AN ANALYSIS OF NOISE-AWARE FEATURES IN COMBINATION WITH THE SIZE AND DIVERSITY OF TRAINING DATA FOR DNN-BASED SPEECH ENHANCEMENT

Robert Rehr, Timo Gerkmann

Signal Processing Group, Department of Informatics, Universität Hamburg, Germany
robert.rehr@uni-hamburg.de, timo.gerkmann@uni-hamburg.de

ABSTRACT

In this work, the generalization of speech enhancement algorithms based on deep neural networks (DNNs) for training datasets that differ in size and diversity is analyzed. For this, we compare noise aware training (NAT) features and signal-to-noise ratio (SNR) based noise aware training (SNR-NAT) features. NAT appends an estimate of the noise power spectral density (PSD) to a noisy periodogram input feature, whereas SNR-NAT uses the noise PSD for normalization. We show that the Hu noise corpus (limited size) and the CHiME 3 noise corpus (limited diversity) may result in DNNs which do not generalize well to unseen noises. We construct a large and diverse dataset from freely available data and show that it helps DNNs to generalize. However, we also show that with SNR-NAT features, the trained models are more robust even if a small or less diverse training set is employed. Using t-distributed stochastic neighbor embedding (t-SNE), we demonstrate that using SNR-NAT both the features and the resulting internal representation of the DNN are less dependent on the background noise which facilitates the generalization to unseen noise types.

Index Terms— Deep neural networks, generalization, speech enhancement, noise reduction, input features

1. INTRODUCTION

Speech is commonly used for communication, e.g., to share ideas and emotions with others. With the increase of powerful mobile personal devices, speech plays a central role in many applications, e.g., in telecommunications, hearing aids and personal home assistants. As many devices are used in adverse acoustic situations, e.g., on a noisy street or a noisy household, the microphones do not only capture the desired speech signal, but also noise. To counteract detrimental effects of the noise, speech enhancement algorithms are employed. In this paper, single-channel speech enhancement algorithms are considered. Such methods can be used to enhance a noisy signal captured by a single microphone or to improve the output of a spatial filter.

Single-channel speech enhancement is an active research field since many decades [1–5]. As a consequence, many different approaches have been presented to solve the problem. Often, the noisy input signal is represented in a time-frequency domain using the short-time Fourier transform (STFT). In this domain, a real-valued mask is applied which suppresses the coefficients which mainly contain noise and preserves the coefficients dominated by speech. This effectively results in a time-varying filtering where the frequency bands that mainly contain noise are attenuated. Many conventional approaches [1, 3, 6] are embedded in a statistical framework which is used to derive optimal noise suppression filters that minimize a given error criterion. For the derivations, it is often assumed that the speech and noise coefficients follow a known probability density distribution parameterized by the speech power spectral density (PSD) and noise PSD, respectively. The PSDs are estimated blindly from the noisy observation using algorithms derived based on statistical models and signal processing models [1, 7, 8]. Noise PSD estimators are typically based on the assumption that background noise varies slower than speech. While this makes these approaches applicable for a wide range of acoustic conditions, highly non-stationary noises, e.g., the cutlery in a restaurant, are not suppressed.

The shortcomings of conventional speech enhancement algorithms motivated the application of machine-learning (ML) for speech enhancement. In contrast to the conventional methods, ML-based methods learn the statistics of the two components prior to the enhancement process. In recent years, many approaches leverage the learning capacity of deep neural networks (DNNs) for speech enhancement, e.g., [5, 9, 10]. Even though learning based approaches show the potential to suppress highly non-stationary noise types, their generalization to unseen conditions is not guaranteed [11–16]. Thus in [11, 14, 16], noise aware training (NAT) [17] has been proposed for improving the robustness of DNN-based speech enhancement methods in unseen acoustic conditions. This approach appends an estimated noise PSD spectrum to noisy periodogram input features. In contrast, we proposed signal-to-noise ratio (SNR) based noise aware training (SNR-NAT) in [18] where the noise PSD is used for normalization instead of appending it. The SNR-NAT features are related to the *a priori* SNR and the *a posteriori* SNR and result in DNNs that are more robust in unseen acoustic conditions [18].

In this paper, we continue this work and conduct experiments on various training sets, which differ in the amount and diversity of the background noise data. Our contributions are as follows: (1) We show that the enhancement performance of a feed-forward DNN strongly depends on the type of training data if the NAT features are employed. (2) We show that, in this case, datasets like the Hu noise corpus [19] or the CHiME 3 noise corpus [20] do not allow the resulting DNN to generalize well to unseen acoustic conditions. (3) We further show that the generalization can be improved using a large and diverse training set which we propose to construct using sounds from the freesound.org website. Even though the Hu noise corpus [19] and the CHiME 3 noise corpus [20] do not lead to a general DNN if the NAT features are used, the limited datasets are sufficient to learn a general DNN if the SNR-NAT features are employed. (4) For better understanding, we analyze the input features, as well as, the internal representation of the trained DNNs using t-distributed stochastic neighbor embedding (t-SNE) [21] for visualization. The graphical analysis gives a simple intuition of the features' behavior and shows that the NAT features and the resulting internal representation are less dependent on the background noise as compared to the NAT features. This property appears to play an important role in increasing the robustness for DNN-based speech enhancement.

The paper is structured as follows: Section 2 describes the employed DNN-based enhancement scheme and its input features. Section 3 describes the employed training data and explains the training procedure. In Section 4, the considered enhancement schemes are evaluated using wideband Perceptual Evaluation of Speech Quality (WB-PESQ) [22] and the feature analysis is conducted in Section 5.

2. DNN-BASED SPEECH ENHANCEMENT ALGORITHM

In this paper, we analyze the behavior of a DNN-based enhancement algorithm for five different input features. In this work, the time-domain signal is sampled using a rate of 16 kHz. The considered speech enhancement algorithm leverages the STFT. Correspondingly, the time-domain signal is split into overlapping segments which are transformed using discrete Fourier transform after a square-root Hann window has been applied. This procedure yields the time-frequency representation of the clean speech signal $S_{k,\ell}$, the noise signal $N_{k,\ell}$ and the noisy input signal $Y_{k,\ell}$. The symbols k and ℓ represent the frequency index and the time index, respectively. In our experiments, the segment length is set to 32 ms, i.e., 512 samples, and the segments overlap by 50 %.

The signal is enhanced in the spectral domain. For this, we employ a feed-forward neural network which comprises three hidden layers with 1024 rectifying linear units (ReLUs) and an output layer with 257 sigmoid units. The network is used to map features, which are extracted from the noisy observation $Y_{k,\ell}$, to a masking function $G_{k,\ell}$. In our study, we employ an ideal ratio mask (IRM) which has been proposed for speech enhancement in [23]

$$G_{k,\ell}^{\text{IRM}} = \frac{|S_{k,\ell}|^2}{|S_{k,\ell}|^2 + |N_{k,\ell}|^2}. \tag{1}$$

The masking function is used to estimate the clean speech spectrum as

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

In (2), $\hat{G}_{k,\ell}^{\text{IRM}}$ denotes an estimate of the IRM which is obtained from a trained DNN. The quantity $G_{\min}$ introduces a lower limit on the IRM which has been proposed in [24] to reduce speech distortions and artifacts. In this work, $G_{\min}$ is set to -20 dB. The estimated clean speech signal is obtained using an overlap-add procedure after the enhanced spectra have been transformed to the time-domain and a square-root Hann window has been applied for synthesis.

In the remainder of this section, the input features of the DNN are considered. First, we describe the logarithmized noisy periodogram which is the basis of the NAT features. The feature vector of a single frame ℓ of the logarithmized periodogram is given by

$$\mathbf{v}_\ell^{(Y)} = \left[\log(|Y_{0,\ell}|^2), ..., \log(|Y_{K,\ell}|^2)\right]^T. \tag{3}$$

Here, $\cdot^T$ is the vector transpose and K corresponds to the number sampling points of the discrete Fourier transform, where the mirror spectrum is omitted. As consequence, the feature contains only the first K sampling points, i.e., 257 points for the employed sampling rate and STFT parameters.

NAT has been considered in [5,11,14,16,17] to improve the robustness of DNN-based speech enhancement approaches in acoustic conditions not seen during training. These features are constructed by appending an estimate of the noise PSD $\Lambda_{k,\ell}^{\text{n}}$ to the logarithmized noisy input periodogram. The noise PSD is also logarithmized and the coefficients of a frame are stacked in a vector as

$$\mathbf{v}_\ell^{(\Lambda^{\text{n}})} = \left[\log(\Lambda_{0,\ell}^{\text{n}}), ..., \log(\Lambda_{K,\ell}^{\text{n}})\right]^T. \tag{4}$$

The NAT features are then given by the concatenation of $\mathbf{v}_\ell^{(Y)}$ and $\mathbf{v}_\ell^{(\Lambda^{\text{n}})}$ as

$$\mathbf{v}_\ell^{(\text{NAT})} = [(\mathbf{v}_\ell^{(Y)})^T, (\mathbf{v}_\ell^{(\Lambda^{\text{n}})})^T]^T. \tag{5}$$

Due to the concatenation of two input feature vectors, the dimensionality of the input features doubles to 514. For estimating $\Lambda_{k,\ell}^{\text{n}}$, we use the conventional noise PSD estimator described in [8]. The noise estimator leverages a conventionally estimated speech presence probability [8] and allows to track moderately changing background noises, e.g., passing cars on a busy street.

username	list of ids
Robinhood76	3238, 3246, 3667, 3668, 3729, 3830, 3870, 3873, 3971, 3979, 3980, 4024, 4025, 4026, 4036, 4058, 4065, 4149, 4364, 5589
rutgermuller	20158

Table 1. Sound packs that form the large and diverse freesound.org noise dataset. The packs can be downloaded by replacing <username> and <id> in freesound.org/people/<username>/packs/<id> by the data above.

Additionally, we also consider the SNR-NAT features, i.e., the logarithmized *a priori* SNR and the logarithmized *a posteriori* SNR. These features have been used, e.g., in [25, 26] for data-driven speech enhancement approaches which, however, did not consider DNNs. In [18], it has been shown that the features result in more robust DNN-based enhancement algorithms if the employed training dataset is limited in size and diversity. The *a priori* SNR is given by $\xi_{k,\ell} = \Lambda_{k,\ell}^{\text{s}} / \Lambda_{k,\ell}^{\text{n}}$ while the *a posteriori* SNR is given by $\gamma_{k,\ell} = |Y_{k,\ell}|^2 / \Lambda_{k,\ell}^{\text{n}}$. The feature vector of the SNR-NAT features is correspondingly given by

$$\mathbf{v}_\ell^{(\text{SNR})} = [\log(\xi_{0,\ell}), ..., \log(\xi_{K,\ell}), \log(\gamma_{0,\ell}), ..., \log(\gamma_{K,\ell})]^T. \tag{6}$$

The speech PSD $\Lambda_{k,\ell}^{\text{s}}$ is also estimated using a conventional approach. For this, we use the cepstral smoothing techniques which has been described in [7]. In contrast to the well known decision-directed approach [1], this approach causes less artifacts in the estimate of the speech PSD.

For all features, a context over three previous segments is included. For this, the vectors for the respective feature are stacked into a super-vector $\tilde{\mathbf{v}}_\ell$ as

$$\tilde{\mathbf{v}}_\ell = [\mathbf{v}_\ell^T, ..., \mathbf{v}_{\ell-3}^T]^T. \tag{7}$$

By using the context, the dimensionality is raised by factor 4, i.e., from 257 to 1028 or from 514 to 2056, respectively.

3. EXPERIMENTAL SETUP

For the training of the DNN-based enhancement scheme above, we use background noises from three different noise sets: the Hu noise corpus [19] with the extension presented in [27], the CHiME 3 noise corpus [20] and a handcrafted noise set created from sound packs available from freesound.org. For simplicity, we refer to the Hu noise corpus and its extension just as Hu noise corpus. The noise sets vary in the amount and diversity of the data as we illustrate in more detail below.

The Hu noise corpus [19] contains 100 non-speech sounds and the extension[1] presented in [27] adds another 15 sounds. Even though many noise types are included, most of the noise recordings are rather short with a length shorter than ten seconds. The total duration of the noise content is about 14 minutes.

For the CHiME 3 noise corpus [20], recordings from four acoustic environments have been obtained using a tablet equipped with six microphones. As we consider single-channel speech enhancement, we use only the audio material from the first microphone. The environments include the ride on a bus, the interior of a cafe, a pedestrian area and a street junction. As only four different environments are included, the diversity of the dataset is relatively low, but the total available noise data is duration quite large and amounts to about 8.5 hours.

The third dataset is constructed from various sound packs published on https://www.freesound.org. Table 1 gives an overview of the used packs

[1] http://staff.ustc.edu.cn/~jundu/The%20team/yongxu/demo/115noises.html

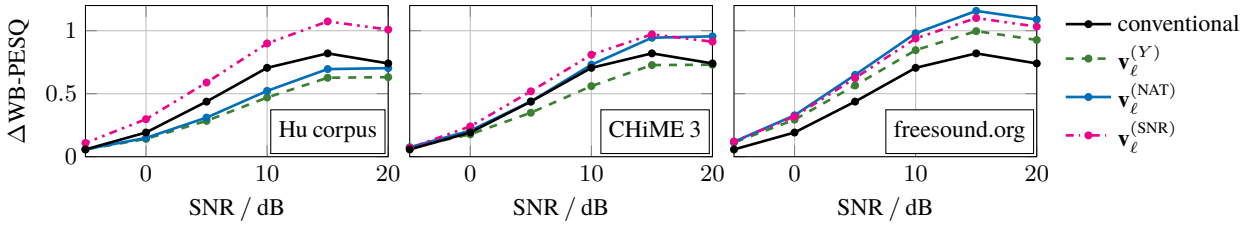

Fig. 1. WB-PESQ improvements averaged over all testing noise types in dependence of the input SNR, the training dataset and the input feature. Further, the results of a conventional speech enhancement approach have been added.

and provides the links to the respective downloads. From these packs, we exclude all sound files whose duration is shorter than 30 seconds. After this, 282 sound files from many different acoustic environments remain and the total duration of the audio material is about 13.5 hours. In contrast to the Hu noise corpus [19, 27] and the CHiME 3 noise corpus [20], the created dataset has both a large amount of data and also a large diversity, i.e., recordings from many different environments.

The noisy training data is generated by artificially corrupting noise free sentences of the TIMIT training set [28] using the background noise of one of the noise sets described above. We use 3992 sentences where we ensure that the number of sentences spoken by male and female speakers is the same. The total duration of the speech training data is about two hours. With these sentences, a training dataset of about 100 hours is generated by allowing each sentence to be reused 49 times. Each sentence, including the reused ones, is randomly embedded in a background noise of the employed noise set. For this, a different excerpt is selected for each sentence. To allow the conventional noise PSD estimator to adapt to the acoustic environment, it is ensured that the first two seconds contain only background noise. After the feature extraction, this part is removed from the training data. For many speech communication applications like hearing aids or telephony, this two second initialization is not an issue. By not excluding it, the results would be strongly impacted by initialization artifacts and thus not be representative for many speech communication applications. The input SNR is randomly chosen between -10 dB and 15 dB for each sentence to allow the DNN to learn how to cope with strongly and weakly corrupted input signals. Additionally, the time-domain peak level of each sentence is varied between -26 dB and -3 dB before being corrupted by the background noise. This changes the overall level for each sentence during training. The variations are included to make the DNN-based enhancement independent of the overall level of the input signal. For approximately 10 % of the overall training data, it is ensured that only the background noise is present to enable the DNN to reject noise only segments.

The generated data is split into a training and a validation set. The validation set corresponds to 15 % of the overall data while the remaining data is used for training. The parameters of the DNNs are initialized using the method described in [29]. After this, we use stochastic gradient descent to reduce the squared error between the predicted IRM and the true IRM, i.e.,

$$J = \sum_k \sum_\ell \left| \hat{G}_{k,\ell}^{\mathrm{IRM}} - G_{k,\ell}^{\mathrm{IRM}} \right|^2. \tag{8}$$

The learning rate is reduced from 0.4 to 0.1 over the training epochs using an exponential decay as $\mathrm{LR} = \max(0.4 \cdot 0.95^{E-1}, 0.1)$. Here, LR is the learning rate and E is the current training epoch. All models are trained for 100 epochs and we select the model with the lowest error on the validation set for testing.

4. INSTRUMENTAL EVALUATION

For the instrumental evaluation, we use speech and noise material that has not been used during training. For testing, ten noise types taken from the NOISEX-92 database [30] and freesound.org are employed. We use the "babble", "factory 1", "f16" and "hfchannel" environment from the NOISEX-92 database. Additionally, we modulate the amplitude of the white and the pink noise from the NOISEX-92 database with a 0.5 Hz sinusoid and include the modified signals in the evaluation. The remaining four noise types are an aircraft interior noise (freesound.org/s/188810), an overpassing propeller plane (freesound.org/s/115387), traffic noise (freesound.org/s/252216) and a vacuum cleaner (freesound.org/s/67421).

The noise data are used to artificially corrupt 128 sentences taken from the TIMIT test set [28]. Again, it is ensured that the number of sentences spoken by female and male speaker is the same. Each sentence is corrupted by all background noises at six input SNRs ranging from -5 dB to 20 dB in 5 dB steps. Similar to the training, the overall level of the test mixtures is varied by changing the speech peak level in the time-domain between -26 dB and -3 dB.

The noisy signals are enhanced using the DNN-based speech enhancement algorithm which are trained using the different input features from Section 2 and the three different noise data sets described in Section 3. Additionally, we include the results of a conventional speech enhancement approach which estimates the clean speech coefficients using the Wiener filter. For estimating the noise PSD and the speech PSD, the same conventional PSD estimators are used which are also used for extracting the SNR-NAT features, i.e., [7, 8].

The performance of the enhancement algorithms is compared using WB-PESQ [22], an instrumental measure to predict the perceived quality of a processed speech signal. Fig. 1 shows the WB-PESQ improvements averaged over all testing noise types in dependence on the input SNR, the training data and the used input feature. The results show that the performance of the DNN-based enhancement approach depends on the training data. The performance may considerably deteriorate if the training data set lacks size (Hu noise corpus [19]) or diversity (CHiME 3 noise corpus [20]) and the NAT or the periodogram features are employed. Only using the proposed large and diverse freesound.org training dataset, the DNN-based enhancement algorithm with the periodogram or NAT features is able to clearly improve the quality over the conventional approach. However, using the SNR-NAT features as input makes the training more robust and less susceptible to insufficient training data. As a result, the performance of the DNN-based enhancement scheme is almost independent of the training data with the SNR-NAT features. From these observations, we conclude that (1) a robust model can be trained if appropriate training data are available and (2) that the normalization considerably improves the robustness of the DNN-based speech enhancement approach in unseen acoustic conditions even if insufficient training data are available. In the next section, we provide further insights on this behavior by analyzing the training features and the internal representations of the DNN.

5. ANALYSIS

In this part, we interpret the WB-PESQ results obtained in Section 4 by analyzing the input features, the internal states and the output of the DNN.

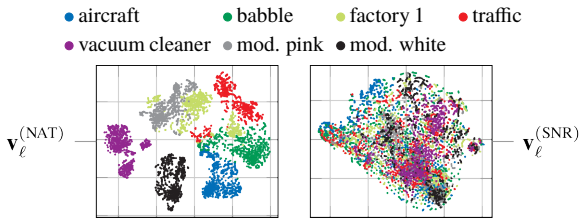

Fig. 2. t-SNE of the input feature vectors extracted from four sentences with a peak level of −6 dB and corrupted by seven different noise types at an SNR of 5 dB. The color of the data points indicates the noise type.

For this, we employ t-SNE [21] which is a method to embed data vectors from a high dimensional space in a low dimensional space. For this, we extracted the NAT features $\mathbf{v}_\ell^{(\mathrm{NAT})}$ and the SNR-NAT features $\mathbf{v}_\ell^{(\mathrm{SNR})}$ from artificially corrupted sentences of two female and two male speakers. For this analysis, the maximum peak level of the speech signal is set to a fixed level of −6 dB and the input SNR is set to 5 dB. The four sentences are corrupted by seven different noise types as shown in Fig. 2. Each point in the plots corresponds to a high-dimensional feature vector and its color indicates the noise type from which the vector has been extracted.

Fig. 2 depicts the t-SNE of the raw input features. For improving the clarity of the embeddings, the context of the features is omitted, i.e., we do not include the three previous frames as in (7). For the NAT features $\mathbf{v}_\ell^{(\mathrm{NAT})}$, it can be observed that the feature vectors form clusters based on the noise type. Further, there is little overlap between the clusters such that the data points appear to be easily separable by the noise type. From this, we conclude that the features strongly depend on the background noise type. In contrast to the NAT features $\mathbf{v}_\ell^{(\mathrm{NAT})}$, the clustering is considerably weaker for the embedding of the SNR-NAT features $\mathbf{v}_\ell^{(\mathrm{SNR})}$ and a straight-forward separation of the noise types feature is not possible. As a consequence, the SNR-NAT features are less dependent on the noise type than the NAT features.

For Fig. 3, the same corrupted speech files as in Fig. 2 are used, but here t-SNE is applied to the output of the DNN's second last layer, i.e., an internal representation, and the estimated clean speech coefficients. The former is shown in the upper two rows of Fig. 3 while the latter is depicted in the lower two rows. The figure shows the embeddings for the three training datasets described in Section 3. The NAT features $\mathbf{v}_\ell^{(\mathrm{NAT})}$ lead to an internal representation that also depends on the background noise type, while the internal representation appears to be less dependent on the noise type if the SNR-NAT features $\mathbf{v}_\ell^{(\mathrm{SNR})}$ are employed. These observations hold for all training datasets.

In the third row of Fig. 3, the embeddings of the estimated speech coefficients are depicted that have been obtained using the NAT features. The structure of these embeddings clearly depends on the training dataset. Considering the Hu noise corpus [19] (small size) or the CHiME 3 noise corpus [20] (low diversity), i.e., the first two columns, the embeddings form clusters based on the noise type. Note that this is typically an undesired effect as the estimated speech coefficients should ideally be independent of the underlying background noise. The clustering is weaker for the speech estimates that are obtained from a model trained on the freesound.org data and correspondingly this model yields higher WB-PESQ scores as shown in Fig. 1. In the last row of Fig. 3, the embeddings of the estimated speech coefficients that are obtained using SNR-NAT features are shown. Here, the structure of the embeddings is similar for all training datasets and resembles the structure that is obtained from the best performing model which is trained on the freesound.org dataset.

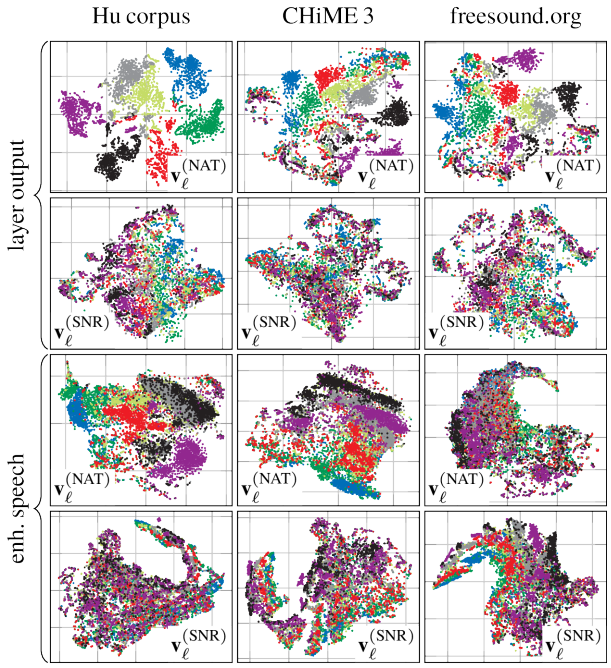

Fig. 3. Embedding of the internal representation (row 1 and 2) and the enhanced speech signal (row 3 and 4). The same four sentences and the same color code as in Fig 2 are used. The columns show the results for the training sets described in Section 3. Row 1 and 3 are based on the NAT features and the row 2 and 4 on the SNR-NAT features.

From these observations, we follow that using noise dependent input features such as the NAT features, leads to a noise dependent internal representation. But only a large and diverse training dataset allows the mapping from the noise dependent representation to the IRM to be learned appropriately. Using, however, the SNR-NAT features, this mapping is established more easily. As a consequence, the SNR-NAT feature are more robust to issues in the design of the training dataset or applications where only limited training data is available.

6. CONCLUSIONS

In this paper, we compared SNR-NAT and NAT features on various training datasets that differ in size and diversity. We show that a large and diverse training dataset is required to make the considered DNN-based enhancement scheme generalize to unseen noise types if the NAT features are employed. On the one hand, we found that the background noises included in the Hu noise corpus [19] or the CHiME 3 noise corpus [20] may not be sufficient to obtain a model that generalizes to unseen noise types. Only using a large and diverse dataset, which we construct from sounds of the freesound.org website, a model that generalizes well to unseen conditions is obtained. On the other hand, we found that using the SNR-NAT features more robust models are obtained even if the training data is limited in size or diversity. We analyzed embeddings of the feature data and the internal representations and showed that the feature data as well as the internal representation depend less on the noise type. This property appears to be the key for simplifying the mapping from the feature space to the masking function which has to be learned by the DNN.

7. 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] 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.

[3] 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.

[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] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1741–1752, Aug. 2007.

[7] 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.

[8] 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.

[9] S. R. Park and J. W. Lee, "A fully convolutional neural network for speech enhancement," in *Interspeech*, Stockholm, Sweden, Aug. 2017, pp. 1993–1997.

[10] K. Qian, Y. Zhang, S. Chang, X. Yang, D. Florêncio, and M. Hasegawa-Johnson, "Speech enhancement using Bayesian wavenet," in *Interspeech*, Stockholm, Sweden, Aug. 2017, pp. 2013–2017.

[11] 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, pp. 2670–2674.

[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] 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.

[15] 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.

[16] 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.

[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] R. Rehr and T. Gerkmann, "Robust DNN-based speech enhancement with limited training data," in *ITG Conference on Speech Communication*, Oldenburg, Germany, Oct. 2018.

[19] G. Hu, "A corpus of nonspeech sounds," http://web.cse.ohio-state.edu/pnl/corpus/HuNonspeech/HuCorpus.html, 2005.

[20] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Analysis and outcomes," *Computer Speech & Language*, vol. 46, pp. 605–626, Nov. 2017.

[21] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, Nov. 2008.

[22] "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.

[23] 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.

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

[26] T. Fingscheidt, S. Suhadi, and S. Stan, "Environment-optimized speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 825–834, May 2008.

[27] Y. Xu, J. Du, Z. Huang, D. Li-Rong, and C.-H. Lee, "Multi-Objective Learning and Mask-Based Post-Processing for Deep Neural Network Based Speech Enhancement," in *Interspeech*, Dresden, Germany, Sep. 2015.

[28] 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.

[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.