise-Robust Automatic Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, Apr. 2014.

[5] P. Vary and R. Martin, *Digital speech transmission: Enhancement, Coding and Error Concealment*. Chichester, West Sussex, UK: Wiley & Sons, 2006.

[6] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[7] 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.

[8] 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.

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

[10] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook Driven Short-Term Predictor Parameter Estimation for Speech Enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 163–176, Jan. 2006.

[11] 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.

[12] 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. 2008, pp. 4897–4900.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] J. Porter and S. Boll, "Optimal estimators for spectral restoration of noisy speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, San Diego, CA, USA, Mar. 1984, pp. 53–56.

[18] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[19] 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.

[20] 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. 2008, pp. 4037–4040.

[21] 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.

[22] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement: A Survey of the State of the Art*, ser. Synthesis Lectures on Speech and Audio Processing, 2013, vol. 9, no. 1.

[23] M. N. Schmidt and R. K. Olsson, "Single-Channel Speech Separation using Sparse Non-Negative Matrix Factorization," in *Interspeech*, Pittsburgh, PA, USA, Sep. 2006, pp. 1652–1655.

[24] S. Srinivasan, J. Samuelsson, and W. Kleijn, "Codebook-Based Bayesian Speech Enhancement for Nonstationary Environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 441–452, Feb. 2007.

[25] D. Burshtein and S. Gannot, "Speech Enhancement Using a Mixture-Maximum Model," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 6, pp. 341–351, Sep. 2002.

[26] 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.

[27] M. Sun, Y. Li, J. F. Gemmeke, and X. Zhang, "Speech Enhancement Under Low SNR Conditions Via Noise Estimation Using Sparse and Low-Rank NMF with Kullback-Leibler Divergence," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 7, pp. 1233–1242, Jul. 2015.

[28] P. S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint Optimization of Masks and Deep Recurrent Neural Networks for Monaural Source Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2136–2147, Dec. 2015.

[29] J. Chen, Y. Wang, S. E. Yoho, D. Wang, and E. W. Healy, "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *The Journal of the Acoustical Society of America*, vol. 139, no. 5, pp. 2604–2612, 2016.

[30] Q. He, F. Bao, and C. Bao, "Multiplicative Update of Auto-Regressive Gains for Codebook-Based Speech Enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 457–468, Mar. 2017.

[31] 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.

[32] A. Kumar and D. Florencio, "Speech Enhancement in Multiple-Noise Conditions Using Deep Neural Networks," in *Interspeech*, San Francisco, CA, USA, 2016, pp. 3738–3742.

[33] 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.

[34] 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.

[35] V. Nair and G. E. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines," in *International Conference on Machine Learning*, Haifa, Israel, Jun. 2010, pp. 807–814.

[36] 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.

[37] 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.

[38] 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.

[39] J. C. Duchi, E. Hazan, and Y. Singer, "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization," *Journal of Machine Learning Research*, vol. 12, pp. 2121–2159, 2011.

[40] "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.

[41] T. Lotter and P. Vary, "Speech Enhancement by MAP Spectral Amplitude Estimation Using a Super-Gaussian Speech Model," *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 7, pp. 1–17, 2005.

[42] 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.

[43] "BS.1534-3: Method for the subjective assessment of intermediate quality levels of coding systems," International Telecommunication Union, ITU-T recommendation, Oct. 2015.

[44] A. Field, *Disocvering Statistics Using SPSS*, 3rd ed. SAGE Publications Ltd., 2009.

[45] S. W. Greenhouse and S. Geisser, "On methods in the analysis of profile data," *Psychometrika*, vol. 24, no. 2, pp. 95–112, Jun. 1959.

[46] S. S. Shapiro and M. B. Wilk, "An Analysis of Variance Test for Normality (Complete Samples)," *Biometrika*, vol. 52, no. 3/4, pp. 591–611, 1965.

[47] J. W. Mauchly, "Significance Test for Sphericity of a Normal n-Variate Distribution," *The Annals of Mathematical Statistics*, vol. 11, no. 2, pp. 204–209, Jun. 1940.

[48] S. Holm, "A Simple Sequentially Rejective Multiple Test Procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979.