Speech Intelligibility Potential of General and Specialized Deep Neural Network Based Speech Enhancement Systems

Morten Kolbæk, Zheng-Hua Tan, *Senior Member, IEEE*, and Jesper Jensen

***Abstract*—In this paper, we study aspects of single microphone speech enhancement (SE) based on deep neural networks (DNNs). Specifically, we explore the generalizability capabilities of state-of-the-art DNN-based SE systems with respect to the background noise type, the gender of the target speaker, and the signal-to-noise ratio (SNR). Furthermore, we investigate how specialized DNN-based SE systems, which have been trained to be either noise type specific, speaker specific or SNR specific, perform relative to DNN based SE systems that have been trained to be noise type general, speaker general, and SNR general. Finally, we compare how a DNN-based SE system trained to be noise type general, speaker general, and SNR general performs relative to a state-of-the-art short-time spectral amplitude minimum mean square error (STSA-MMSE) based SE algorithm. We show that DNN-based SE systems, when trained specifically to handle certain speakers, noise types and SNRs, are capable of achieving large improvements in estimated speech quality (SQ) and speech intelligibility (SI), when tested in matched conditions. Furthermore, we show that improvements in estimated SQ and SI can be achieved by a DNN-based SE system when exposed to unseen speakers, genders and noise types, given a large number of speakers and noise types have been used in the training of the system. In addition, we show that a DNN-based SE system that has been trained using a large number of speakers and a wide range of noise types outperforms a state-of-the-art STSA-MMSE based SE method, when tested using a range of unseen speakers and noise types. Finally, a listening test using several DNN-based SE systems tested in unseen speaker conditions show that these systems can improve SI for some SNR and noise type configurations but degrade SI for others.**

***Index Terms*—Deep neural networks, generalizability, ideal ratio mask, intelligibility, speech enhancement.**

I. Introduction

IMPROVING quality and intelligibility of noisy speech signals is of great interest in a vast amount of applications such as mobile communications, speech recognition systems, and hearing aids. In a single-microphone setting, improving Speech Quality (SQ) and especially Speech Intelligibility (SI) is a challenging task and is an active topic of research [1]–[3]. Traditionally, single microphone Speech Enhancement (SE) has been addressed by statistical model based methods such as the Wiener filter [2] and Short-Time Spectral Amplitude Minimum Mean Square Error (STSA-MMSE) estimators, e.g., [4]–[6]. However, recent advances in Computational Auditory Scene Analysis (CASA) and machine learning have introduced new methods, e.g. Deep Neural Network (DNN), Gaussian Mixture Model (GMM), and Support Vector Machine (SVM) based methods, which address single-microphone SE and speech segregation in terms of advanced statistical estimators. These estimators aim at estimating either a clean speech Time-Frequency (T-F) representation directly or a T-F mask that is applied to the T-F representation of the noisy speech to arrive at an estimate of the clean speech signal [7]–[15]. For some potential future applications, e.g. DNN based SE algorithms for hearing aids or mobile communications, the range of possible acoustic situations which can realistically occur is virtually endless. It is therefore important to understand how such methods perform in different acoustic situations, and how they perform, when they are exposed to ”unseen” situations, i.e. acoustic scenarios not encountered during training. Despite the obvious importance of this generalizability question, it is currently not well understood.

In this study we focus on situations where a single target speaker is present in additive noise and the aim of the SE algorithm is to enhance the speech signal and attenuate the noise using a single-microphone recording. Generally, when evaluating generalizability of machine learning based SE algorithms, there are at least three dimensions in which the input signal can vary: i) the noise type dimension, ii) the speaker dimension and iii) the Signal to Noise Ratio (SNR) dimension. Therefore, evaluation of DNN based SE methods should cover each of these dimensions in a way similar to what is expected to be experienced in a real life scenario. For mobile communication devices and hearing aid systems, evaluation should hence encompass a wide range of noise types, a wide range of speakers and a wide range of SNRs, in order to give a realistic estimate of the expected performance of the algorithm in real life scenarios. On the other hand, for applications where the typical usage situation is much more well-defined, e.g. voice-controlled devices to be used in a car cabin situation, training and testing might involve only car cabin noises at a narrow SNR range for a single particular speaker.

The exploration of these three dimensions is motivated by the fact that no matter how many noise types, SNRs, and speakers

Manuscript received July 20, 2016; revised November 6, 2016; accepted November 6, 2016. Date of publication November 15, 2016; date of current version December 2, 2016. This work was supported in part by the Oticon Foundation. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Richard Christian Hendriks.

M. Kolbæk and Z.-H. Tan are with the Department of Electronic Systems, Aalborg University, Aalborg 9220, Denmark (e-mail: mok@es.aau.dk; zt@es.aau.dk).

J. Jensen is with Oticon A/S, Smørum 2765, Denmark, and also with the Department of Electronic Systems, Aalborg University, Aalborg 9220, Denmark (e-mail: jesj@oticon.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2016.2628641

a SE system is exposed to during training, in a real life scenario, sooner or later the system will be exposed to an unseen noise type, an unseen speaker or an unseen SNR. However, if the system is well trained, one might expect that the system has captured some general acoustic characteristics from these dimensions and hence generalizes well to unseen conditions. Furthermore, if any *a priori* knowledge about the noise type, speaker characteristics or SNR is available, it is important to know what performance gain can be achieved by utilizing this *a priori* knowledge.

Several studies have investigated aspects of generalizability of SE algorithms based on DNNs, SVMs, and GMMs, e.g. [7]–[24]. However, these models are fundamentally different in both training schemes and architectures [25] and since DNNs are currently state-of-the-art in a large number of applications [26] and have outperformed SVMs and GMMs in SE tasks [8]–[10], [12], SVMs and GMMs are less suited for the current study and are therefore not considered.

Common for all the studies, based on DNNs [7], [9], [12]–[14], [16]–[20], is that during training or testing one or more of the generalizability dimensions defined above are held fixed, while others are varied. To the authors knowledge no study exists which explores the simultaneous variation of all the three dimensions - a situation which is realistic for many real life applications. Furthermore, interpretation of existing studies is sometimes complicated by the fact that the training and test signals, for the dimensions which *are* varied, are not described in all details. For example, the distribution of males and females is often not reported [7], [16], [18], [19] and it is hence not clear if the system is mostly a gender specific or gender general system. Several studies [7], [16], [18] use the TIMIT corpus [27], which is approximately 70% male and 30% female. Furthermore, the duration of the different training noise types is typically not considered when the training data is constructed, hence the exact distribution of the noise types is unknown. For example, in [8], [16], [18], [19] noise sequences with highly varying duration are used, which makes it unclear to which extent these systems are noise specific or noise general. Another issue related to the noise dimension is concerning the construction of training and test data. In several studies [8], [10]–[12], [18], [19] , the exact same noise realizations were used for training and testing. In [28] this training/testing paradigm was analyzed, and it was found to erroneously give remarkably better performance compared to the realistic scenario, where the actual noise sequence is unknown. Furthermore, the systems presented in [7], [9], [12]–[14], [16]–[20] are based on various network architectures, training methods, testing methods, speech corpora, noise databases, feature representations, target representations etc. As a consequence of these differences, their results cannot be directly related and it is therefore unclear how a state-of-the-art DNN based SE algorithm perform when the generalizability dimensions mentioned above are considered simultaneously. Finally, it is unclear to what extent state-of-the-art DNN based SE algorithms provide improvements over existing non-DNN based SE methods. In [7], [16], [29] a DNN based SE method similar to the one studied here outperforms several different non-DNN based methods such as statistical MMSE based methods [6], [30]–[33] and non-Negative Matrix Factorization (NMF) methods [7], [34]. However, since the DNNs used in [7], [16], [29] have not been trained across all three generalizability dimensions, the comparison may not yield a true picture of the actual performance difference. This is particularly true with the statistical MMSE based methods [1], [2], which have not been trained to handle any specific noise types or speakers but merely rely on basic statistical assumptions with respect to Short-Time Fourier Transform (STFT) coefficients and might perform worse than a system trained on a given speaker or noise type.

The goal of this paper is therefore to conduct a systematic evaluation of the generalizability capabilities of a state-of-the-art DNN based SE algorithm in terms of estimated SQ and SI. Specifically, we investigate how a state-of-the-art DNN based SE method performs when it is trained to be noise type specific vs. noise type general, speaker specific vs. speaker general, and SNR specific vs. SNR general. Furthermore, we study the performance drop, if any, for systems which are specialized in one or more of the three generalizability dimensions, compared to a completely general DNN based SE system, which relies on essentially no prior knowledge with respect to speaker characteristics, noise type, and SNR. Additionally, it is investigated how this general system performs compared to a state-of-the-art non-machine learning based method namely the STSA-MMSE estimator employing generalized gamma priors as proposed in [6], [32], [33][1]. This is of interest since the STSA-MMSE method relies on very little prior knowledge compared to conventional DNN based SE methods [8], [9], [13]. Furthermore, given that the computational and memory complexity associated with DNN type of systems is typically orders of magnitude larger than that associated with simple STSA-MMSE based systems it is of obvious interest to understand the performance gain of this technology. Finally, a listening test is conducted, using both specialized and general DNN based SE systems, to investigate if such systems improve SI, when tested in different matched and unmatched conditions.

It is important to note that this paper emphasizes on the generalizability properties of DNN based SE algorithms in terms of estimated and measured SI, since these properties has not yet been rigorously investigated in the current literature [7], [9], [12]–[14], [16]–[20]. To do so, we rely on a specific implementation of a feed-forward DNN, whose architecture and training procedure resemble those of a large range of existing DNN based SE methods [7], [9], [16], [19], [35]. This allows us to expect that our findings are representative not only for our particular implementation but are generally valid for DNN based SE methods. The fact that the DNN based SE method under study is a representative member of a larger class of algorithms also implies that this particular implementation does not necessarily outperform all existing methods with respect to estimated SQ and SI.

Furthermore, obviously, the three chosen generalizability dimensions are not the only dimensions for which mixing scenarios can vary. Other such dimensions include reverberation conditions, e.g. in terms of varying room impulse responses, or

[1]http://insy.ewi.tudelft.nl/content/software-and-data

digital signal processing conditions, e.g. in terms of signal sampling rate, number of bits with which each sample is represented, microphone characteristics, compression/coding schemes, etc. Furthermore, for DNN based SE algorithms the DNN architecture can also be varied and considered as a dimension. We have chosen the speaker dimension, the noise type dimension and the SNR dimension for this particular work since these are dimensions most often encountered in the SE literature [7]–[24]. Furthermore, in most papers related to DNN based SE algorithms only a single speaker is considered, so it is of interest to study how well these algorithms generalize to unknown speakers. Finally, the performance of non-machine learning based SE algorithms such as STSA-MMSE and Wiener filtering based approaches are known to be highly dependent on the noise type, and SNR, but not the speaker. Hence, it is of interest to study how a DNN based SE algorithm performs in a large range of noise types, speakers and SNRs.

The paper is organized as follows: Section II describes the DNN architecture, training procedure and speech material used for conducting the desired experiments. Section III describes and discusses the experimental setups and results. Finally, in Section IV the findings are concluded.

II. Speech Enhancement Using Neural Networks

A. Speech Corpus and Noisy Mixtures

The phonetically balanced Danish speech corpus *Akustiske Databaser for Dansk* (ADFD)[2] is used as target speech material for training and testing all DNN based SE systems considered in this paper. This corpus consists of two sets: One set (set 1) consisting of 56 speakers with 986 spoken utterances for each speaker and another set (set 2) with 560 speakers and 311 spoken utterances, and males and females are approximately equally distributed among the two sets. The majority of the text material is based on conversational speech, but also single words, numbers and sequences of numbers are included, and each utterance has an average duration of approximately 5 seconds.

The training, validation, and test sets, were constructed such that no sentence appears more than once in the combined training, validation, and test set. The sampling frequency was 16 kHz and all files were normalized to have unit Root Mean Square (RMS) power.

The noisy mixtures for all experiments were constructed by adding a noise signal to a clean speech signal at a certain SNR. The noise signal was scaled to achieve the desired SNR based on the speech active region of the speech signal, i.e. the silence parts in the beginning and in the end of the speech signal were omitted in SNR computation. Omitting the silence parts for the SNR computation is crucial since the inclusion of these parts will effectively decrease the energy estimate of the clean speech, hence a lower noise power is required to achieve the same SNR, than if the silence regions were omitted. The difference in SNR between these two approaches of constructing noisy mixtures can be more than one dB and is typically not considered in the literature [9], [16], [18], even though it is of importance if results from different studies are to be related. Alternatively, a Voice Activity Detection (VAD) algorithm could have been used to exclude all silent regions, which would be highly beneficial for practical applications. However, for simplicity and to be in-line with existing literature [7]–[24], we excluded the VAD for all experiments. As before, the global SNR based approach were chosen from a practical perspective and to be in-line with existing literature [7]–[24], where global SNR is by far the most common.

B. Features and Labels

The choice of training targets for supervised speech enhancement have been widely studied [2], [7], [36]–[41]. Recent studies [7], [9], [38], [40] suggest that continuous targets such as the Ideal Ratio Mask (IRM) are preferable over binary targets such as the Ideal Binary Mask (IBM) [39], [40]. Therefore, the DNN studied in this paper is trained in a supervised fashion to estimate the IRM from a feature representation of a noisy speech signal.

The T-F representation used to construct the IRM is based on a gammatone filter bank with 64 filters linearly spaced on a MEL frequency scale from 50 Hz to 8 kHz and with a bandwidth equal to one Equivalent Rectangular Bandwidth (ERB) [42][3]. The output of the filter bank is divided into 20 ms frames with 10 ms overlap and with a sampling frequency of 16 kHz, each T-F unit represents a vector of 320 samples.

Let $\mathbf{x}(n,\omega)$ denote the Time-Frequency (T-F) unit of the clean speech signal at frame n and frequency channel ω, and let $\mathbf{d}(n,\omega)$ denote the corresponding T-F unit of the noise signal. Then the IRM is computed as [7]

$$\mathrm{IRM}(n,\omega) = \left(\frac{||\mathbf{x}(n,\omega)||^2}{||\mathbf{x}(n,\omega)||^2 + ||\mathbf{d}(n,\omega)||^2}\right)^{\beta}$$

where $||\mathbf{x}(n,\omega)||^2$ is the squared 2-norm, i.e. the clean speech energy, of T-F unit n in frequency channel ω. Likewise, $||\mathbf{d}(n,\omega)||^2$ is the noise energy of a T-F unit n in frequency channel ω. The variable β is a tunable parameter and has for all experiments in this paper been set to $\beta = 0.5$, which in [7] was found empirically to provide good results.

To have discriminative and noise robust features, each frame is transformed into a 1845-dimensional feature vector inspired by [3], [8], [9], [12], [43]–[45]. Although, very recent works use only magnitude spectra [13], [20], [46] a large context of several hundred milliseconds is used, which is undesirable for real time applications. The chosen feature vector was found to outperform features of magnitude spectra when these were based on only a small context. The features used are 31 Mel Frequency Cepstrum Coefficients (MFCC), 15 Amplitude Modulation Spectrogram (AMS), 13 Relative Spectral Transform - Perceptual Linear Prediction (RASTA-PLP) and 64 Gammatone Filter bank Energies (GFE). Furthermore delta and double delta features are computed and a context of 2 past and 2 future frames are utilized, hence arriving at a 1845-dimensional feature vector. All feature vectors are normalized to zero mean and unit variance.

[2]http://www.nb.no/sbfil/talegjenkjenning/

[3]http://web.cse.ohio-state.edu/pnl/shareware/cochleagram

C. Network Architecture and Training

The DNNs used in this paper follow a feed-forward architecture with a 1845-dimensional input layer and three hidden layers, each with 1024 hidden units, and 64 output units [7], [9]. The activation functions for the hidden units are Rectified Linear Units (ReLUs) [47] and for the output units the sigmoid function is applied. The hidden layers are initialized using the "GlorotUniform" approach [48]. Furthermore, the DNN has approximately 4M tunable parameters in terms of weights and biases. The values of the parameters are found using Stochastic Gradient Descent (SGD) following the AdaGrad approach [49]. The gradients are computed using backpropagation based on the Mean Square Error (MSE) error function using a batch size of 1024 [25]. Furthermore, 20% dropout has been applied to all hidden layers during training to reduce overfitting [50]. In order to further reduce overfitting, an early-stopping training scheme is applied, which stops the training, when the MSE of the validation set has not decreased with more than 1% for more than 20 epochs. Although used in [8], [16], unsupervised DNN pretraining using Deep Belief Networks [26], [51], [52] was found not to significantly improve performance and has therefore not been applied in the reported results.

Finally, it is well known that increasing the network size or changing the network architecture can improve performance of DNN based algorithms [13], [16], [20], [46], [53], [54]. However, it is not practically feasible to include network architecture as a dimension in our experiments. Furthermore, although absolute performance might be better with a different architecture, the conclusions drawn from using a fixed-sized feed-forward DNN are expected to be valid for a broader range of DNN architectures, since the underlying assumptions are practically the same.

D. Signal Enhancement

After DNN training, the IRM is estimated for a given test signal by forward propagating its feature representation, for all frames, through the DNN. The output of the DNN represents the estimated IRM, $\widehat{\mathrm{IRM}}(n,\omega)$ for the given frame. The estimated IRM can then be applied to the T-F representation of the noisy speech signal by multiplying the given entry of the mask to all 320 noisy signal samples of a T-F unit. All T-F units in each frequency channel are then concatenated and all overlapping parts are summed. Afterwards, the 64 frequency channels can be synthesized into a time domain signal by first compensating for the different group delays in the different channels and then adding the frequency channels. The group delay compensation is performed by time reversing the signals, passing them through the gammatone filter bank and then time reversing the signals once again [42].

E. Evaluation of Enhancement Performance

Speech signals enhanced with the DNN based SE algorithm studied in this paper were evaluated using the Short-Time Objective Intelligibility (STOI) [55] measure and the wideband extension of the Perceptual Evaluation of Speech Quality (PESQ) measure [56], [57]. The STOI measure estimates SI and PESQ estimates SQ and both have been found to be highly correlated with human listening tests [2], [55]. STOI is defined in the range $[-1, 1]$ and PESQ is defined in the range $[1, 4.5]$ and for both measures higher is better. We used the implementations of STOI and PESQ available from [55] and [2], respectively.

Although other performance measures exists such as Signal to Distortion Ratio (SDR), Signal to Interferences Ratio (SIR), and Signal to Artifact Ratio (SAR) [2], [58] we report only PESQ and STOI to limit the number of tables. Furthermore, PESQ and STOI are by far the most used speech quality and speech intelligibility estimators in the literature [7]–[24].

III. Experimental Results and Discussion

To investigate the generalizability capability of DNN based SE systems with relation to: i) the noise type dimension, ii) the speaker dimension and iii) the SNR dimension, three experimental setups, one for each dimension, have been designed. When a dimension is explored the remaining two dimensions are held fixed. For example, when exploring the SNR dimension, the SNR dimension is varied but only a single speaker and a single noise type is used for both training and testing. Furthermore, a fourth setup has been constructed where a general system has been designed. This system was trained using a wide range of speakers, noise types and SNRs, hence the system relies on a minimum of *a priori* knowledge. This "general" system is compared against the three experiments previously described, as well as a state-of-the-art non-machine learning based SE algorithm.

A. SNR Dimension

The purpose of the SNR experiments is to investigate the impact on the performance of DNN based SE systems, when training is performed based on a single SNR vs. a wide range of SNRs, i.e. constructing a SNR specific or a SNR general system. The SNR dimension is explored using speech material based on 986 spoken utterances from a single female speaker from the ADFD set 1. These 986 utterances were divided such that 686 were used for training, 100 for validation and 200 for testing. Two noise types have been investigated, a stationary Speech Shaped Noise (SSN) and a non-stationary Babble (BBL) noise. The SSN sequence is constructed by filtering a 50 min. Gaussian white noise sequence through a 12th-order all-pole filter with coefficients found from Linear Predictive Coding (LPC) analysis of 100 randomly chosen TIMIT sentences [27]. The BBL noise is also based on TIMIT. The corpus, which consists of a total of 6300 spoken sentences, is randomly divided into 6 groups of 1050 concatenated utterances. Each group is then truncated to equal length followed by addition of the six groups. This results in a BBL noise sequence with a duration of over 50 min. The SSN and BBL sequences were both divided such that 40 min. were used for training, 5 min. were used for validation and 5 min. for testing, hence there is no overlapping samples in the noise segments used for training, validation and test. To investigate how the performance of DNN based SE systems depends on the SNR dimension, eight systems were trained with eight different SNR settings for both SSN and BBL noise. All 16 systems were tested using eight SNRs ranging from -15 dB to 20 dB with

TABLE I
STOI IMPROVEMENT FOR THE SNR DIMENSION. EIGHT DNN BASED SE SYSTEMS TRAINED ON DIFFERENT SNR RANGES AS INDICATED IN THE FIRST ROW. THE NOISE TYPE DIMENSION IS HELD CONSTANT USING SSN ONLY AND THE SPEAKER DIMENSION IS HELD CONSTANT USING A SINGLE FEMALE SPEAKER. THE SYSTEMS ARE EVALUATED USING STOI FOR TEST SIGNALS WITH 8 DIFFERENT SNRS RANGING FROM −15DB TO 20 DB. THE SECOND COLUMN PRESENTS THE STOI SCORE FOR THE UNPROCESSED NOISY CMIXTURES. COLUMNS 3−10 PRESENT STOI IMPROVEMENTS

	Noisy	−5 dB	−5 dB – 0 dB	−5 dB – 5 dB	−10 dB – 5 dB	−15 dB – 5 dB	−15 dB – 10 dB	−15 dB – 15 dB	−15 dB – 20 dB
−15 dB	0.354	0.016	0.019	0.028	0.063	0.074	0.075	0.075	0.072
−10 dB	0.417	0.170	0.166	0.165	0.186	0.186	0.185	0.183	0.179
−5 dB	0.519	**0.219**	**0.218**	**0.218**	**0.219**	**0.216**	**0.215**	**0.213**	**0.210**
0 dB	0.642	0.180	0.186	0.187	0.185	0.183	0.183	0.182	0.181
5 dB	0.756	0.115	0.125	0.130	0.128	0.126	0.128	0.127	0.127
10 dB	0.844	0.058	0.070	0.078	0.077	0.076	0.079	0.079	0.078
15 dB	0.905	0.016	0.030	0.040	0.039	0.039	0.044	0.045	0.044
20 dB	0.944	−0.010	0.005	0.015	0.014	0.014	0.020	0.023	0.023

TABLE II
AS TABLE I BUT FOR PESQ

	Noisy	−5 dB	−5 dB – 0 dB	−5 dB – 5 dB	−10 dB – 5 dB	−15 dB – 5 dB	−15 dB – 10 dB	−15 dB – 15 dB	−15 dB – 20 dB
−15 dB	1.133	−0.044	−0.041	−0.044	−0.036	−0.027	−0.029	−0.035	−0.032
−10 dB	1.115	0.025	0.024	0.025	0.038	0.044	0.042	0.041	0.041
−5 dB	1.115	0.198	0.190	0.192	0.202	0.196	0.196	0.191	0.187
0 dB	1.144	0.457	0.425	0.421	0.410	0.400	0.410	0.408	0.405
5 dB	1.234	0.700	0.691	0.655	0.643	0.638	0.630	0.636	0.642
10 dB	1.438	**0.769**	**0.875**	0.879	0.863	0.859	0.831	0.803	0.811
15 dB	1.811	0.583	0.830	**0.942**	**0.925**	**0.911**	**0.948**	**0.902**	**0.878**
20 dB	2.346	0.130	0.518	0.764	0.745	0.733	0.860	0.877	0.848

TABLE III
AS TABLE I BUT FOR BBL

	Noisy	−5 dB	−5 dB – 0 dB	−5 dB – 5 dB	−10 dB – 5 dB	−15 dB – 5 dB	−15 dB – 10 dB	−15 dB – 15 dB	−15 dB – 20 dB
−15 dB	0.292	0.048	0.034	0.033	0.070	0.093	0.095	0.094	0.096
−10 dB	0.369	0.161	0.150	0.146	0.170	0.173	0.173	0.174	0.174
−5 dB	0.480	**0.214**	**0.218**	**0.216**	**0.214**	**0.205**	**0.205**	**0.206**	**0.206**
0 dB	0.608	0.187	0.200	0.202	0.194	0.188	0.189	0.191	0.191
5 dB	0.728	0.128	0.147	0.152	0.146	0.141	0.144	0.147	0.146
10 dB	0.823	0.070	0.091	0.098	0.095	0.091	0.096	0.098	0.097
15 dB	0.890	0.024	0.045	0.056	0.053	0.050	0.057	0.059	0.059
20 dB	0.934	−0.008	0.013	0.026	0.023	0.021	0.029	0.032	0.033

TABLE IV
AS TABLE I BUT FOR PESQ AND BBL

	Noisy	−5 dB	−5 dB – 0 dB	−5 dB – 5 dB	−10 dB – 5 dB	−15 dB – 5 dB	−15 dB – 10 dB	−15 dB – 15 dB	−15 dB – 20 dB
−15 dB	1.201	−0.063	−0.066	−0.058	−0.070	−0.080	−0.066	−0.079	−0.075
−10 dB	1.180	−0.047	−0.060	−0.058	−0.056	−0.052	−0.055	−0.055	−0.054
−5 dB	1.143	0.086	0.090	0.089	0.081	0.074	0.072	0.079	0.075
0 dB	1.162	0.289	0.312	0.319	0.294	0.280	0.279	0.284	0.293
5 dB	1.270	0.493	0.571	0.580	0.543	0.511	0.516	0.527	0.538
10 dB	1.478	**0.636**	0.772	0.805	0.770	0.732	0.740	0.745	0.741
15 dB	1.829	0.621	**0.826**	**0.914**	**0.884**	**0.844**	**0.872**	**0.872**	**0.854**
20 dB	2.312	0.426	0.691	0.863	0.835	0.794	0.863	0.871	0.851

steps of 5 dB. For each noise source, the first system was trained using −5 dB since this is a commonly encountered SNR in the literature [8], [9] where SI is typically degraded and DNN based SE algorithms have been successfully applied [8], [9]. The next system was trained using SNRs from −5 dB to 0 dB with steps of 1 dB. In a similar fashion wider and wider SNR ranges were used for training the remaining systems with the widest range being from −15 dB to 20 dB. The precise intervals are given in Tables I, II, III and IV. For all systems, each training utterance was mixed with different noise realizations 35 times in order to increase the amount of training data. For each noisy mixture, the SNR was drawn from a discrete uniform distribution de-

fined within the given SNR range. Due to the large number of realizations, it is assumed that the distribution of drawn SNRs is approximately uniform. The noise signal used for each noisy mixture was extracted from the whole training noise sequence by using a starting index drawn from a discrete uniform distribution defined over the entire length of the noise sequence. If the starting index is such that there is no room for the whole utterance, the remaining samples are extracted from the beginning of the noise sequence. Following the same procedure, each validation utterance is mixed with different noise realizations 10 times. Using this form of training data augmentation, the total amount of training utterances, used for training each system, is increased to $686 \times 35 = 24010$, which is approximately equal to 33 hours of speech material and is approximately 65% more data than used by [9].

The results of the SNR dimension experiments are presented in Tables I and II for SSN and in Tables III and IV for BBL noise. From Tables I and III it is seen that the SNR specific system of SNR of -5 dB achieves relatively large STOI improvements, for test signal SNRs in the range -10 dB to 5 dB. In general, it can be observed that inclusion of SNRs in the range from -15 dB to 5 dB has a larger positive impact on the performance than inclusion of SNRs above 5 dB. This might be explained partly by the fact that intelligibility is almost 100% (STOI ≈ 1) for test signal SNRs above 5 dB, and partly by the limited noise energy, which makes it more difficult for the DNN to actually learn important noise characteristics. Tables II and IV show a somewhat similar picture. The inclusion of training signals with SNRs around 0 dB in general improves performance, but extending the training SNR range from 5 dB to 20 dB does not further improve performance. Furthermore, it is also seen that the system in general cannot improve PESQ for test signals with SNRs below -5 dB.

Based on these experiments it can be concluded that there is generally a good correspondence between SNR ranges used in training and STOI improvement seen during testing. For example, the systems trained in the SNR range from -5 dB to 0 dB perform better at 0 dB than the systems trained using only -5 dB. Furthermore, even at -15 dB, where the noise energy is approximately 40 times larger than the speech energy, STOI is still improved with 0.074 and 0.093 for SSN and BBL noise, respectively, when this particular SNR is included in the training set, and the improvement is almost constant for SSN, and even slightly increasing for BBL noise, when a wider range of positive SNRs are included in the training set. Also, the system trained using the widest SNR range from -15 dB to 20 dB achieves almost similar performance as the -5 dB SNR specialized system, when tested at an SNR of -5 dB, and generally performs better at other SNRs. This observation is in line with related studies [59] and is of large practical importance, as it suggests that DNN based SE systems should simply be trained using as large a training signal SNR range as practically possible.

B. Noise Dimension

The purpose of the noise dimension experiments is to investigate the performance impact when DNN based SE systems are trained on a single noise type vs. a wide range of noise types. In other words, this allows us to compare a noise specific vs. a noise general system. The noise dimension has been explored using the same 986 spoken utterances from the same single female speaker as used in the SNR experiments. Likewise, the partition of the speech material into training, validation and test set is also identical to the SNR experiments. To explore the noise dimension, six distinct noise types were used: SSN (N1) and BBL (N2) from the SNR experiments and four additional noises: street (N3), pedestrian (N4), cafe (N5) and bus (N6), from the CHiME3 dataset[60]. Furthermore, 1260 randomly selected sound effect noises from soundbible.com[4] were used to construct a seventh noise type referred to as the *mix* (N7) noise type. These 1260 noises were first truncated to have a maximum duration of 3 seconds each and then concatenated into one large noise sequence. The sound effects include sounds from animals, singing humans, explosions, airplanes, slamming doors etc. All seven (N1 – N7) noise types used for the noise experiments were first truncated to have a total duration of 50 min. and then divided into a 40 min. training set, a 5 min. validation set and a 5 min. test set, hence there is no overlapping samples in the noise segments used for training, validation and test.

To investigate how the performance of the DNN based SE system depends on the noise dimension, eight systems were trained with eight different noise combinations all at an SNR of -5 dB. Two systems were trained with only one noise type, namely the stationary SSN (N1) and non-stationary BBL (N2). The remaining six systems were trained with an increasing number of noise types starting with N1 – N2 and ending with N1 – N7 as indicated in the second row in Tables V and VI. When noise types were combined, the 40 min. noise sequences were concatenated and similar to the SNR experiment, a noise sequence was extracted based on a randomly chosen starting index within this concatenated noise sequence. Similarly to the SNR experiment, each utterance in the training set was mixed with a randomly chosen noise sequence 35 times, hence a total of $686 \times 35 = 24010$ noise mixtures were constructed. The large number of mixtures and the identical duration of the noise sequences ensures that the noise distribution within the training data is approximately uniform, hence a noise-general system is constructed. All eight systems have been tested with speech signals contaminated by all seven noise types, which ensures that all but the system trained with all seven noises will be tested with at least one unseen and at the most 6 unseen noise types.

The results are presented in Tables V and VI where the first column represents the noise types used for testing and the second row represents the noises used for training. Table V shows that when a system is trained using SSN only (N1) it achieves a relatively large STOI improvement of 0.22, when tested on that particular noise type, but generalizes poorly on the majority of the unseen test noises. Similarly, when a system is trained on BBL (N2), the performance is good in the matched noise case, but the system generalizes poorly to other noise types. Furthermore, when both SSN and BBL noise types are included equally in the training set (N1-N2), the system performs almost as good

[4]http://soundbible.com/free-sound-effects-1.html

TABLE V
STOI IMPROVEMENT FOR THE NOISE TYPE DIMENSION. EIGHT DNN BASED SE SYSTEMS HAVE BEEN TRAINED WITH DIFFERENT COMBINATIONS OF SEVEN DIFFERENT NOISE TYPES (N1-N7) AS GIVEN BY THE FIRST ROW. THE SNR DIMENSION IS HELD CONSTANT AT −5 DB AND THE SPEAKER DIMENSION IS HELD CONSTANT USING A SINGLE FEMALE SPEAKER. THE SYSTEMS HAVE BEEN EVALUATED USING STOI AND TEST SIGNALS CORRUPTED BY ALL SEVEN NOISE TYPES. THE SECOND COLUMN PRESENTS THE STOI SCORE FOR THE NOISY UNPROCESSED MIXTURES. COLUMNS 3-10 PRESENT STOI IMPROVEMENTS

	Noisy	N1	N2	N1–N2	N1-N3	N1–N4	N1–N5	N1–N6	N1–N7
N1: ssn	0.519	**0.220**	0.083	0.207	0.209	**0.208**	**0.206**	**0.206**	**0.203**
N2: bbl	0.482	0.029	**0.217**	**0.210**	**0.211**	0.204	0.202	0.203	0.199
N3: str	0.590	0.122	−0.079	0.080	0.174	0.172	0.172	0.173	0.171
N4: ped	0.504	0.095	−0.008	0.078	0.139	0.157	0.161	0.160	0.158
N5: caf	0.572	0.072	−0.007	0.065	0.143	0.155	0.165	0.167	0.165
N6: bus	0.703	0.071	−0.058	0.003	0.112	0.114	0.118	0.130	0.128
N7: mix	0.685	0.015	0.028	0.038	0.072	0.078	0.092	0.093	0.119

TABLE VI
AS TABLE V BUT FOR PESQ

	Noisy	N1	N2	N1–N2	N1-N3	N1–N4	N1–N5	N1–N6	N1–N7
N1: ssn	1.112	**0.197**	−0.012	**0.175**	0.186	0.174	0.175	0.178	0.173
N2: bbl	1.174	−0.072	0.060	0.048	0.054	0.047	0.039	0.046	0.032
N3: str	1.069	0.114	−0.002	0.071	0.302	0.294	0.294	0.294	0.298
N4: ped	1.099	0.033	−0.025	0.005	0.095	0.118	0.125	0.125	0.120
N5: caf	1.081	0.030	−0.003	0.025	0.191	0.224	0.237	0.247	0.242
N6: bus	1.083	0.125	0.010	0.036	**0.329**	**0.336**	**0.351**	**0.421**	**0.415**
N7: mix	1.143	0.002	**0.067**	0.059	0.126	0.144	0.161	0.180	0.293

as the individual noise specific systems. However, the system does not generalize as well to the unseen noises as N1 did alone, except for the *mix* noise type, that similarly to BBL is highly non-stationary. It is also interesting to notice that the SSN and BBL specific systems achieve very similar performance for test signals contaminated by SSN and BBL, respectively. This is in contrast to STFT-based methods for which non-stationary noise is much more challenging [1]. A different picture is seen when a third noise (N3) is added in the training set (N1-N3). This system performs similarly well in the matched noise type setting, but also for the unseen noises the performance has increased considerably. Similar behavior is seen when the remaining noise types are included in the training set. Furthermore, even though new noise types are included in the training set, the performance of the system is almost constant in the matched noise type setting. One can argue that *str*, *ped*, *caf* and *bus* are quite similar noise types, but it is seen that the system trained with signals contaminated by all but the mix noise type (N1-N6) generalizes relatively well to the mix noise type, which is a noise type radically different from the others. From Table VI a similar behavior is observed where relatively large PESQ scores are achieved for all testing noises, already after noise type N1 – N3 have been included in the training set. Similar for both Tables V and VI is that there is generally a good correspondence between noise types used for training and STOI and PESQ improvements seen during testing. For example, the systems performing best on SSN and BBL noise are the systems that have been trained on only these noise types. However, a system trained on both noise types show only a slightly decrease in performance. Furthermore, the noise general system (N1 – N7), where all seven noise types are used for training, achieves on average the best performance across all seven noise types, while still being comparable in performance to the more specialized systems where only a single or a few noise types have been used for training. This is similar to the SNR experiments where no particular degradation in performance was observed by extending the SNR range used for training.

C. Speaker Dimension

The purpose of the speaker dimension experiments is to study the impact of using a single speaker vs. a wide range of speakers in the training material, i.e. constructing a speaker specific or speaker general system. The speaker dimension is explored using speech material based on 311 spoken utterances from 41 male and 41 female speakers from the ADFD set 2. The utterances from the 41 females are referenced as F-ID1 – F-ID41 and similarly, the utterances from the 41 males are referenced as M-ID1 – M-ID41. For each of the speakers of interest, (F/M-ID1 – F/M-ID21) 231 utterances were used for training, 30 for validation and 50 for testing. Furthermore, 50 utterances from each of the 40 remaining speakers (F/M-ID22 – F/M-ID41) were used as testing material for unseen speaker testing. The text material used for the 50 test utterances from each speaker was the same for all 82 speakers used for these experiments. A total of 10 systems were trained. Five systems using speech material corrupted with SSN at an SNR of -5 dB and five systems with speech material corrupted with BBL noise at an SNR of -5 dB. For each noise type, speakers F-ID1 and MID-1 were used to train two individual speaker specific systems. Furthermore, speakers F-ID2 – 21 and M-ID2 – 21 were used to train two individual gender specific systems and finally the speakers F-ID2 – 21 and

TABLE VII
TRAINING AND VALIDATION DATA AUGMENTATION SCHEME USED FOR RESULTS REPORTED IN SUBSECTION III-C TO ENSURE ALL SYSTEMS USE THE SAME AMOUNT OF DATA. THE FORMAT IS THE FOLLOWING: $\#speakers \times \#utterances \times \#repetitions = \#mixtures$

System	#Training Utterances	#Validation Utterances
Speaker Specific	$1 \times 231 \times 80 = 18480$	$1 \times 30 \times 40 = 1200$
Gender Specific	$20 \times 231 \times 4 = 18480$	$20 \times 30 \times 2 = 1200$
Speaker General	$40 \times 231 \times 2 = 18480$	$40 \times 30 \times 1 = 1200$

TABLE VIII
STOI IMPROVEMENT FOR THE SPEAKER DIMENSION. FIVE DNN BASED SE SYSTEMS HAVE BEEN TRAINED ON A VARIYING NUMBER OF SPEAKERS OF BOTH GENDERS AS GIVEN BY THE FIRST ROW. THE SYSTEMS HAVE BEEN TESTED IN BOTH SPEAKER MATCHED AND UNMATCHED CONDITIONS. THE NOISE TYPE DIMENSION IS HELD CONSTANT USING SSN FOR TRAINING AND TESTING AND THE SNR DIMENSION IS HELD CONSTANT USING AN SNR OF -5 DB FOR TRAINING AND TESTING. THE SYSTEMS HAVE BEEN EVALUATED USING STOI. THE SECOND COLUMN PRESENTS THE STOI SCORE FOR THE NOISY UNPROCESSED MIXTURES. COLUMNS 3-7 PRESENT STOI IMPROVEMENTS

Test\Train	Noisy	F-ID1	M-ID1	F-ID2 – 21	M-ID2 – 21	F/M-ID2 – 21
F-ID1	0.564	**0.168**	–	–	–	–
M-ID1	0.460	–	**0.204**	–	–	–
F-ID2-21	0.532	–	–	**0.175**	–	**0.170**
F-ID22-41	0.530	0.127	0.062	0.170	0.119	0.166
M-ID2-21	0.543	–	–	–	**0.174**	0.167
M-ID22-41	0.538	0.067	0.114	0.124	0.160	0.163
F/M-ID2–21	0.538	–	–	–	–	0.167
F/M-ID22–41	0.535	0.097	0.089	0.147	0.140	0.164

TABLE IX
AS TABLE VIII BUT FOR PESQ

Test\Train	Noisy	F-ID1	M-ID1	F-ID2 – 21	M-ID2 – 21	F/M-ID2 – 21
F-ID1	1.062	**0.160**	–	–	–	–
M-ID1	1.078	–	**0.185**	–	–	–
F-ID2-21	1.068	–	–	**0.168**	–	0.158
F-ID22-41	1.065	0.108	0.043	0.160	0.058	0.150
M-ID2-21	1.110	–	–	–	**0.219**	0.208
M-ID22-41	1.118	0.070	0.149	0.141	0.199	**0.213**
F/M-ID2–21	1.096	–	–	–	–	0.175
F/M-ID22–41	1.093	0.087	0.095	0.149	0.126	0.180

TABLE X
AS TABLE VIII BUT FOR BBL

Test\Train	Noisy	F-ID1	M-ID1	F-ID2 – 21	M-ID2 – 21	F/M-ID2 – 21
F-ID1	0.535	**0.131**	–	–	–	–
M-ID1	0.433	–	**0.184**	–	–	–
F-ID2-21	0.498	–	–	**0.121**	–	0.110
F-ID22-41	0.496	0.046	−0.107	0.117	−0.059	0.108
M-ID2-21	0.511	–	–	–	**0.140**	0.112
M-ID22-41	0.507	−0.093	0.039	−0.007	0.125	**0.115**
F/M-ID2–21	0.505	–	–	–	–	0.110
F/M-ID22–41	0.501	−0.025	−0.034	0.054	0.032	0.111

TABLE XI
AS TABLE VIII BUT FOR PESQ AND BBL

Test \ Train	Noisy	F-ID1	M-ID1	F-ID2 – 21	M-ID2 – 21	F/M-ID2 – 21
F-ID1	1.094	**0.065**	–	–	–	–
M-ID1	1.159	–	**0.029**	–	–	–
F-ID2-21	1.129	–	–	**0.001**	–	−0.007
F-ID22-41	1.130	−0.029	−0.049	−0.007	−0.064	−0.017
M-ID2-21	1.168	–	–	–	**0.024**	−0.005
M-ID22-41	1.181	−0.074	−0.040	−0.077	0.002	−0.007
F/M-ID2–21	1.141	–	–	–	–	**0.001**
F/M-ID22–41	1.164	−0.059	−0.051	−0.050	−0.037	−0.019

M-ID2 – 21 were combined (F/M-ID2 – 21) and used to train a single speaker general system. All systems were evaluated in both a seen speaker and an unseen speaker scenario using the test material from speaker F/M-ID1 – F/M-ID21 and F/M-ID22 – F/M-ID41, respectively. However, the systems trained using only one speaker is tested using 20 speakers, instead of only one speaker, to give more realistic unseen-speaker results. Since the number of distinct utterances used for training vary between the different systems, due to the varying number of speakers, a fixed total number of 18480 training utterances were used for training all systems. This is done to ensure that all systems are presented to the same amount of noise material. Using the same argument a total number of 1200 utterances were used for validation during the training of all systems. To achieve 18480 training mixtures, and 1200 validation mixtures, for each system, each distinct utterance was mixed with unique noise realizations multiple times as given by Table VII.

The results with SSN are presented in Tables VIII and IX, and the results with BBL noise are presented in Tables X and XI. The first column presents the speaker IDs used for testing and the second row represents speaker IDs used for training. From Table VIII it is seen that speaker specific systems trained on a single speaker achieves a STOI improvement of 0.168 and 0.204 for same-gender-same-speaker testing, for the female (F-ID1) and male (M-ID1) specific systems, respectively. However, if these systems are tested with new speakers of same gender, i.e. same-gender-new-speaker testing, the STOI improvements are reduced to 0.127 and 0.114 for the female (F-ID1) and male (M-ID1) specific systems, respectively. Furthermore, if the systems are tested on opposite gender the STOI improvement decreases to 0.067 and 0.062 for the female (F-ID1) and male (M-ID1) specific systems, respectively. Similar behavior, but with larger variations, is seen from Table X where the systems have been trained using utterances corrupted with BBL noise instead of SSN. Table X shows that systems trained using F-ID1 and M-ID2 improve STOI with 0.131 and 0.184 for same-gender-same-speaker testing, for the female (F-ID1) and male (M-ID1) systems, respectively. However, these improvements reduce to 0.046 and 0.039 for same-gender-new-speakers testing, and to -0.093 and -0.107 for new-gender testing, for the female (F-ID1) and male (M-ID1) systems, respectively. From these results it can be concluded that systems which are trained using only a single speaker generalizes very well to unseen utterances from the same speaker but not as good to unseen utterances from new speakers of same gender and even worse to opposite gender. Especially for BBL noise, the systems even

degrade the signals when evaluated using opposite gender. If gender specific systems are trained with 20 speakers instead of only a single speaker it is seen from Table VIII that the STOI improvements in the same-gender-same-speakers testing case are 0.175 and 0.174 for the female (F-ID2 – F-ID21) and male (M-ID2 – M-ID21) systems, respectively. Furthermore, the STOI improvements in the same-gender-new-speakers testing case are 0.170 and 0.160 for the female (F-ID2 – F-ID21) and male (M-ID2 – M-ID21) systems, respectively. Compared to the systems trained using a single speaker, the systems trained using 20 speakers of same gender generalize considerably better to the same-gender-new-speaker testing case. Also in the new-gender testing case Table VIII shows STOI improvements of 0.124 and 0.119 for the female (F-ID2 – F-ID21) and male (M-ID2 – M-ID21) systems, respectively. However, Table X shows that STOI is degraded when the female (F-ID2 – F-ID21) and male (M-ID2 – M-ID21) systems are tested in the new-gender testing case. Finally, if a speaker general system is trained using both males and females (F/M-ID2 – 21) and is tested in an unseen speaker setting based on both genders (F/M-ID22 – 41) using respectively SSN and BBL noise, the STOI improvements are 0.164 and 0.111, respectively. This shows that the speaker general system in terms of STOI generalizes considerably better than the speaker specific and gender specific systems to unseen speakers of both genders, for both a stationary and non-stationary noise type. Importantly, it is seen that the loss from a gender specific system to a gender general system is almost zero.

One interesting observation is the decrease in performance, when compared to the experiments exploring the noise type dimension in subsection III-B. For example, Table V shows that a system specialized to a single female speaker using BBL noise at an SNR of -5 dB achieves a STOI improvement of 0.217. Table X shows that a similar system (F-ID1) trained with a different female speaker using BBL noise at an SNR of -5 dB achieves a STOI improvement of 0.131, which is a considerable difference. There is one major difference between these two systems. For the experiments used to produce Table V, the speaker was represented by 686 distinct spoken utterances, whereas for the experiments used to produce Table X only 231 distinct spoken utterances were used. This indicates that not only the number of speakers but also the variability in speech material from each speaker is crucial to achieve good generalizability.

In general, it can be observed that a DNN based SE system trained using a single speaker becomes speaker specific and performs well, in terms of estimated SI, when evaluated using the same speaker. If a large number of speakers, of the same gender, are used for training, the system becomes gender specific and generalizes well to unseen speakers of same gender. Furthermore, if a large number of male and female speakers are used for training, the system becomes speaker general and generalizes well to unseen speakers of both genders. This applies for systems trained using training signals corrupted with either SSN or BBL noise. In terms of estimated SQ a similar behaviour can only be observed for systems trained with training signals corrupted with SSN whereas for the systems trained using training signals corrupted with BBL noise no, or only minor, improvements were found as shown by Tables IX and XI.

D. Combined Dimensions

The purpose of the combined dimension experiments is twofold. First, we wish to determine the performance decrease, if any, of a general DNN based SE system vs. the specialized systems considered in the three previous subsections, where only one dimension was varied at a time. Such experiments can be used to relate results previously reported in the literature, where at least one dimension has been fixed, to the more general case where all three dimensions are varied. Secondly, we wish to investigate how such a general DNN based SE method performs relative to a state-of-the-art non-machine learning based method, namely the STSA-MMSE method proposed in [6]. This is done in an attempt to give a realistic picture of the performance difference between these two classes of algorithms, which utilize different kinds of prior knowledge.

Alternatively, we could have compared the performance with a NMF based SE approach, which is another popular SE algorithm. However, several studies [3], [61]–[63] show that DNN based SE algorithms outperform NMF based approaches on several tasks. Furthermore, the NMF based approach can be viewed as a single hidden layer DNN. Hence, comparing the performance of the DNN based SE algorithm investigated in this paper to a NMF based SE algorithm is less interesting than a comparison with the STSA-MMSE based SE approach, which is from a completely different class of algorithms.

STSA-MMSE type of methods such as [6], [64] are very general and make only few assumptions about the target and noise signals and are therefore often used in practice [1]. Furthermore, the performance of these simple non-machine learning based algorithms in terms of speech intelligibility improvements are well studied in the literature, e.g. [37], [40], [65], [66]. Although deep neural network based speech enhancement algorithms have shown impressive performance, they are often trained and tested in narrow settings using either a few noise types [9], [43] or a single speaker [13]. It is therefore of interest to identify if/when a deep neural network based speech enhancement algorithm can outperform a non-machine learning based method, when approximately the same type of general a priori information is utilized: given that the computational and memory complexity associated with deep neural network type of systems is typically orders of magnitude larger than that associated with simple STSA-MMSE based systems, it is of obvious interest to understand the performance gain one can expect from the increased memory and computational complexity.

The comparison is based on a "General" DNN based SE system trained using all the noise types from the noise dimension experiments at all the SNRs from the SNR dimension experiments, and using all the speakers from the speaker dimension experiments. This means that the system is trained using 7 different and equally distributed noise types mixed with 20 female and 20 male speakers at SNRs from -15 dB to 20 dB. To encompass the increased variability of this dataset compared

TABLE XII
AVERAGE STOI PERFORMANCE IMPROVEMENT SCORES USING A STATE-OF-THE-ART STSA-MMSE ESTIMATOR. THE SCORE IN THE PARENTHESIS IS FOR THE NOISY UNPROCESSED SIGNALS. THE TEST MATERIAL IS BASED ON 2000 UTTERANCES EVENLY DISTRIBUTED AMONG 20 MALES AND 20 FEMALES MIXED WITH 10 DIFFERENT NOISE TYPES FROM THE DEMAND NOISE CORPUS AT SEVEN SNRS IN THE RANGE FROM −10 DB TO 20 DB

	−10 dB	−5 dB	0 dB	5 dB	10 dB	15 dB	20 dB
DM bus	−0.003 (0.819)	−0.003 (0.884)	−0.003 (0.927)	−0.003 (0.955)	−0.003 (0.972)	−0.003 (0.983)	−0.003 (0.991)
DM cafe	−0.026 (0.521)	−0.011 (0.643)	−0.003 (0.756)	−0.001 (0.843)	−0.002 (0.902)	−0.003 (0.939)	−0.003 (0.962)
DM cafeter	−0.043 (0.459)	−0.022 (0.58)	−0.006 (0.706)	−0.001 (0.811)	−0.002 (0.884)	−0.003 (0.928)	−0.003 (0.955)
DM car	0.008 (0.913)	0.005 (0.945)	0.002 (0.966)	0.000 (0.979)	−0.001 (0.987)	−0.002 (0.992)	−0.002 (0.996)
DM metro	0.002 (0.62)	0.007 (0.73)	0.006 (0.821)	0.002 (0.886)	0.000 (0.929)	−0.001 (0.955)	−0.002 (0.972)
DM resto	−0.054 (0.395)	−0.031 (0.496)	−0.012 (0.623)	−0.004 (0.746)	−0.003 (0.84)	−0.004 (0.902)	−0.004 (0.939)
DM river	0.011 (0.55)	0.020 (0.655)	0.020 (0.755)	0.013 (0.838)	0.006 (0.897)	**0.002** (0.936)	**0.000** (0.961)
DM square	−0.008 (0.651)	−0.001 (0.761)	−0.000 (0.846)	−0.002 (0.904)	−0.003 (0.94)	−0.003 (0.962)	−0.003 (0.977)
DM station	0.008 (0.496)	**0.022** (0.614)	**0.023** (0.733)	**0.016** (0.829)	**0.008** (0.894)	**0.002** (0.934)	**0.000** (0.958)
DM traffic	**0.019** (0.611)	0.021 (0.724)	0.016 (0.819)	0.009 (0.887)	0.003 (0.93)	0.001 (0.957)	−0.001 (0.974)
Average	−0.009 (0.604)	0.001 (0.703)	0.004 (0.795)	0.003 (0.868)	0.000 (0.917)	−0.001 (0.949)	−0.002 (0.968)

TABLE XIII
AS TABLE XII BUT FOR A STATE-OF-THE-ART DNN BASED SE ALGORITHM

	−10 dB	−5 dB	0 dB	5 dB	10 dB	15 dB	20 dB
DM bus	0.033 (0.819)	0.021 (0.884)	0.011 (0.927)	0.004 (0.955)	−0.001 (0.972)	−0.004 (0.983)	−0.006 (0.991)
DM cafe	0.034 (0.521)	0.058 (0.643)	0.054 (0.756)	0.038 (0.843)	0.022 (0.902)	0.010 (0.939)	0.002 (0.962)
DM cafeter	−0.011 (0.459)	0.038 (0.58)	0.056 (0.706)	0.045 (0.811)	0.027 (0.884)	0.014 (0.928)	0.005 (0.955)
DM car	0.018 (0.913)	0.008 (0.945)	0.002 (0.966)	−0.002 (0.979)	−0.004 (0.987)	−0.006 (0.992)	−0.007 (0.996)
DM metro	0.040 (0.62)	0.046 (0.73)	0.038 (0.821)	0.024 (0.886)	0.012 (0.929)	0.004 (0.955)	−0.002 (0.972)
DM resto	−0.017 (0.395)	0.046 (0.496)	0.078 (0.623)	**0.069** (0.746)	**0.043** (0.84)	**0.022** (0.902)	**0.009** (0.939)
DM river	0.077 (0.55)	0.089 (0.655)	0.074 (0.755)	0.048 (0.838)	0.026 (0.897)	0.011 (0.936)	0.001 (0.961)
DM square	0.064 (0.651)	0.054 (0.761)	0.036 (0.846)	0.021 (0.904)	0.010 (0.94)	0.003 (0.962)	−0.003 (0.977)
DM station	0.076 (0.496)	**0.096** (0.614)	**0.080** (0.733)	0.051 (0.829)	0.027 (0.894)	0.013 (0.934)	0.004 (0.958)
DM traffic	**0.085** (0.611)	0.072 (0.724)	0.048 (0.819)	0.027 (0.887)	0.013 (0.93)	0.004 (0.957)	−0.002 (0.974)
Average	0.040 (0.604)	0.053 (0.703)	0.048 (0.795)	0.032 (0.868)	0.018 (0.917)	0.007 (0.949)	0.000 (0.968)

to the previous datasets the training set size is increased to $40 \times 231 \times 12 = 110880$ utterances. To make a fair comparison to the STSA-MMSE method, which does not strongly rely on prior speaker, SNR, or noise type knowledge, 10 unseen noises, 20 unseen females and 20 unseen males are used for evaluating the performance at SNRs from −10 dB to 20 dB. The 10 noises are taken from the DEMAND noise database[5] and represent a wide range of both stationary and non-stationary noise types.

The STSA-MMSE method relies on the assumption that noise free Discrete Fourier Transform (DFT) coefficients are distributed according to a generalized gamma distribution with parameters $\gamma = 2$ and $\nu = 0.15$ [1], [6]. The *a priori* SNR estimator used by the STSA-MMSE method is the Decision-Directed approach [64] using a smoothing factor of 0.98 and a noise Power Spectral Density (PSD) estimate based on the noise PSD tracker reported in [67][6]. For each utterance, the noise tracker was initialized using a noise PSD estimate based on a noise only region prior to speech activity.

The results of the experiments are presented in Tables XII and XIV for the STSA-MMSE method and in Tables XIII and XV for the general DNN based SE system. The performance scores for the noisy unprocessed mixtures are given in parenthesis and the average across all 10 noises at each SNR is given in the last row. From Tables XII and XIII it is seen that for all SNRs, the DNN based SE system outperforms the STSA-MMSE method in terms of STOI. Similar behavior is seen from Tables XIV and XV, where the systems are evaluated using PESQ. However, at high SNRs the STSA-MMSE method achieves comparable results with the DNN based SE method and for some noise types such as *DM station* and *DM traffic*, the STSA-MMSE even achieves slightly better PESQ scores at SNRs above 5 dB. This might be explained by the fact that the STSA-MMSE algorithm uses prior knowledge in terms of an ideal noise PSD estimate based on a noise only signal region prior to speech activity. This prior knowledge could be particularly beneficial for stationary noise types, where the initial noise PSD estimate remains correct throughout the utterance. The DNN based SE method explored in this paper does not utilize such prior knowledge. However, in [16], [68] noise PSD estimates obtained prior to speech activity were used in combination with traditional features to train a DNN based SE system and it was shown that performance was improved, when such prior knowledge was utilized. It is also seen that the STSA-MMSE method on average does not improve STOI, whereas the general DNN based SE method does. For some conditions such as *DM station* at an SNR of −5 dB the improvement is as high as 0.096. In general, it can be observed that a DNN based SE system trained across

[5]http://parole.loria.fr/DEMAND
[6]http://insy.ewi.tudelft.nl/content/software-and-data

TABLE XIV
AS TABLE XII BUT FOR PESQ

	−10 dB	−5 dB	0 dB	5 dB	10 dB	15 dB	20 dB
DM bus	0.172 (1.26)	0.278 (1.51)	0.325 (1.93)	0.336 (2.46)	0.264 (3.05)	0.129 (3.61)	−0.001 (4.04)
DM cafe	−0.035 (1.13)	0.013 (1.11)	0.067 (1.20)	0.142 (1.42)	0.206 (1.83)	0.222 (2.37)	0.174 (2.97)
DM cafeter	−0.061 (1.17)	−0.001 (1.11)	0.056 (1.16)	0.129 (1.32)	0.206 (1.65)	0.228 (2.14)	0.177 (2.73)
DM car	**0.363** (1.21)	**0.500** (1.45)	**0.682** (1.81)	**0.742** (2.35)	0.677 (2.94)	0.459 (3.53)	0.192 (4.01)
DM metro	0.019 (1.13)	0.134 (1.17)	0.264 (1.33)	0.356 (1.64)	0.370 (2.12)	0.297 (2.70)	0.180 (3.28)
DM resto	−0.130 (1.25)	−0.046 (1.13)	0.029 (1.11)	0.113 (1.20)	0.227 (1.43)	0.305 (1.84)	0.295 (2.40)
DM river	0.009 (1.07)	0.061 (1.09)	0.183 (1.17)	0.410 (1.36)	0.605 (1.75)	0.632 (2.30)	0.500 (2.92)
DM square	0.039 (1.08)	0.102 (1.14)	0.203 (1.29)	0.303 (1.61)	0.344 (2.10)	0.330 (2.68)	0.247 (3.29)
DM station	−0.008 (1.09)	0.093 (1.07)	0.271 (1.13)	0.511 (1.30)	0.688 (1.63)	**0.705** (2.14)	**0.565** (2.75)
DM traffic	0.054 (1.07)	0.188 (1.09)	0.406 (1.19)	0.615 (1.43)	**0.706** (1.86)	0.700 (2.43)	0.558 (3.05)
Average	0.042 (1.14)	0.132 (1.19)	0.249 (1.33)	0.366 (1.61)	0.429 (2.04)	0.401 (2.57)	0.289 (3.14)

TABLE XV
AS TABLE XII BUT FOR PESQ WITH A STATE-OF-THE-ART DNN BASED SE ALGORITHM

	−10 dB	−5 dB	0 dB	5 dB	10 dB	15 dB	20 dB
DM bus	0.414 (1.26)	0.550 (1.51)	0.596 (1.93)	0.543 (2.46)	0.401 (3.05)	0.225 (3.61)	0.080 (4.04)
DM cafe	0.007 (1.13)	0.149 (1.11)	0.300 (1.20)	0.425 (1.42)	0.479 (1.83)	0.467 (2.37)	0.372 (2.97)
DM cafeter	−0.047 (1.17)	0.066 (1.11)	0.196 (1.16)	0.350 (1.32)	0.471 (1.65)	0.499 (2.14)	0.423 (2.73)
DM car	**0.811** (1.21)	**1.021** (1.45)	**1.119** (1.81)	**1.005** (2.35)	**0.763** (2.94)	0.447 (3.53)	0.167 (4.01)
DM metro	0.095 (1.13)	0.242 (1.17)	0.373 (1.33)	0.463 (1.65)	0.483 (2.12)	0.408 (2.70)	0.274 (3.28)
DM resto	−0.140 (1.25)	−0.009 (1.13)	0.157 (1.11)	0.338 (1.20)	0.499 (1.43)	0.571 (1.84)	**0.523** (2.4)
DM river	0.111 (1.07)	0.268 (1.09)	0.464 (1.17)	0.639 (1.36)	0.682 (1.75)	**0.615** (2.30)	0.445 (2.92)
DM square	0.170 (1.08)	0.327 (1.14)	0.484 (1.29)	0.572 (1.61)	0.564 (2.10)	0.489 (2.69)	0.343 (3.29)
DM station	0.059 (1.08)	0.199 (1.07)	0.356 (1.13)	0.513 (1.30)	0.610 (1.63)	0.603 (2.14)	0.478 (2.75)
DM traffic	0.172 (1.07)	0.347 (1.09)	0.513 (1.19)	0.620 (1.43)	0.636 (1.86)	0.575 (2.43)	0.427 (3.05)
Average	0.165 (1.14)	0.316 (1.19)	0.456 (1.33)	0.547 (1.61)	0.559 (2.04)	0.490 (2.57)	0.353 (3.14)

all three generalizability dimensions using a large number of noise types, speakers and SNRs, outperforms a state-of-the-art non-machine learning based method, even though this method utilizes prior knowledge in terms of ideal initial noise PSD estimates. However, the performance of the general DNN based SE system is on average considerably reduced compared to the specialized systems where only one generalizability dimension was varied at a time. From this, it can be concluded that if the usage situation of a SE algorithm is well-defined e.g., in terms of speaker characteristics, noise type, or SNR range, considerably performance improvements can be achieved using a DNN based SE algorithm that has been specifically trained to fit the application. On the other hand, for more general applications where the acoustic usage situation cannot be narrowed down in one or more of these dimensions, the advantage of DNN based SE methods is much smaller, while they may still offer improvements over current state-of-the-art non-machine learning based methods.

E. Listening Test

To investigate how the DNN based SE system performs in practice, an intelligibility test, using 10 normal-hearing Danish graduate students, has been conducted. The gender distribution among the 10 students was 3 females and 7 males with ages from 20 to 28 years and a mean age of 24. Five systems have been designed for the SI test and their training specifications are given by Table XVI.

The systems are designed to investigate if a female specific system, in different noise and SNR conditions (DNN-1–DNN-4), can improve SI, when exposed to an unseen female speaker. This is an extension of the experiments in [9] where the system was tested in matched speaker and matched SNR conditions only.

Furthermore, DNN-5, which is a "general" system that has been trained on a wide range of speakers, noise types and SNRs, is included in the experiments to investigate if such a general system can improve SI, when exposed to both an unseen speaker and noise type.

The noise types used for training DNN-5 include white Gaussian noise (WGN), babble noise (BBL-ADFD) and N3–N7 from the noise dimension tests described in subsection III-B. The BBL-ADFD noise is constructed using the procedure for BBL, as described in subsection III-A, but with three males and three females from the unused part of the ADFD corpus. Each test subject was exposed to five repetitions of 32 test conditions (2 noise types $\times$ 4 SNRs $\times$ 4 processing conditions), hence each test subject was exposed to a total of 160 sentences. The two noise types are SSN (N1) and BBL (N2) noise and the four SNRs are -13 dB, -9 dB, -5 dB and -1 dB. This SNR range was chosen to cover SNRs where SI is close to 0% (-13 dB) and close to 100% (-1 dB). The four processing conditions for each noise type were unprocessed corrupted speech, and corrupted speech processed by DNN-1, DNN-2, and DNN-5, for SSN and DNN-3, DNN-4, and DNN-5, for BBL noise.

TABLE XVI
DNN BASED SE SYSTEMS USED FOR THE INTELLIGIBILITY TEST PRESENTED IN FIGS. 1 AND 2. THE FIRST COLUM SHOWS THE SYSTEM ID AND THE REMAINING COLUMNS SHOW THE TRAINING CRITERIA

system ID	Noise Dim.	SNR Dim.	Speaker Dim.
DNN-1	SSN	−5 dB	20 Female
DNN-2	SSN	−15 dB – 20 dB	20 Female
DNN-3	BBL	−5 dB	20 Female
DNN-4	BBL	−15 dB – 20 dB	20 Female
DNN-5	N3–N7, WGN, BBL-ADFD	−15 dB – 20 dB	20 Female, 20 Male

Immediately prior to the listening test, each test subject performed a familiarization test using 24 noisy utterances from a left out test set. The speech material used for the SI test was based on the Danish Dantale-II speech corpus [69]. Each utterance, which is spoken by a female, consists of five words from five different word classes appearing in the following order: name, verb, numeral, adjective and a noun and the test subject was asked to identify the spoken words via a computer interface. There are a total of 10 different words within each word class, hence the Dantale-II corpus is based on a total of 50 different words. All sentences are constructed such that they are syntactically correct but semantically unlikely, which makes it difficult to predict one word based on another, hence the corpus is suitable for intelligibility tests. The SI test was performed in an audiometric booth using a set of beyerdynamic DT 770 headphones and a Focusrite Scarlett 2i2 sound card

The results are presented in Figs. 1 and 2 for SSN and BBL noise, respectively. Figs. 1 and 2 show that DNN-5, which is the speaker, noise type, and SNR general system, is unable to improve SI at any of the four SNRs of BBL noise as well as the SNRs at −13 dB, −9 dB, and −1 dB of SSN. A paired-sample t-test shows that this SI degradation is statistical significant, i.e. $p < 0.05$, for all these results. It is also seen that DNN-5 improves SI with a small amount for SSN at an SNR of −5 dB. However, this improvement is not statistically significant ($p = 0.44$). For DNN-2 and DNN-4, which are the female and noise type specific, but SNR general systems, a somewhat different picture is observed. In general both DNN-2 and DNN-4 perform better than DNN-5. For SSN, DNN-2 manages to improve SI over the unprocessed signals at SNR −9 dB, while DNN-4 improves SI at SNRs of −5 dB and −1 dB. However, none of these improvements are statistical significant ($p = 0.10, p = 0.10, p = 0.25$, respectively)

Finally, for DNN-1 and DNN-2, which are the female, noise type, and SNR specific systems, DNN-1 improves over DNN-2, whereas DNN-3 in general performs worse than DNN-4. Especially at an SNR of −5 dB DNN-3 performs significantly ($p < 0.001$) worse than DNN-4 ($p = 0.10$) relative to the unprocessed signals. This is surprising since DNN-3 is trained at only −5 dB SNR, while DNN-4 had been trained using the SNR range from −15 dB to 20 dB. Furthermore, the observed SI improvement, especially for DNN-4 and DNN-5 using BBL, is lower than one would expect based on the STOI scores for related models in Sec. III. This discrepancy between STOI scores and observed SI, especially for highly modulated noise signals,

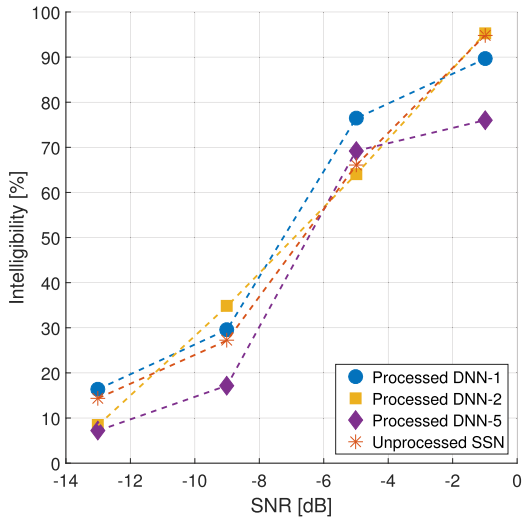


Fig. 1. SI test results for 3 different DNN based SE systems processing SSN corrupted speech signals based on 10 Danish test subjects.

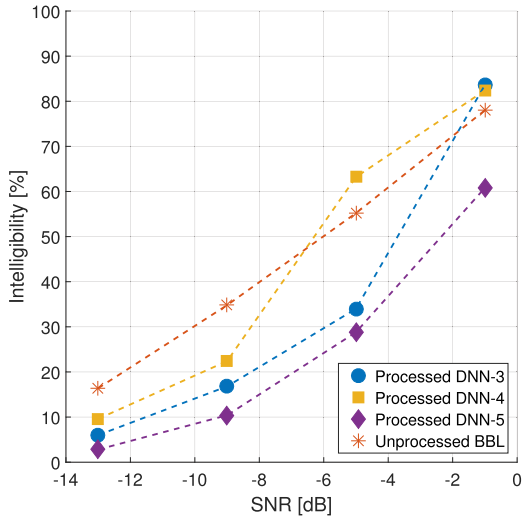


Fig. 2. SI test results for 3 different DNN based SE systems processing BBL corrupted speech signals based on 10 Danish test subjects.

has previously been observed [9], [13], [55], [70]. For DNN-1 a statistically significant improvement of 10.4 percentage points ($p = 0.011$) in SI is observed at an SNR of −5 dB, which also corresponds to the SNR at which DNN-1 is trained. To the authors knowledge, SI improvements achieved by a female specific DNN based SE system tested on an unseen female speaker has not yet been reported. Furthermore, the system outperforms a wide range of previously reported SI test results by non-machine learning based methods reported in [65], [66] and is comparable with the SI results reported in [37] where a single continuous-gain MMSE method was used.

IV. CONCLUSION

In this paper the generalizability of a state-of-the-art Deep Neural Network (DNN) based Speech Enhancement (SE) method has been investigated. Specifically, it has been

investigated how noise specific, speaker specific and Signal-to-Noise Ratio (SNR) specific systems perform in relation to noise general, speaker general and SNR general systems, respectively. Furthermore, it has been investigated how such systems perform in relation to a single DNN based SE system which has been designed to be speaker, noise type and SNR general. Also, a comparison between this general DNN based SE system and a state-of-the-art Short-Time Spectral Amplitude Minimum Mean Square Error (STSA-MMSE) based SE method has been conducted. In general, a positive correspondence between training data variability and generalization was observed. Specifically, it was found that DNN based SE systems generalize well to both unseen speakers and unseen noise types given a large number of speakers and noise types were included in the training set. Furthermore, it was found that specialized DNN based SE systems trained on only one noise type, one speaker or one SNR, outperformed DNN based SE systems trained on a wide range of noise types, speakers, and SNRs in terms of both estimated Speech Quality (SQ) and estimated Speech Intelligibility (SI). In addition, a general DNN based SE algorithm trained using a large number of speakers, a large number of noise types at a large range of SNRs, outperformed a state-of-the-art STSA-MMSE SE algorithm in terms of estimated SQ and SI. However, the performance of this general DNN based SE system, was considerably reduced compared to the specialized systems, that have been optimized to only a single noise type, a single speaker or a single SNR. Finally, it was found that a DNN based SE system trained to be female, noise type and SNR specific, was able to improve SI when tested with an unseen female speaker for particular SNR and noise type configurations, although degrading SI for others.

In general, it can be concluded that DNN based SE systems do have potential to improve SI in a broader range of usage situations than investigated in [9], [13]. Furthermore, the experiments conducted in this paper, indicate that matching the noise type is critical in acquiring good performance for DNN based SE algorithms, whereas matching the SNR dimension is the least critical followed by the speaker dimension for which good generalization can be achieved with a modest amount of training speakers. Also, it can be concluded that considerable improvement in performance can be achieved if the usage situation is limited such that the DNN based SE method can be optimized towards a specific application.

Even though the results reported in this paper are considered general, there is some experimental evidence [13], [16], [20], [46], [53], [54] showing that generalizability performance of DNN based SE algorithms, and DNNs in general, improves when more data and larger networks are being applied, hence SQ and SI performance of DNN based SE systems are expected to improve in the future, when more data and computational resources become available.

ACKNOWLEDGMENT

The authors would like to thank A. H. Andersen for providing software used to conduct the SI tests, and NVIDIA Corporation for the donation of a Titan X GPU.

REFERENCES

[1] 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*, (Synthesis Lectures on Speech and Audio Processing). San Rafael, CA, USA: Morgan & Claypool, Jan. 2013, vol. 9, no. 1, pp. 1–80.

[2] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL, USA: CRC Press, 2013, vol. 2013.

[3] Y. Wang, "Supervised speech separation using deep neural networks," Ph.D. dissertation, Dept. Comput. Sci. Eng., Ohio State Univ., Columbus, OH, USA, 2015.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[5] R. Martin, "Speech enhancement based on minimum mean-square error estimation and super-Gaussian priors," *IEEE Speech Audio Process.*, vol. 13, no. 5, pp. 845–856, Sep. 2005.

[6] J. Erkelens, R. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 6, pp. 1741–1752, Aug. 2007.

[7] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.

[8] E. W. Healy, S. E. Yoho, Y. Wang, and D. Wang, "An algorithm to improve speech recognition in noise for hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 134, no. 4, pp. 3029–3038, Oct. 2013.

[9] E. W. Healy, S. E. Yoho, J. Chen, Y. Wang, and D. Wang, "An algorithm to increase speech intelligibility for hearing-impaired listeners in novel segments of the same noise type," *J. Acoust. Soc. Am.*, vol. 138, no. 3, pp. 1660–1669, Sep. 2015.

[10] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1486–1494, Sep. 2009.

[11] K. Han and D. Wang, "A classification based approach to speech segregation," *J. Acoust. Soc. Am.*, vol. 132, no. 5, pp. 3475–3483, Nov. 2012.

[12] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1381–1390, Jul. 2013.

[13] 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," *J. Acoust. Soc. Am.*, vol. 139, no. 5, pp. 2604–2612, May 2016.

[14] J. Chen, Y. Wang, and D. Wang, "Noise perturbation for supervised speech separation," *Speech Commun.*, vol. 78, pp. 1–10, 2016.

[15] K. Han and D. Wang, "Towards generalizing classification based speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 1, pp. 168–177, Jan. 2013.

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

[17] T. Lee and F. Theunissen, "A single microphone noise reduction algorithm based on the detection and reconstruction of spectro-temporal features," *Proc. R. Soc. Lond. A, Math. Phys. Sci.*, vol. 471, no. 2184, Dec. 2015, http://rspa.royalsocietypublishing.org/content/471/2184/20150309.full

[18] 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 Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 12, pp. 2136–2147, Dec. 2015.

[19] D. Liu, P. Smaragdis, and M. Kim, "Experiments on deep learning for speech denoising," in *Proc. INTERSPEECH*, 2014, pp. 2685–2689.

[20] Y. Wang, J. Chen, and D. Wang, "Deep neural network based supervised speech segregation generalizes to novel noises through large-scale training," Dept. Comput. Sci. Eng., Ohio State Univ., Columbus, OH, USA, Tech. Rep. OSU-CISRC-3/15-TR02, 2015.

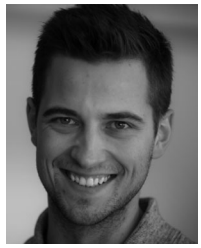
[21] S. Gonzalez and M. Brookes, "Mask-based enhancement for very low quality speech," in *Proc. 2014 IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2014, pp. 7029–7033.

[22] J. Chen and D. Wang, "Long short-term memory for speaker generalization in supervised speech separation," in *Proc. 17th Annu. Conf. Int. Speech Commun. Assoc.*, 2016, pp. 3314–3318.

[23] M. Delfarah and D. Wang, "A feature study for masking-based reverberant speech separation," in *Proc. 17th Annu. Conf. Int. Speech Commun. Assoc.*, 2016, pp. 555–559.

[24] A. Kumar and D. Florencio, "Speech enhancement in multiple-noise conditions using deep neural networks," arXiv:1605.02427 [cs], May 2016.

[25] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006, vol. 2006.
[26] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.
[27] J. Garofolo *et al.*, "TIMIT acoustic-phonetic continuous speech corpus LDC93S1. Web Download," *Philadelphia: Linguistic Data Consortium*, 1993.
[28] T. May and T. Dau, "Requirements for the evaluation of computational speech segregation systems," *J. Acoust. Soc. Am.*, vol. 136, no. 6, pp. EL398–EL404, Dec. 2014.
[29] Y. Wang and D. Wang, "A deep neural network for time-domain signal reconstruction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2015, pp. 4390–4394.
[30] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.
[31] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Process.*, vol. 81, no. 11, pp. 2403–2418, Nov. 2001.
[32] J. Erkelens, R. Hendriks, and R. Heusdens, "On the estimation of complex speech DFT coefficients without assuming independent real and imaginary parts," *IEEE Signal Process. Lett.*, vol. 15, pp. 213–216, 2008, http://ieeexplore.ieee.org/abstract/document/4443129
[33] R. Hendriks, J. Erkelens, and R. Heusdens, "Comparison of complex-DFT estimators with and without the independence assumption of real and imaginary parts," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2008, pp. 4033–4036.
[34] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2000, vol. 13, pp. 556–562.
[35] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, Jan. 2014.
[36] D. Wang, "Time-frequency masking for speech separation and its potential for hearing aid design," *Trends Amplif.*, vol. 12, no. 4, pp. 332–353, Dec. 2008.
[37] J. Jensen and R. Hendriks, "Spectral magnitude minimum mean-square error estimation using binary and continuous gain functions," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 92–102, Jan. 2012.
[38] C. Hummersone, T. Stokes, and T. Brookes, "On the ideal ratio mask as the goal of computational auditory scene analysis," in *Blind Source Separation* (ser. Signals and Communication Technology), G. R. Naik and W. Wang, Eds. Berlin, Germany: Springer, 2014, pp. 349–368.
[39] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. New York, NY, USA: Springer, 2005, pp. 181–197.
[40] N. Madhu, A. Spriet, S. Jansen, R. Koning, and J. Wouters, "The potential for speech intelligibility improvement using the ideal binary mask and the ideal Wiener filter in single channel noise reduction systems: Application to auditory prostheses," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 1, pp. 63–72, Jan. 2013.
[41] D. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 3, pp. 483–492, 2016.
[42] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. New York, NY, USA: Wiley, 2006.
[43] E. W. Healy, S. E. Yoho, Y. Wang, F. Apoux, and D. Wang, "Speech-cue transmission by an algorithm to increase consonant recognition in noise for hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 136, no. 6, pp. 3325–3336, Dec. 2014.
[44] J. Chen, Y. Wang, and D. Wang, "A feature study for classification-based speech separation at low signal-to-noise ratios," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1993–2002, Dec. 2014.
[45] Y. Wang, K. Han, and D. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 2, pp. 270–279, Feb. 2013.
[46] D. Amodei *et al.*, "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin," *CoRR*, Dec. 2015, http://arxiv.org/abs/1512.02595
[47] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 807–814.
[48] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.
[49] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, pp. 2121–2159, Jul. 2011.
[50] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *CoRR*, Jul. 2012, http://arxiv.org/abs/1207.0580
[51] A. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 14–22, Jan. 2012.
[52] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, U. D. Montral, and M. Qubec, "Greedy layer-wise training of deep networks," in *Proc. Adv. Neural Inf. Process. Syst. 19*, 2007, pp. 153–160.
[53] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *IEEE Intell. Syst.*, vol. 24, no. 2, pp. 8–12, Mar. 2009.
[54] A. Hannun *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *CoRR*, Dec. 2014, http://arxiv.org/abs/1412.5567
[55] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.
[56] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2001, pp. 749–752 vol. 2.
[57] "P.862.2: Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs," 2005. [Online]. Available: https://www.itu.int/rec/T-REC-P.862.2-200511-S/en
[58] C. Fvotte, R. Gribonval, and E. Vincent, "BSS EVAL Toolbox User Guide Revision 2.0," Institut de Recherche en Informatique et Systmes Alatoires, Inria Rennes—Bretagne Atlantique, Rennes, France, Tech. Rep. inria-00564760, 2011. [Online]. Available: https://hal.inria.fr/inria-00564760
[59] X. L. Zhang and D. Wang, "A deep ensemble learning method for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 5, pp. 967–977, May 2016.
[60] J. Barker, M. Ricard, V. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," *2015 IEEE Workshop Automat. Speech Recognit. Understanding*, 2015, pp. 504–511.
[61] D. Williamson, Y. Wang, and D. Wang, "Deep neural networks for estimating speech model activations," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2015, pp. 5113–5117.
[62] M. Kolbæk, Z.-H. Tan, and J. Jensen, "Speech enhancement using long short-term memory based recurrent neural networks for noise robust speaker verification," in *Proc. 2016 IEEE Spoken Lang. Technol. Workshop*, 2016, http://www.slt2016.org/Papers/AcceptedPapers.asp
[63] D. S. Williamson, Y. Wang, and D. Wang, "Estimating nonnegative matrix model activations with deep neural networks to increase perceptual speech quality," *J. Acoust. Soc. Am.*, vol. 138, no. 3, pp. 1399–1407, Sep. 2015.
[64] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.
[65] Y. Hu and P. C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," *J. Acoust. Soc. Am.*, vol. 122, no. 3, pp. 1777–1786, Sep. 2007.
[66] H. Luts *et al.*, "Multicenter evaluation of signal enhancement algorithms for hearing aids," *J. Acoust. Soc. Am.*, vol. 127, no. 3, pp. 1491–1505, Mar. 2010.
[67] R. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2010, pp. 4266–4269.
[68] M. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2013, pp. 7398–7402.
[69] K. Wagener, J. L. Josvassen, and R. Ardenkjaer, "Design, optimization and evaluation of a Danish sentence test in noise," *Int. J. Audiol.*, vol. 42, no. 1, pp. 10–17, Jan. 2003.
[70] S. Jørgensen, R. Decorsre, and T. Dau, "Effects of manipulating the signal-to-noise envelope power ratio on speech intelligibility," *J. Acoust. Soc. Am.*, vol. 137, no. 3, pp. 1401–1410, Mar. 2015.



Morten Kolbæk received the B.Eng. degree in electronic design from Aarhus University, Business and Social Sciences, Herning, Denmark, in 2013, and the M.Sc. degree in signal processing and computing from Aalborg University, Aalborg, Denmark, in 2015. He is currently working toward the Ph.D. degree at the Section for Signal and Information Processing, Department of Electronic Systems, Aalborg University, under the supervision of Z.-H. Tan and J. Jensen. His main research interests include speech enhancement, deep learning, and intelligibility improvement of noisy speech for hearing-aids applications.

Zheng-Hua Tan (M'00–SM'06) received the B.Sc. and M.Sc. degrees in electrical engineering from Hunan University, Changsha, China, in 1990 and 1996, respectively, and the Ph.D. degree in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 1999. He is currently an Associate Professor in the Department of Electronic Systems, Aalborg University, Aalborg, Denmark. He is also a cofounder of the Centre for Acoustic Signal Processing Research, Aalborg University. He was a Visiting Scientist at the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA, an Associate Professor in the Department of Electronic Engineering, Shanghai Jiao Tong University, and a Postdoctoral Fellow in the Department of Computer Science, Korea Advanced Institute of Science and Technology, Daejeon, South Korea. His research interests include speech and speaker recognition, noise-robust speech processing, multimedia signal and information processing, human–robot interaction, and machine learning. He has published extensively in these areas in refereed journals and conference proceedings. He has served as an Editorial Board Member/Associate Editor for Elsevier *Computer Speech and Language*, Elsevier *Digital Signal Processing*, and Elsevier *Computers and Electrical Engineering*. He was a Lead Guest Editor of the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING. He has served as a Chair, Program Co-chair, Area and Session Chair, and Tutorial Speaker of many international conferences.

Jesper Jensen received the M.Sc. degree in electrical engineering and the Ph.D. degree in signal processing from Aalborg University, Aalborg, Denmark, in 1996 and 2000, respectively. From 1996 to 2000, he was with the Center for Person Kommunikation, Aalborg University, as a Ph.D. student and as an Assistant Research Professor. From 2000 to 2007, he was a Postdoctoral Researcher and an Assistant Professor with Delft University of Technology, Delft, The Netherlands, and an External Associate Professor with Aalborg University. He is currently a Senior Researcher with Oticon A/S, Copenhagen, Denmark, where his main responsibility is scouting and development of new signal processing concepts for hearing-aid applications. He is a Professor with the Section for Information Processing, Department of Electronic Systems, Aalborg University. He is also a cofounder of the Centre for Acoustic Signal Processing Research, Aalborg University. His main interests include the area of acoustic signal processing, including signal retrieval from noisy observations, coding, speech and audio modification and synthesis, intelligibility enhancement of speech signals, signal processing for hearing aid applications, and perceptual aspects of signal processing.