Training Data-Driven Speech Intelligibility Predictors on Heterogeneous Listening Test Data

Pedersen, Mathias Bach; Andersen, Asger H.; Jensen, Soren Holdt; Tan, Zheng Hua; Jensen, Jesper

Published in:
IEEE Access

DOI (link to publication from Publisher):
10.1109/ACCESS.2022.3184785

Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Pedersen, M. B., Andersen, A. H., Jensen, S. H., Tan, Z. H., & Jensen, J. (2022). Training Data-Driven Speech Intelligibility Predictors on Heterogeneous Listening Test Data. *IEEE Access*, *10*, 66175-66189. https://doi.org/10.1109/ACCESS.2022.3184785

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 11, 2022, accepted June 11, 2022, date of publication June 21, 2022, date of current version June 24, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3184785

Training Data-Driven Speech Intelligibility Predictors on Heterogeneous Listening Test Data

MATHIAS BACH PEDERSEN[1], (Member, IEEE), ASGER H. ANDERSEN[2,3], SØREN HOLDT JENSEN[4], (Senior Member, IEEE), ZHENG-HUA TAN[1], (Senior Member, IEEE), AND JESPER JENSEN[1,2]

[1]Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark
[2]Demant A/S, 2765 Smørum, Denmark
[3]WS Audiology A/S, 3540 Lynge, Denmark
[4]The Danish Ministry of Defence Estate Agency, 9800 Hjoerring, Denmark

Corresponding author: Mathias Bach Pedersen (mbp@es.aau.dk)

This work was supported by the Independent Research Fund Denmark under Grant DFF-7017-00017.

ABSTRACT Prediction of Speech Intelligibility (SI) is a topic of interest for most speech processing applications, where intelligibility is of any importance, e.g., speech coding, transmission and enhancement. Traditionally, SI predictors have been based on signal processing methods and heuristics, but more recently, an increasing number of data-driven SI-predictors have been proposed. Data-driven prediction of SI requires large quantities of labelled data, ideally from many listening tests. Listening tests differ in factors such as vocabulary, talker, listener's task, etc. collectively referred to as the paradigm. A naïve strategy of training SI-predictors directly on stimuli, pooled from different listening tests, is futile because the exact map from the stimulus to SI is determined, not only by the stimulus, but also by the paradigm. Data-driven SI-predictors trained in this way become specialized to the paradigms of the training data by erroneously attributing all paradigm influences on SI to the stimulus. The problem is fundamental and persists even in the idealized situation where training data is abundant. We propose a strategy for training data-driven SI-predictors that is independent of the paradigms, underlying the training data. The proposed strategy is to concatenate an SI-predictor and a layer of trainable dataset-specific mapping functions, each corresponding to a single paradigm in the training data. These mapping functions are trained jointly with the SI-predictor and serve to efficiently approximate the psychometric functions implied by each paradigm. The mapping functions prevent the predictor from specializing to these paradigms during training. We present an SI-predictor with a novel architecture that incorporates a convolutional network and an ESTOI back-end, train it with this strategy, compare it to naïve training and a range of existing non-data-driven predictors. The proposed training strategy and architecture results in higher performance overall and increased robustness to unseen paradigms.

INDEX TERMS Neural networks, psychometric functions, speech intelligibility prediction.

I. INTRODUCTION

Speech Intelligibility (SI) is an important concept for speech communication devices, such as hearing aid systems or devices for communicating under extreme acoustic conditions, such as aeroplane cockpits or emergency response situations. Because of this, SI is repeatedly measured during the development of these devices. The most reliable measurements of SI come from listening tests, where human listeners respond to examples of the noisy or processed speech in question. Since many human listeners need to be involved, these listening tests are significant time-sinks, slowing down iterative development of speech processing methods, concerned with SI.

To speed up this development, SI-prediction has become a popular and valuable tool. SI-prediction refers to algorithms or models, designed to predict the SI of noisy or processed speech signals, as it would be rated by a panel of human listeners. SI-prediction offers fast and reproducible results and can significantly increase the speed of development for speech processing systems, when used in place of listening tests. A potential disadvantage is that SI-predictors, like any

The associate editor coordinating the review of this manuscript and approving it for publication was Shaikh Anowarul Fattah.

predictor or estimator, may exhibit variable accuracy, depending on variations in the signals under study. These variations include, for instance, the type and intensity of noise or distortions deteriorating the signals and the type of processing applied, if any. We refer to a particular combination of these variations as a *listening condition*. Applying an SI-predictor to listening conditions, on which it has not been validated by a listening test, can give misleading results. Robustness to a wide variety of listening conditions is thus an important quality in SI-prediction.

Data driven SI-predictors are designed using machine learning methods, such as neural networks. These predictors usually have a large number of parameters, which are optimized through training on labelled speech in different listening conditions. When data-driven SI-predictors are trained on speech data from a set of listening test conditions, it makes sense to refer to these listening conditions as *seen* conditions for that predictor. This is in contrast to *unseen* conditions, which refers to conditions not represented in the training set. Data-driven SI-predictors have demonstrated performance improvements over state-of-the-art classical predictors in seen conditions, but not in unseen conditions.

Listening test paradigms are important to consider, when dealing with SI prediction. The paradigm of a listening test refers to factors other than the physical stimuli, such as different talkers, languages, vocabulary, sentence structure, lexical redundancy, test scoring methods, listening equipment and more that have an impact on the measured SI. The effects of a given paradigm can be approximated well by an s-shaped curve, which maps predictions of SI to absolute measured SI. This curve is called a psychometric function, a type of function that relates human responses on a test to some physical quantity, e.g., the SI experienced by the listener vs. signal to noise ratio of the stimulus. Psychometric functions for SI are typically modelled by a sigmoid function, where the parameters depend on the paradigm and the SI-predictor [1]. Note that a difference of slope between the psychometric functions of two listening tests implies that a similar change in a physical quantity, such as SNR, results in different changes to SI.

We use the term ''pooling'' to refer to constructing a dataset that contains speech stimuli and SI labels from multiple listening tests. However, naïve pooling of listening test data may be a questionable approach, because different listening tests have different underlying paradigms and psychometric functions associated with them. The speech stimuli alone do not completely account for the specific SI measurements of a listening test. For instance, the loudspeakers or headphones, used in two different listening tests, could make a difference in the subjective scores of the test subjects. Furthermore, some languages might be easier or harder to understand, under certain noise types. Similarly, coherent sentences allow some words to be inferred by context, which leads to higher SI scores than randomly constructed sentences, devoid of context, in the same listening conditions. These influences on the SI scores of different listening tests result in different parameters of the psychometric function.

Many studies of classical SI-predictors apply listening test specific mapping functions to convert the predictor output to absolute SI in performance tests, e.g., STOI [2], SIIB [3] and SII [4]. This is done in order to take the psychometric functions specific to each listening test into account, when evaluating predictor performance, and thus facilitate comparisons of predictions and performance across different listening tests. The predictions prior to these mappings are typically called *SI indices*, since they are, ideally, related monotonically to the subjectively measured SI, or *absolute SI*. SI indices can be meaningfully compared within the same paradigm, with a higher index corresponding to a higher absolute SI, but indices from different paradigms can not, since the psychometric function, and thus the map from SI index to absolute SI, changes with the paradigm.

When a data-driven SI-predictor is trained on a dataset of pooled listening tests, a fundamental problem arises. The input signals, used to train the SI-predictor, i.e., the speech stimuli, do not contain the complete information that determines the shape of the psychometric functions. With the information available in the training inputs and labels, the predictor can learn the specific psychometric functions underlying the training data, but it can not learn how to adapt to new unseen psychometric functions. This means that the predictor specializes in the paradigms underlying the training data.

We propose and investigate a method for training data-driven predictors, which allows the use of pooled listening test data from different paradigms, by taking the differences in psychometric functions in the training data into account. In particular, the method introduces trainable mapping functions with dataset dependent parameters. These mapping functions, which we call *Dataset-Specific Mapping Functions* (DSMF's), serve to model the psychometric functions specific to each individual listening test in the training data. We apply the training strategy to an SI-predictor[1] consisting of a Convolutional Neural Network (CNN) with a back-end inspired by ESTOI. This CNN is trained with pooled data consisting of speech datasets with SI-labels from different listening tests. The parameters of the trainable mapping functions are learned independently for each dataset. Their purpose is to approximate the psychometric function of each dataset, separately from the SI-predictor. After the training is complete, the trained DSMF's are discarded, because the information they contain, namely an approximation of the psychometric functions of the training sets, is generally not useful, when predicting the SI of unseen datasets and paradigms. The trained SI index predictor is simply the remaining CNN-ESTOI network depicted in Figure 1.

We show that training a data-driven SI-predictor with this strategy prevents it from learning an internal representation of

[1]The implementation of this SI-predictor can be found at https://github.com/Mapede/DSMF_SI_Predictor

the psychometric functions, inherent in the training data. It is demonstrated that this enables the proposed data-driven predictor to reach higher performance for seen conditions, and also to be more robust to new unseen test-paradigms. First, two SI-predictors are trained using the same architecture and pooled data, one using the proposed strategy, the other trained naïvely. This experiment shows that the proposed strategy leads to higher performance on average. Secondly, a series of hold-one-out cross validation experiments are conducted, where SI predictors are trained according to the proposed strategy, using all the available datasets except for one. The dataset, held out of training, is instead used for testing. In these experiments, the average performance of the trained predictors, on their respective unseen datasets, is higher than that of the classical predictors used for comparison.

The paper is organized as follows. Section II goes into detail on existing SI-predictors, both classical and data-driven. Section III describes the architecture of the proposed SI predictor and details of the proposed training procedure. Section IV describes the datasets used to train and test the proposed SI predictor, as well as the training procedure and hyper parameters. Section V describes the experiments, and presents a performance evaluation of the proposed SI predictor. Finally, Section VI contains the conclusions of the work.

II. RELATED WORK

SI-predictors may be roughly divided into classical, or data-driven methods. Classical SI-predictors, e.g., the Articulation Index (AI) [5], the Extended Speech Intelligibility Index (ESII) [6], the Speech-to-Reverberation Modulation energy Ratio (SRMR) [7], the Short-Time Objective Intelligibility (STOI) [2], the Spectro-Temporal Modulation Index (STMI) [8], the Extended Short-Time Objective Intelligibility (ESTOI) [9], the Speech Intelligibility In Bits (SIIB) [3] and the Hearing Aid Speech Perception Index (HASPI) [10], are hand-crafted models, often inspired by models of auditory perception, with only few parameters optimized for listening data. Data-driven SI-predictors, e.g., Non-Intrusive Speech Assessment (NISA) [11], a twin hidden Markov model [12], the data-driven STI estimator proposed by [13], the neural network proposed by [14], the convolutional neural network proposed by [15], and the convolutional neural network proposed by [16], learn a prediction model primarily, or in full, by a process of optimization on a dataset of speech with labels of measured, or in some cases predicted, SI.

Another mode of classification for SI-predictors is, whether they are intrusive or non-intrusive. Intrusive predictors use both the clean reference signal and the noisy/processed test signal, whereas non-intrusive predictors only require the noisy/processed test signal. The advantages of intrusive SI predictors is that they are given more information than their non-intrusive counterparts, and can, in principle, reach a higher accuracy. The advantage of non-intrusive predictors is that they can be used when the clean reference is unavailable.

The AI [5] is perhaps the first classical method, and has served as inspiration for many following predictors. The AI performs a frequency weighted comparison of the long-term intensities of the underlying clean speech and the noise to estimate SI. The primary focus of the AI was speech in additive noise, and it was also designed for calculation by hand. The Speech Transmission Index (STI) [17] analyzes a set of probe signals passed through the transmission channel or processing algorithm of interest. In particular, the preservation of the probe signal modulations are measured, and used to quantify SI. Assuming that the channel is known, the STI supports non-additive distortions, such as clipping, filtering and reverberation.

The Speech Intelligibility Index (SII) [4] and Extended SII (ESII) [6] compute a weighted average of Signal to Noise Ratios (SNR) of specific frequency bands. The SII was proposed as an updated version of the AI, suitable for calculation by computer. In ESII, the SNR is computed in short time frame averages, rather than the long-term average used in SII. This improves its performance for speech signals in fluctuating noise [6]. The STMI [8] decomposes the signal under study into spectro-temporal components, and makes a comparison to the clean reference via cross correlation.

STOI [2] and Extended STOI (ESTOI) [9] use averages of sample correlations between the test signal and clean reference in short time segments in the 1/3 octave band magnitude domain. These sample correlations predict SI well when the time-frequency tiles are independent of each other. Since this is not generally the case, STOI and ESTOI normalize the signal segments before the sample correlations are computed. In STOI each segment is normalized across time, whereas in ESTOI they are also normalized across the 1/3 octave bands. This allows ESTOI to better handle temporally fluctuating noise, compared to STOI, [9]. SIIB [3] provides an estimate of SI via an estimate of the mutual information between the clean speech and noisy/processed speech. The idea of using mutual information to predict SI has been used earlier, see e.g., Speech Intelligibility using Mutual Information (SIMI) [18], the AI [19], [20] and Mutual Information Variational Bayes (MI-VB), MI K Nearest Neighbours (MI-KNN) and MI Expectation Maximization (MI-EM) [21].

HASPI [10] computes an intelligibility score based on an auditory model, including both spectral envelope features and coherence. HASPI is also able to account for hearing impairment.

Data driven SI-predictors can be categorized by the type of labels used for training. The predictors proposed in [14]–[16] and [22], which are all different types of neural networks, are trained to estimate actual listening test results. Other data-driven SI-predictors are trained to emulate existing classical predictors in circumstances, where the classical predictor in question can not be used. In these cases, the labels are SI predictions produced by the classical predictor. For instance, the Non-Intrusive Speech Assessment (NISA) method [11] is trained to predict the outcome of STOI, without the clean reference that STOI normally requires. The important

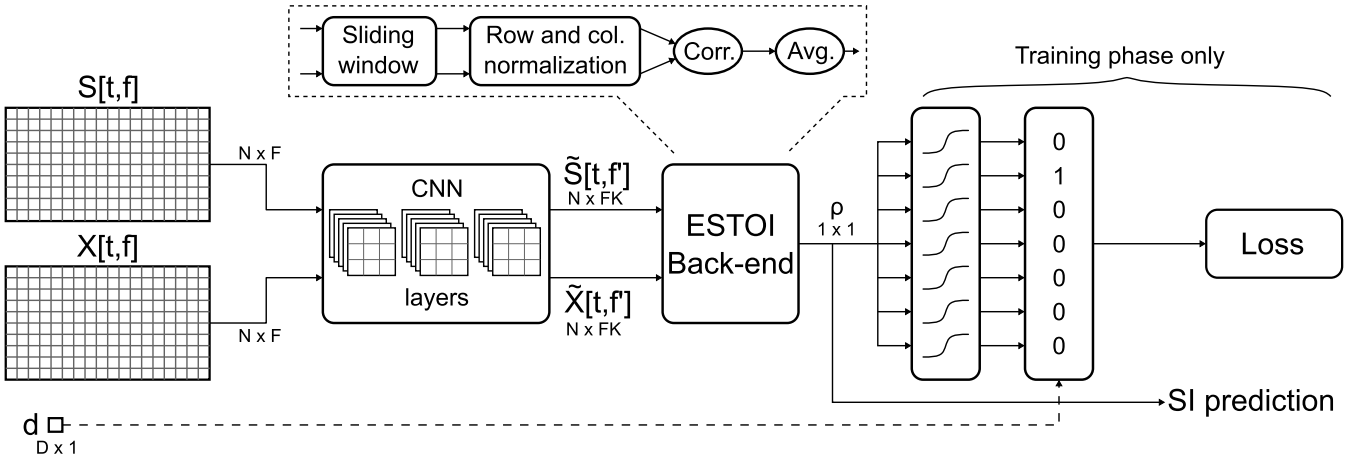

FIGURE 1. Proposed SI prediction architecture. From left to right are the 1/3 octave band inputs $S[t,f]$ of clean speech and $X[t,f]$ of noisy/processed speech, the dataset selection vector, d, used to choose the mapping function matching the listening test, the CNN layers which are applied to both $S[t,f]$ and $X[t,f]$ using the same kernels, yielding $\tilde{S}[t,f]$ and $\tilde{X}[t,f]$, the ESTOI back-end consisting of normalization and correlation performed on a sliding window, and an average across frames resulting in the SI index prediction, ρ. In the training phase, the SI index, ρ, i.e., the output of the ESTOI back-end, is mapped to a prediction of absolute SI using the logistic function indicated by d.

distinction is that NISA is trained using labels generated by STOI rather than a listening test. This circumvents the limitations imposed by the scarcity of listening test data, but also imposes the performance of STOI as an upper bound on the performance of NISA. Other examples include the predictors described in [13], a convolutional neural network emulating the STI, and [12] a hidden Markov model emulating STOI.

The data-driven methods proposed by [12], [14], and [16] are not evaluated on unseen conditions. The methods proposed by [11] and [13] are tested on unseen conditions, though these conditions are in the same category as the seen data, additive noise for [11], and reverberation from convolution with room impulse responses for [13]. Furthermore, these methods were trained using labels generated by classical predictors, STOI and STI for [11] and [13] respectively, rather than measured SI. Finally, the methods proposed in [15] and [22] were tested on unseen datasets, revealing highly dataset dependent performance.

III. ARCHITECTURE AND MAPPING FUNCTIONS

The data-driven SI-predictor proposed in this paper is a Convolutional Neural Network (CNN) with inspirations from ESTOI [9]. The architecture is shown in Figure 1. The model takes two inputs: a potentially noisy and/or processed speech signal, $X[t,f]$, and the corresponding time-aligned clean speech signal, $S[t,f]$. In the training phase the model is also given a third input, the paradigm selector vector, d, which is a vector with a 1 in the entry corresponding to the listening test from which the training sample, i.e., $X[t,f]$ and $S[t,f]$, was drawn, out of a total set of D listening tests used for training. This vector is used to select the appropriate mapping function. Spectrograms $X[t,f]$ and $S[t,f]$ are 1/3 octave band representations of the time-domain speech signals $x[\tau]$ and $s[\tau]$, respectively. To obtain $X[t,f]$ and $S[t,f]$, both $x[\tau]$ and $s[\tau]$ are resampled to 20 kHz. Then a Short-Time Fourier Transform (STFT) is performed, followed by a 1/3 octave band transform, similar to that of ESTOI, yielding $X[t,f]$ and $S[t,f]$. For the STFT a 50% overlapping Hann window, W samples in length, and zero padding to $2W$ samples is used. The input signals are processed in a number of CNN layers, followed by an ESTOI back-end, which performs the comparison between the signal under study and the clean reference. During network training, the output of the ESTOI back-end is mapped to absolute SI by the mapping function corresponding to the listening test from which the inputs and SI label were obtained.

A. NETWORK DESIGN

The goal in designing the network is to increase robustness to unseen datasets. The architecture is designed to be relatively small, in order to mitigate overfitting to the seen datasets. The proposed architecture has fewer than 10^4 trainable parameters, whereas network sizes used in [22] range from 10^5 to 10^6 parameters. These large models showed signs of overfitting, as the performance was drastically lower for certain unseen datasets. This is also the reason why we have chosen to incorporate part of ESTOI into the network, i.e., to reduce the required number of trainable parameters.

We choose ESTOI, specifically, because of its simplicity and performance. The ESTOI back-end provides an anchor point of performance, in that the network should be able to perform at least as well as ESTOI on the training set. Hence, with this network design we expect performance on par with or better than ESTOI for seen conditions. The trainable part of the network, i.e., the CNN layers, is placed before the ESTOI normalization for a number of reasons. First, it guarantees that when the studied signal is in fact clean, i.e., $x[\tau] = s[\tau]$, the predicted SI is maximized. This is due to the fact that $x[\tau]$ and $s[\tau]$ are subject to the exact same mathematical operations, because they share the same CNN layers.

A CNN architecture was chosen, because it allows processing of variable lengths of input signals [15], and because CNN's have proven efficient for speech processing tasks in general, see e.g., [15], [22]–[24] or [25]. In a preliminary experiment we tested other architectures, particularly including a trainable weighted averaging across frequency bands in the back-end. This weighted average was found to have no significant impact on performance. The proposed training procedure is not limited to the architecture described here. It can be applied to the training of any data driven SI-predictor.

B. 1/3 OCTAVE BAND TRANSFORM

The 1/3 octave band transform is applied as presented in [2]. First, the STFT, given by:

$$\hat{X}[t,k] = \frac{1}{\sqrt{2\pi}} \sum_{s=0}^{W-1} x\left[\frac{tW}{2} + s\right] \mathrm{w}[s]\, \mathrm{e}^{-jks}, \tag{1}$$

is applied, where $\hat{X}[t,k]$ is the STFT of x at time t, and frequency k, $\mathrm{w}[\cdot]$ is a Hann analysis window of length W, and j denotes the imaginary unit. Then, the magnitudes of each 1/3 octave band are computed as follows:

$$X[t,f] = \sqrt{\sum_{s=k_l[f]}^{k_h[f]} |\hat{X}[t,s]|^2}, \tag{2}$$

where $X[t,f]$ is the 1/3 octave band representation of x at time t, and 1/3 octave band f, and where $k_l[f]$ and $k_h[f]$ are the indices of the lowest and highest frequency bands of $\hat{X}$ within the f'th 1/3 octave band. Similar operations are applied to $s[\tau]$ to obtain $S[t,f]$. For more details we refer to [2].

C. CNN LAYERS

The 1/3 octave band transformed signals, $X[t,f]$ and $S[t,f]$, are now run independently through the same CNN layers, cf. Figure 1. We use L CNN layers of K kernels with Rectified Linear Unit (ReLU) activation functions. The signals are zero padded to preserve their size after each convolution.

D. ESTOI BACK-END

The CNN layers produce K outputs, $\tilde{X}[t,f,0], \ldots, \tilde{X}[t,f,K-1]$, each corresponding to one kernel in the final layer. These K outputs are concatenated along the frequency-axis:

$$\tilde{X}[t,f'] = \Big[\tilde{X}[t,f,0] \ldots \tilde{X}[t,f,K-1]\Big], \tag{3}$$

where f' is used to index the new concatenated frequency axis. This concatenation results in a computationally convenient representation of $\tilde{X}$ for the next step. Following the CNN layers are a series of operations from ESTOI, as illustrated in the details of the "ESTOI Back-end" in Figure 1 [9]. A sliding rectangular window, N samples wide, is applied along the temporal axis, splitting the input spectrograms into short overlapping matrices. The n'th of these matrices is given by:

$$\tilde{X}_n[t,f'] = \Big[\tilde{X}[n,f']^\top \ldots \tilde{X}[n+N-1,f']^\top\Big]^\top. \tag{4}$$

For each n, $\tilde{X}_n[t,f']$ is normalized across time and frequency as follows. First, the mean is subtracted across time:

$$\tilde{X}_{n,2}[t,f'] = \tilde{X}_n[t,f'] - \frac{1}{N}\sum_{s=0}^{N-1} \tilde{X}_n[s,f']. \tag{5}$$

Then, the variance is normalized across time:

$$\tilde{X}_{n,3}[t,f'] = \tilde{X}_{n,2}[t,f'] \Big/ \sqrt{\sum_{s=0}^{N-1} \tilde{X}_{n,2}^2[s,f']}. \tag{6}$$

Now, the mean across frequency is subtracted:

$$\tilde{X}_{n,4}[t,f'] = \tilde{X}_{n,3}[t,f'] - \frac{1}{N}\sum_{s=0}^{F-1} \tilde{X}_{n,3}[t,s]. \tag{7}$$

Finally, the variance is normalized across frequency:

$$\tilde{X}_{n,5}[t,f'] = \tilde{X}_{n,4}[t,f'] \Big/ \sqrt{\sum_{s=0}^{F-1} \tilde{X}_{n,4}^2[t,s]}. \tag{8}$$

$\tilde{S}_{n,5}[t,f']$ is computed similarly. The correlation coefficient between each corresponding matrix of the noisy/processed and clean speech signals is now given by:

$$\rho_n = \frac{1}{N}\sum_{t=0}^{N-1}\sum_{f=0}^{F-1} \tilde{X}_{n,5}[t,f']\tilde{S}_{n,5}[t,f']. \tag{9}$$

The average across frames, ρ, of these correlation coefficients is the output of the network.

E. DATASET-SPECIFIC MAPPING FUNCTIONS

Ideally, the trained SI index predictor should be independent of the paradigms specific to the listening tests included in the training data. To achieve this, we append a number of DSMF's to the architecture used exclusively for the network training and validation phases. This is marked as "Training phase only" in Figure 1. During network training, an additional input is given. This input, d, is a vector with a 1 in the entry corresponding to the index of the dataset from which the inputs $s[t]$ and $x[t]$ originate, and 0's in all other entries. The DSMF's used in this study are logistic functions, defined as:

$$\sigma(x) = \frac{1}{1+\mathrm{e}^{-(ax+b)}}, \tag{10}$$

where x is the input, and a and b are the trainable parameters. Conveniently, computing these functions corresponds to a single fully connected layer, with a number of nodes equal to the number of listening tests, followed by a sigmoid activation function. The parameters a and b then, respectively, correspond to the weights and biases of the fully connected layer. This layer is designed to apply all the DSMF's to the network output in parallel during training, which is represented

TABLE 1. Overview of the datasets used for training and testing of the proposed SI predictor. The datasets have been split into files of equal duration of approximately 6.6s of speech. The column labelled #subj. list the number of participating listeners, and the column labelled #cond. lists the number of different listening conditions resulting from the various noise types and SNR's as well as processing types and settings.

Dataset		Size			Content	
No.	Ref.	#subj.	#files	#cond.	Speech material	Noise & processing types
DS0	[26, Sec. VI]	11	278	9	Dantale II (closed)	BBL, Beamforming
DS1	[27, Sec. III-C]	14	241	20	Dantale II (closed)	BFN, ITFS
DS2	[9, Sec. IV-1]	12	684	60	Dantale II (closed)	Noisex, SSN, BBL, Temporal modulation
DS3	[28, Sec. II]	15	7808	168	Dantale II (open)	SSN, BBL, café, car, ITFS
DS4	[29, Sec. IV]	9	64	52	Dutch Hagerman test (closed)	SSN, pre-noise enhancement
DS5	[2, Sec. III-C]	7	390	35	IEEE database (open)	SSN, BBL, ITFS w. artificial errors
DS6	[15, Sec. III-D_4]	8	2139	327	ADD (open)	SSN, Low- and high-pass filtering
DS7	[30, Sec. III]	15	976	24	CLUE database (open)	ICRA, Speech segregation
DS8	[31, Sec. III-B]	16	547	18	Dutch Hagerman test (closed)	SSN, BBL, pre-noise filtering
DS9	[9, Sec. IV-5]	13	4333	20	Dutch Hagerman test (closed)	SSN, Single channel noise reduction

by the block filled with s-shaped curves in Figure 1. The inner product is now taken between the outputs of the fully connected layer and the selector vector d, in order to select the relevant DSMF. Thus, only the DSMF corresponding to the dataset indicated by d is passed through this operation. This particular implementation was chosen because it is differentiable, which allows for back propagation. In this way, the network can be simultaneously trained on multiple pooled datasets, while the mapping functions absorb the different psychometric functions, which the network could otherwise only account for by over-fitting. The choice to use logistic functions as DSMF's is inspired by the fact that logistic functions are often used to model psychometric functions for classical SI-predictors, see e.g., STOI [2], ESTOI [9], SIIB [3], CNN [15], SII [4] or the survey of psychometric functions for SI in [1]. Importantly, because of the choice to train with logistic DSMF's it can be expected that the network outputs SI-indices that are logistically related to absolute SI.

The DSMF training procedure is designed to give the network a parameter efficient way to represent the psychometric functions that arise from the training data. The psychometric functions are thus learned separately from the CNN, which means that the internal parameters of the network can be utilized more efficiently, leading to better SI-prediction performance even though the DSMF's themselves are discarded in the end.

The result of the proposed DSMF training procedure is a network that outputs an unmapped SI-index, ρ, which correlates highly with absolute SI. In practice, for unseen data, ρ would be used as the SI-prediction. In general, SI-indices produced by this network are not predictions of absolute SI, because the network does not account for psychometric functions. In special cases, however, where the listening test paradigm is known, i.e., when the data comes from a known listening test, the corresponding DSMF could be appended to produce predictions of absolute SI. In the interest of facilitating a fair comparison with competing predictors, however, we will not be using the trained DSMF's in the test phase.

IV. DATASET DESCRIPTION AND TRAINING PROCEDURE

A. DATASETS

The experiments described in this paper are based on a pooled dataset consisting of the results from ten listening tests. Table 1 describes the datasets with a few keywords pertaining to the speech material, noise types and processing in each listening test. The noisy/processed speech stimuli, $x[t]$, and the clean reference signals, $s[t]$, from each noise/processing condition in each listening test were extracted. The label for each pair of signals was taken to be the average fraction of correct words across all listeners within the given condition. It would have been desirable to use more granular SI-labels, e.g., binary labels indicating whether each individual word was correctly identified in the corresponding listening test. However, for the vast majority of the datasets we use, particularly DS3 through DS9, only the average SI is available. For the sake of consistency, we use the average SI labels for all datasets.

All ten listening tests were conducted with normal hearing native speakers. The listening tests were either conducted with a closed set, which allowed participants to select each word from a list, or an open set, which required the participants to either write down or repeat each word without a list of candidate words. For more detailed descriptions of the datasets and listening tests, we refer to the respective sources listed in Table 1. In Table 1, Dantale II refers to the Danish matrix test speech corpus described in [32]. ADD refers to Akustiske Databaser for Dansk,[2] which contains meaningful Danish sentences. CLUE refers to the Danish speech corpus described in [33]. The Dutch Hagerman matrix test speech corpus is described in [34], and the IEEE database contains English speech. These speech datasets each contain speech signals from a single talker, apart from ADD, which contains speech signals from multiple talkers. The Noisex database is described in [35], and contains various recorded noise types. Speech shaped noise (SSN) refers to white Gaussian noise, filtered to match the long term spectral envelope of speech. Babble (BBL) noise refers to the mixture of a number of competing talkers. The number of competing talkers varies from 2 to 20 depending on the dataset. Bottle factory noise (BFN) refers to recorded noise of bottles clinking against each other on a conveyor belt. ICRA is a database of noise signals, constructed to mimic the short term modulations of speech [36]. Ideal time-frequency segregation (ITFS) is a method for enhancing a signal in the time-frequency

[2]http://www.nb.no/sbfil/dok/nst_taledat_dk.pdf

domain by utilizing the true signal to noise ratio (SNR) for each time-frequency tile, in order to, for example, compute ideal gains or cut-off thresholds [37]. The signals in DS6 have been recreated using a different speech database than the one used in the original listening test, the full details and verification experiments can be found in [15, Sec. III].

B. TRAINING

Each dataset was split randomly into 80% training, 10% validation and 10% test data. This was done to ensure that all datasets would be represented in the test set. The data was partitioned into training samples of equal duration, to enable the construction of mini batches. The duration of 512 frames, corresponding to approximately 6.6 seconds, which is long enough to accommodate one to two sentences, was chosen. This fixed duration resulted in some training samples spanning two listening test conditions. The labels for these samples were computed as the weighted average of the measured SI for those two conditions, with weights equal to the number of frames from each condition in that training sample. A batch size of 32 was found to give the best compromise between GPU-memory, training speed and end-performance. The network was trained on batches from the training dataset using the Adam optimizer [38], and the Mean Squared Error (MSE) loss function. An early stopping scheme was used, where the learning rate was halved for every 25 epochs without a new global minimum in validation cost, and the training was stopped early if this continued for 35 epochs. Training was allowed to proceed for a maximum of 300 epochs. Training of the models with DSMF involves forward passing training samples, i.e., triplets of $X[t,f]$, $S[t,f]$ and d, through the CNN layers, the ESTOI back-end and finally the DSMF's, after which the loss function is evaluated. For the test phase, the trained DSMF's are discarded. To take the psychometric functions into account for the evaluation, logistic functions are fitted for each listening test in the test data by least squares for all the evaluated predictors. This is done in order to facilitate a fair comparison between the DSMF trained networks and the classical predictors. This also allows the DSMF trained networks to be tested on unseen datasets. The architecture was implemented using Tensorflow 2.1 [39].

C. NETWORK PARAMETERS

We trained the networks with the following parameters. The window length of the 1/3 octave band transform is $W = 512$. A preliminary experiment showed that $L = 3$ CNN layers with $K = 20$, 3×3 kernels resulted in the best performance. Networks with 1, 2, 3 and 4 CNN layers and 5, 10, 15 and 20 kernels per layer were tested. Due to memory constraints, we were unable to test higher numbers of kernels. The window length of the ESTOI back-end is $N = 30$, cf. [9]. The lowest 1/3 octave band is centred around 150 Hz, and the highest around 6050 Hz, for a total of $F = 17$ bands. This is an increase from the conventional ESTOI, which uses 15 bands. According to the band importance function of the SII, [4], this frequency range accounts for most of the intelligibility of speech. In total the architecture has 7, 460 trainable parameters.

V. PERFORMANCE EVALUATION

Two experiments, A and B, are performed to investigate the properties of the proposed DSMF training strategy and the resulting SI predictor. In Experiment A the goal is to validate that the DSMF's absorb the information related to the different psychometric functions, and result in improved prediction performance over plain pooling with no DSMF's. In Experiment B the goal is to investigate the robustness of the network and training method to new or unseen listening conditions and test paradigms. The Spearman and Pearson correlation coefficients, along with the Mean Squared Error (MSE) values, are used as evaluation metrics.

A. EXPERIMENT A

In order to evaluate the efficacy of the DSMF training procedure, models were trained both with and without DSMF. Both models have the same number of parameters in the CNN layers, but since the DSMF's should be able to represent the psychometric functions of each dataset, we expect the DSMF trained model to utilize these parameters more efficiently. As a result, the DSMF trained model is expected to reach higher performance than the one without DSMF. These models are both tested against each other, and against a variety of classical predictors, i.e., ESTOI, SIIB, HASPI, STOI and SI-SDR. We remind the reader that the trained DSMF's are not used in the test phase. Instead, as part of the evaluation of the performance of each predictor on the test data, logistic psychometric functions are fitted to the test data using least squares, and used to transform the outputs of the predictors to absolute SI, before computing the Spearman and Pearson correlations as well as the MSE values. These logistic functions should not be confused with the trained DSMF's, and we stress that they are solely used as part of the evaluation of the SI-predictors, facilitating the comparison between predictions and measured absolute SI. This has no impact on the Spearman coefficient, because it is invariant under monotonically increasing transforms, i.e., the fitted logistic functions. It does affect the Pearson correlation and MSE, however, since the logistic fitting attempts to map predictions onto a straight line, which should increase the Pearson correlation, and reduce the MSE. This facilitates a fair comparison between the trained and classical predictors, and better reflects the performance that can be expected in practice. Specifically, if-hypothetically-the trained DSMF's were used in the test phase, the proposed network might have an advantage specific to the datasets used in this work, but this advantage would not generalize, since trained DSMF's only exist for seen datasets.

Table 2 shows the Spearman correlations, as computed dataset-wise, for each predictor. The predictors trained in this experiment are marked with (seen) in Table 2. The prediction for each listening test condition was made by concatenating all the speech signals available in the test set for that particular

TABLE 2. Spearman correlations between the mapped predictions and the measured SI of the 10 datasets. In the (unseen) rows, each column represents a different permutation of training, validation and test data, where the corresponding dataset has been excluded from training and validation. The rightmost column shows the average of the Spearman coefficients across the datasets.

Spearman ×100

Predictor	DS0	DS1	DS2	DS3	DS4	DS5	DS6	DS7	DS8	DS9	Average
Net w. DSMF (seen)	97.62	**97.94**	86.37	**97.51**	89.66	66.78	93.46	68.43	88.85	**98.61**	**88.52**
Net w.o. DSMF (seen)	95.24	89.47	**89.26**	97.06	85.65	61.50	92.18	71.13	85.35	97.56	86.43
Net w. DSMF (unseen)	97.62	92.78	80.19	89.39	88.82	64.00	90.02	60.52	**97.52**	95.64	85.65
Net w.o. DSMF (unseen)	97.62	92.36	70.77	82.56	85.67	51.11	**93.88**	58.52	90.51	97.74	82.07
STOI	97.62	94.22	36.29	94.70	88.70	54.16	81.47	61.57	91.95	98.31	79.90
ESTOI	97.62	97.73	85.50	95.07	89.39	41.21	87.66	62.09	87.20	96.73	84.02
SIIB	**100.00**	96.70	79.28	91.05	**93.78**	35.52	92.33	**76.43**	92.78	97.93	85.58
HASPI	-2.38	63.88	68.13	68.79	62.02	38.03	85.65	65.91	2.37	82.21	53.46
SI-SDR	-83.83	91.74	54.83	43.73	29.28	**84.75**	66.43	66.43	33.54	95.75	48.27

TABLE 3. Pearson correlations between the mapped predictions and the measured SI of the 10 datasets. In the (unseen) rows, each column represents a different permutation of training, validation and test data, where the corresponding dataset has been excluded from training and validation. The rightmost column shows the average of the pearson coefficients across the datasets. The values marked with * are not significantly different compared to the best predictor on the given dataset.

Pearson ×100

Predictor	DS0	DS1	DS2	DS3	DS4	DS5	DS6	DS7	DS8	DS9	Average
Net w. DSMF (Seen)	**99.68***	**97.95***	87.10*	**98.07***	93.67*	64.04	**93.20***	84.06*	91.85	99.00*	**90.86**
Net w.o. DSMF (Seen)	99.21*	89.76	**88.95***	96.65	91.15	60.56	91.93	82.84*	83.55	98.53*	88.31
Net w. DSMF (Unseen)	99.63*	93.49	78.32*	89.18	92.09	62.07	88.63	75.93	**96.95***	96.57*	87.28
Net w.o. DSMF (unseen)	98.93*	93.30	69.01	82.82	90.48	47.36	90.83	55.58	92.07*	98.21*	81.86
STOI	99.20*	92.11	34.16	93.99	92.62	49.64	80.47	76.69	92.34*	**99.01***	81.02
ESTOI	99.54*	97.79*	84.78*	94.61	90.77	36.05	86.47	76.34	83.73	97.52*	84.76
SIIB	99.18*	94.46	79.23*	89.94	**95.86***	29.37	91.44	**91.79***	94.91*	96.95*	86.31
HASPI	1.39	60.17	61.67	67.19	69.75	22.45	84.57	32.40	1.67	78.87	48.01
SI-SDR	-32.07	92.59	46.65	41.67	25.79	**83.34***	60.37	73.14	34.48	97.15*	52.31

TABLE 4. Mean squared error between the mapped predictions and the measured SI of the 10 datasets. In the (unseen) rows, each column represents a different permutation of training, validation and test data, where the corresponding dataset has been excluded from training and validation. The rightmost column shows the average mean squared error across the datasets. Note that all mean squared errors in this table have been scaled by a factor of 100 for better formatting.

Mean squared error ×100

Predictor	DS0	DS1	DS2	DS3	DS4	DS5	DS6	DS7	DS8	DS9	Average
Net w. DSMF (Seen)	**0.052**	**0.335**	2.355	**0.420**	0.976	3.868	1.549	1.135	0.496	**0.075**	**1.126**
Net w.o. DSMF (Seen)	0.127	1.587	2.023	0.729	1.334	4.146	1.811	1.209	0.959	0.109	1.404
Net w. DSMF (Unseen)	0.060	1.028	3.218	2.247	1.204	4.026	2.509	1.634	0.191	0.253	1.637
Net w.o. DSMF (unseen)	0.171	1.060	4.365	3.438	1.439	5.081	2.059	2.631	0.483	0.133	2.086
STOI	0.127	1.238	6.430	0.827	1.143	4.361	1.673	1.539	**0.182**	0.088	1.761
ESTOI	0.071	0.361	**0.987**	1.071	1.401	5.370	1.959	1.663	0.566	0.162	1.361
SIIB	0.135	0.899	2.007	2.145	**0.619**	5.707	**0.797**	**0.488**	0.235	0.252	1.328
HASPI	7.755	5.212	6.046	5.995	4.072	6.217	3.334	3.399	3.168	1.416	4.661
SI-SDR	0.460	1.173	5.867	9.000	7.332	**2.009**	7.190	1.793	2.807	0.208	3.784

condition, resulting in one pair of inputs for each condition. These pairs were given to the predictors as inputs yielding one scalar SI-prediction per condition as output. The correlations between the predictions of all conditions within each dataset and the corresponding measured SI from the listening tests were then computed. The performance in terms of Pearson correlations is computed in a similar way and seen in Table 3. Additionally, the mean squared error of each predictor is reported in Table 4. We noticed no loss in performance as a consequence of the relatively longer test signals. This is likely because the architecture has a very small receptive field because of the small kernels in the CNN layers.

As expected, the DSMF trained network reaches a higher performance than the non-DSMF trained network in terms of

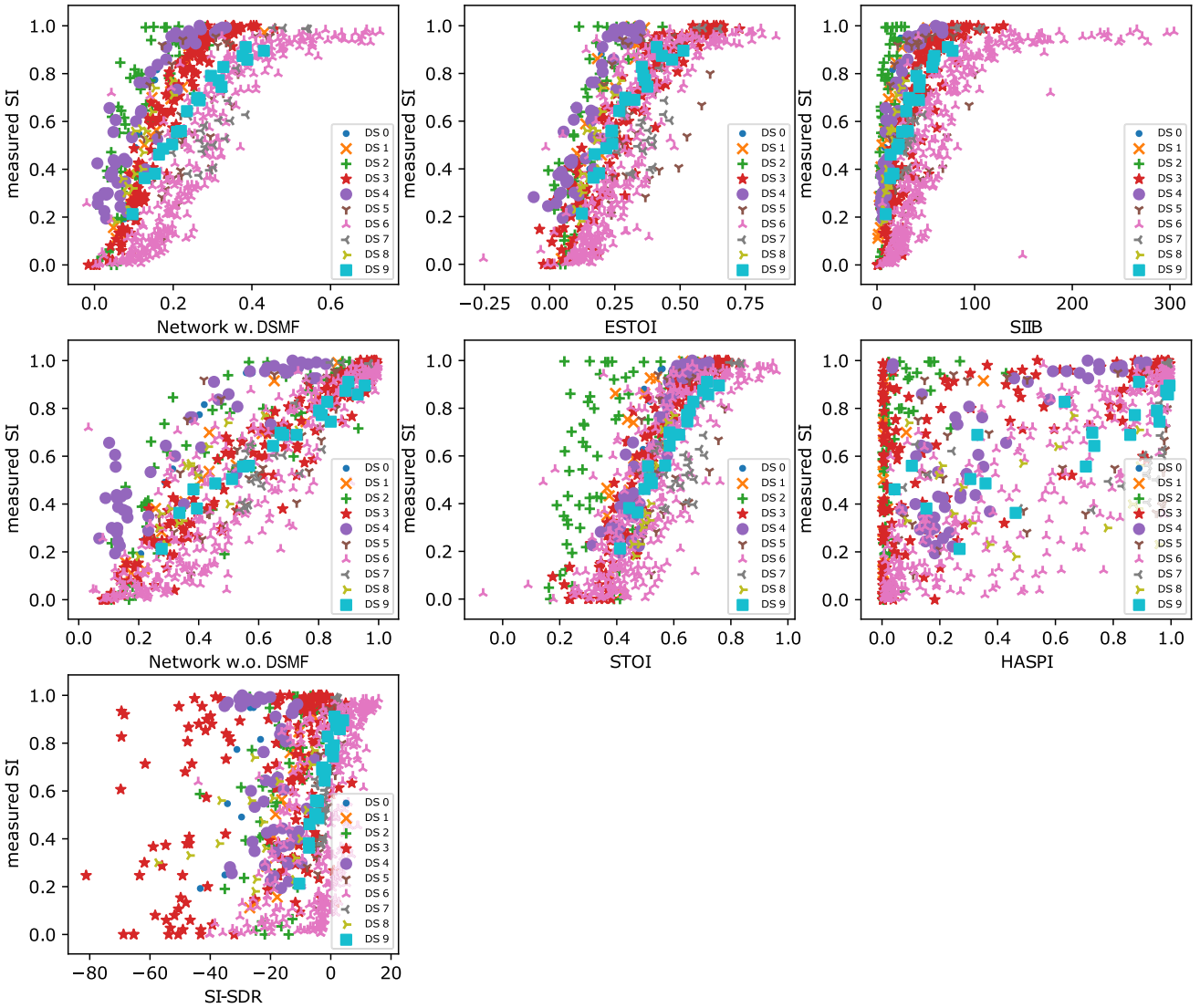

FIGURE 2. Experiment A: Absolute measured SI vs. raw SI indices output by the networks and classical predictors, i.e., with no fitted logistic functions.

both Spearman and Pearson correlation. Since the only difference between these two networks is the presence of DSMF's during training, it is clear that training with DSMF's has a positive effect on the final performance of the SI-predictor, indicating that the DSMF's are working as intended. In particular, the performance average across datasets is higher with DSMF. The exceptions, where the non-DSMF model performs better are DS2 and DS7. A possible explanation for this could be that networks sacrifice performance for some datasets in order to perform better on average. DS2 is a difficult dataset for many SI-predictors, because it contains temporally modulated noise [9]. Note also that STOI performs poorly, whereas ESTOI does particularly well on this dataset. This is an expected result, as ESTOI was proposed in order to improve STOI's performance on speech in temporally modulated noise, and evaluated using DS2 [9]. Figures 2 and 3 show scatter plots of measured SI and predictions, each point representing one listening test condition. Figure 2 shows the raw predictions, or SI indices, i.e., before logistic functions are fitted, vs. measured absolute SI for the various SI predictors. From Figure 2, the proposed SI predictor, ESTOI and SIIB manage to produce fairly concentrated clusters of predictions, whereas STOI and HASPI struggle to do so. This is also reflected in Tables 2 and 3. For DS2 specifically, the predictions show a much wider spread at the high end of the SI-spectrum. This is consistent with the observations in [9] that many SI-predictors tend to underestimate the SI in this dataset. In DS7 there are very few conditions at the extreme ends of the measured SI spectrum, i.e., 0 and 1, where prediction errors are generally smaller. This could explain why many of the SI predictors score relatively low on this dataset. In Figure 2 it can be seen that the network trained without DSMF produces indices with an approximately linear relation to absolute SI, whereas the

network trained with DSMF produces indices with separate, approximately logistic relations to absolute SI. This clearly illustrates the difference between training with and without DSMF; the non-DSMF trained network must necessarily be dedicating internal parameters to recognizing and mapping each of the datasets to absolute SI, i.e., the network has specialized to the training data. Recall that the psychometric functions cannot generally be determined from the network inputs alone. The network trained with DSMF, however, does not appear to have any internal representation of the psychometric functions of the datasets, since each dataset forms a separate s-shaped cluster, indicating that the DSMF's were able to absorb the different psychometric functions of the training data.

Among the classical predictors, ESTOI and SIIB have the best performance, which is in accordance with existing studies, see e.g., SIIB [3] or ESTOI [9]. While the classical predictors are not primarily data-driven, some of the datasets we test on, were used in the development of the classical predictors. Specifically DS3, DS5 and DS9 were used in the development of STOI, DS2, DS3 and DS9 in the development of ESTOI [9], and DS3, DS4 and DS9 in the development of SIIB [3]. This is reflected in the performance of these predictors on those respective datasets, as seen in Tables 2, 3 and 4, where e.g., STOI reaches a Spearman correlation of 0.54 on DS5. These observations are well in line with conclusions drawn in [40] that SI-predictors tend perform better on datasets used during their development. HASPI and SI-SDR show the lowest performance on average. SI-SDR shows drastic variation in performance from one dataset to the next, with high performance on DS1, DS5, DS7 and DS9, and low performance on DS0, DS2, DS3, DS4 and DS8. Note in particular the negative Spearman coefficients on DS0. This negative correlation could be due to the relatively few conditions in DS0, which means that fewer discordant pairs are necessary to significantly reduce the score. Note that high correlations with different signs may be detrimental to any SI-predictor: In order to be reliable in practice, it must be clear whether an increase in predictor output is indicative of an improvement or a decline in SI.

In the case of HASPI this can be attributed to slightly lower scores on most of the datasets and very low scores on DS0 and DS8 in particular. HASPI's very low score on DS8, might also be explained by the fact that DS8 has few conditions.

Figure 2 demonstrates the difference between training with and without DSMF's. In particular, the network trained without DSMF's attempts to force predictions from all the datasets onto the same line between $(0, 0)$ and $(1, 1)$. This is a clear indication that the network has learned an internal representation of the psychometric functions specific to the training datasets. As a consequence, the predictions show a substantial variance. When trained with DSMF's, however, the outputs related to different datasets form separate s-shaped clusters. The differences between these clusters are a result of the paradigm differences, meaning that the network has not learned an internal representation of the psychometric functions of the training data. It is evident that this has resulted in substantially reduced variance in the predictions. Note that the clusters corresponding to each dataset, appear similar for this DSMF trained network and for ESTOI, which could be a result of the similarities between the proposed architecture and ESTOI. These similarities are further evidence that the different clusters represent different psychometric functions, since ESTOI does produce SI indices that map to absolute SI via a logistic psychometric function [9]. Given the similarities between the proposed architecture and ESTOI, it is not surprising to see similarities in their psychometric functions as well.

Figure 3 shows the same results, but each dataset has now been mapped to absolute SI using logistic psychometric functions fitted by least squares. Note that these logistic functions are not the trained DSMF's. These logistic functions are fitted to the test data, as opposed to the DSMF's that are fitted to the training data. The DSMF's are also trained jointly with the network, whereas these logistic functions are only fitted after the network has been trained. Thus, the predictions now ideally cluster around the diagonal line from $(0, 0)$ to $(1, 1)$. From this figure it is easier to compare the performance across the different predictors because the predictions can now be considered absolute SI predictions, rather than SI indices. For instance, the DSMF network appears to be better at predicting low intelligibility than the non-DSMF network, as the clustering is tighter near $(0, 0)$. This could be because DSMF allows training to focus on tightening each dataset cluster, tighter clusters being equivalent to higher precision in predictions, rather than spending degrees of freedom on bringing all the clusters together. In other words, the DSMF network learns to predict SI indices for each dataset, rather than absolute SI, and reaches better performance, because this task is simpler. As a result, the network trained with DSMF reaches the highest performance among the tested predictors.

The listening conditions associated with three sets of predictions with notable errors are listed in Table 5. The conditions are labelled 1-12 in Figure 3. These predictions come from datasets DS2, DS5 and DS6. In the case of DS2 there are

TABLE 5. Marked points from the scatterplot "Network w. DSMF" in figure 3.

Point	Dataset	Condition
1	DS6	High-pass 1122 Hz, -8 dB SSN
2	DS6	Low-pass 3458 Hz, 2 dB SSN
3	DS6	High-pass 1122 Hz, -2 dB SSN
4	DS6	Low-pass 2239 Hz, -2 dB SSN
5	DS6	High-pass 178 Hz, 0 dB SSN
6	DS5	Type-I 20 talker, error rate 0.8
7	DS5	Type-I 20 talker, error rate 0.6
8	DS5	Type-I 20 talker, error rate 0.7
9	DS5	Type-I 20 talker, error rate 0.4
10	DS2	-27 dB SNAM 2 Hz
11	DS2	-19 dB SNAM 8 Hz
12	DS2	-21 dB SNAM 4 Hz

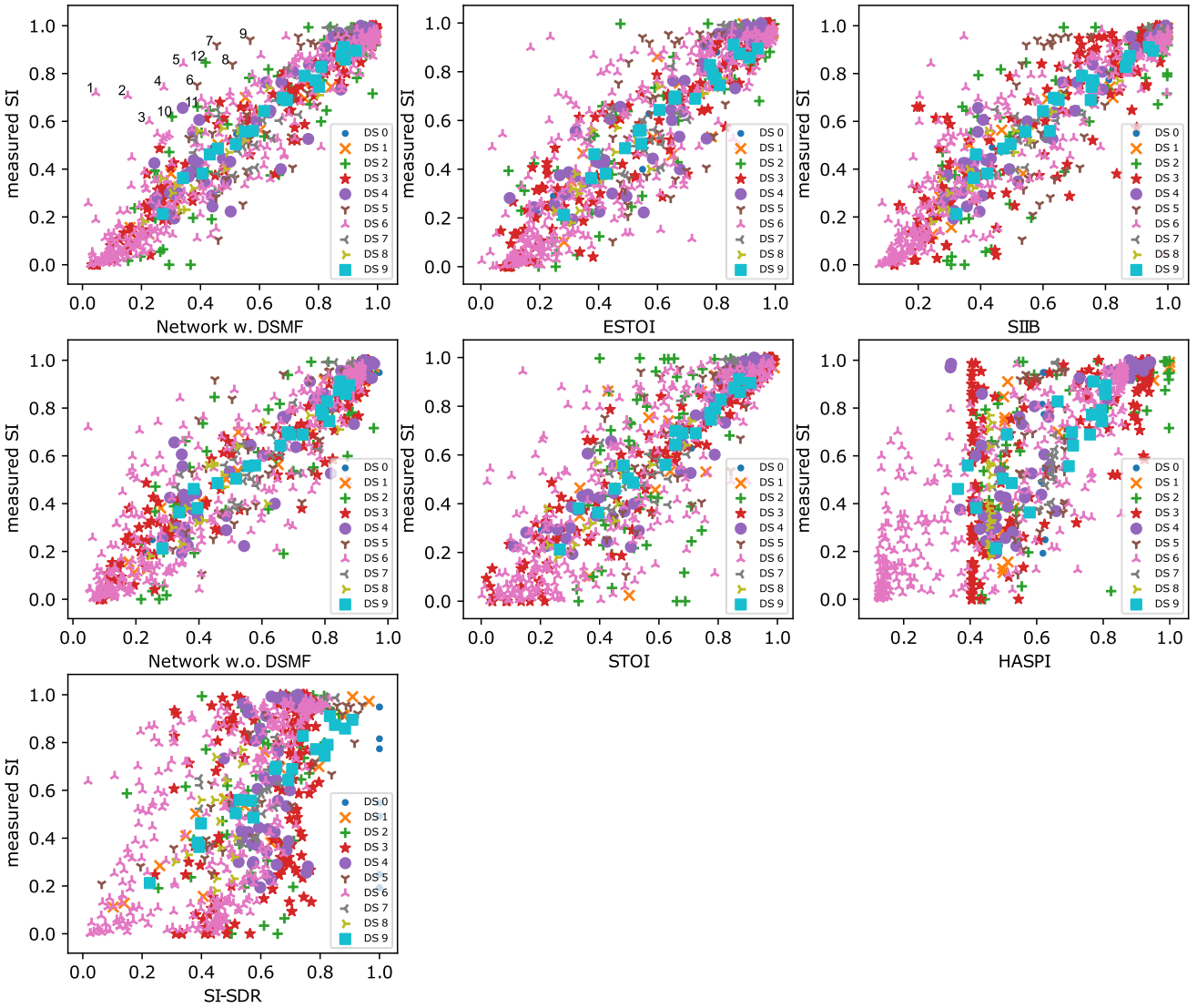

FIGURE 3. Experiment A: Absolute measured SI vs. SI-predictor output transformed by logistic functions fitted to each of the test datasets and predictors for both networks and classical predictors.

three listening conditions, all with the noise type Sinusoidal Noise Amplitude Modulation (SNAM) at various modulation frequencies and low SNR. A possible explanation for why the network struggles with this noise type could be that it is similar to the stationary noise type SSN, which appears very frequently in the training set. However, speech in SNAM may be significantly more intelligible than speech in SSN [9], because the modulated noise allows the listener to "listen in the dips", see e.g., [41] for more details. Looking at the points from DS5, they all come from the same processing scheme involving Ideal Binary Masked (IBM) speech. In particular, this listening test investigated the effect of artificial errors in an IBM speech enhancement system. In this context the Type I error listening condition, cf. Table 5, refers to IBM's where spectro-temporal gains of zero were converted to one, i.e., the enhancement system preserves too much of the noise. It is possible that the network overestimates the impact on SI of this extra noise, especially considering that this noise only appears in spectro-temporal regions which were noise dominated in the first place. For DS6 there does not appear to be any pattern in the listening conditions. The errors here could be due to the fact that the stimuli in this dataset were recreated using a different speech corpus from the original listening test [15].

B. EXPERIMENT B

In general, SI predictor networks should ideally be applicable to other types of listening conditions than the ones used during the training phase. The generalizability of the network proposed in this study is tested in a cross-validation experiment. In this experiment we train the network with ten different initializations on ten different partitions of training, validation and test data, i.e., one hundred networks trained in total. More precisely, we move the training and validation

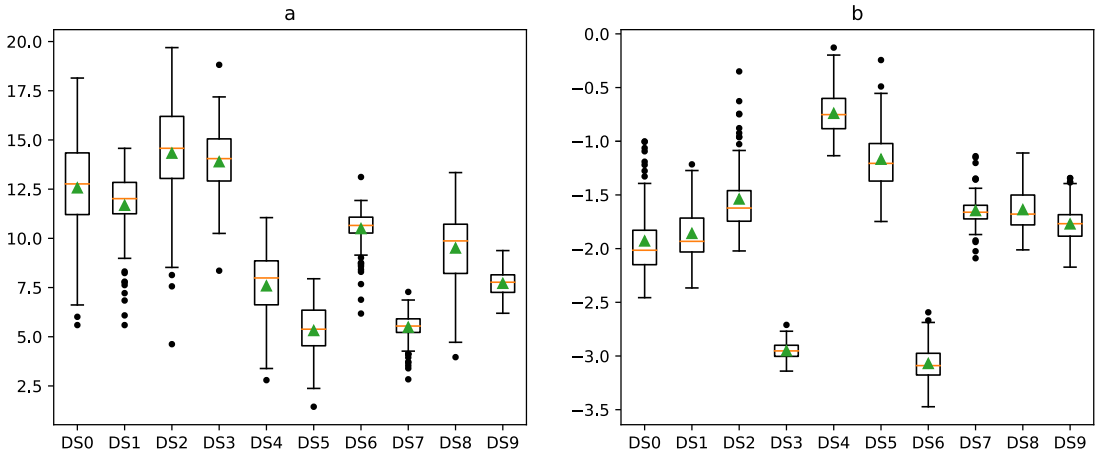

FIGURE 4. Boxplots of the trained DSMF parameters a and b described in eq. (10). The left figure shows a and right shows b, from the DSMF's of 90 differently initialized models. The green triangles denote the mean and the orange lines denote the median. The bottom and top of the boxes mark the 25% and 75% percentiles respectively. The black dots are outliers.

data from one listening test at a time entirely to the test set. This means that each listening test is excluded from the training and validation phases of ten models, and that the dataset is unseen when testing those models. For each partition, the model with the lowest validation loss was selected for the test phase. As such, this experiment gives an indication as to how the networks will perform in unseen conditions, and how they react to unseen listening conditions and test paradigms. As in Experiment A, we expect that the DSMF trained models will reach higher performance than the non-DSMF trained models. This is because the non-DSMF trained models learn an internal representation of the psychometric functions related to the training datasets. Since the psychometric functions related to unseen test data may be completely different from those related to the seen training data, such internal representations are undesirable.

Tables 2, 3 and 4 contain Spearman correlations, Pearson correlations and mean squared errors for the models. For any given dataset, the correlations and MSE values in the rows marked as Net (unseen) are computed for predictions made by a model with that dataset excluded from the training and validation sets. This means that each column describes a separate instance of the model, trained without access to the dataset corresponding to that column.

As expected, the performance for most of the datasets is lower when the dataset is unseen. The models experience the largest drops in performance on DS3, DS5 and DS6 as compared to when the datasets are seen. The reason for this could be that DS3 and DS6 are the largest and most diverse in terms of listening conditions. The exclusion of any of these datasets is a large reduction in the total amount of training data, which could result in the relatively larger loss of performance. Furthermore, large test sets also make it harder to produce a good ranking of a larger number of diverse conditions, as there are more opportunities for mistakes. As for DS5, judging by the relatively low scores, which the classical predictors achieve, it appears to be one of the hardest of these datasets for SI predictors in general. Despite the performance drop when this dataset is left out of training, the given model achieves higher performance than the classical predictors. Exceptionally, DS0, DS1 and DS8 have higher scores on the unseen models compared to the seen. These datasets all consist of few listening conditions, 20 or fewer. The explanation could be similar to the one for the large datasets, i.e., that the models simply perform better in general, when the training set is larger. Removing a small dataset from the training set, would then have only a small impact on performance.

Williams' t-test [42] was used to test for significant differences between the SI-predictors. This is a pairwise hypothesis test designed to detect significant differences in Pearson correlations. The null-hypothesis is that two different predictors have the same Pearson correlation with measured SI. Following the same procedure used in [9], we tested the highest performing predictor on each dataset against the others, and marked those not significantly different with * in Table 3. A significance level of $\alpha = 0.05$ with Bonferroni correction, to account for multiple tests, was used. Note that DS0, DS1, DS8 and DS9 contain 20 or fewer datapoints, i.e., listening conditions, which means that the t-tests could be unreliable on these datasets, according to [42].

On average, the unseen models score slightly higher than the classical predictors, which suggests that the proposed architecture and training scheme generalizes well and produces predictors which perform on par with, or better than the existing classical predictors for listening conditions, on which it has not been trained. We attribute this robust performance to two main factors. First, the proposed network contains as few as 7,460 trainable parameters, which mitigates overfitting. Secondly, the use of DSMF during training facilitates pooling of training data obtained from different listening

tests, effectively increasing the amount of listening test data available for training.

Performance of the proposed SI predictors, when tested with signals from listening conditions similar to those used for training the SI-predictor, is substantially better than existing methods. This improved prediction performance may be advantageous for replacing some listening tests in iterative development of speech processing systems. Assuming that the processing scheme, or the stimuli, are not changed too drastically, then the SI-predictor network can be validated or even retrained in order to benefit from the high performance on seen conditions.

Looking at DS5 in Table 2, there is a larger gap in performance between the (seen) and (unseen) models without DSMF's compared to the models with DSMF's. In particular, the difference in Spearman correlation is 0.0278, for the DSMF trained model and 0.1039, for the non-DSMF trained model. Noting that DS5 is the only dataset which contains English speech, this might be interpreted as the DSMF training successfully increasing the model's robustness to an unseen language. It should be noted, however, that language is not the only paradigm difference in DS5, so the drop in performance of the model without DSMF's might not only be due to the unseen language.

Figure 4 contains box-plots of the trainable parameters, a and b described in eq. (10), of the DSMF's belonging to the seen datasets, i.e., 90 maps per dataset. While there are significant outliers, depending on the initialization, the majority of the DSMF's for each dataset are very similar. This is evident from the boxes which contain the parameters from 50% of the initializations. This is more evidence that the DSMF's are in fact consistently used by the network to model specific information about each dataset. Since the DSMF's are trained jointly with their respective CNN's, variations in DSMF's can be compensated for by the CNN and vice versa, which means that a large spread of parameters, a and b, across initializations is not necessarily indicative of a similar spread in output predictions.

VI. CONCLUSION

We proposed and investigated a training strategy for data-driven speech intelligibility predictors, using dataset-specific mapping functions. The proposed strategy allows the use of pooled listening test datasets during network training, without specializing to the paradigms of those listening tests. Solving this problem is important, because training of data-driven SI predictors almost inevitably involves the use of listening test data obtained from multiple listening tests employing different paradigms. Without these proposed dataset-specific mapping functions, data-driven SI-predictors trained on pooled listening test datasets undesirably learn an internal representation of the psychometric functions particular to the listening test paradigms included in the training data. This can cause the trained SI-predictor to perform poorly, or even fail, when employed on new unseen data. To demonstrate this, ten listening test datasets were used to train, validate and test instances of a data-driven SI predictor using this training strategy. The dataset-specific mapping functions consisted of trainable logistic functions at the output of the architecture, which were designed to absorb the different psychometric functions of the datasets, thus preventing an inefficient internal representation of these functions from being learned. Experiments were designed to test the efficacy of training with these dataset-specific mapping functions, along with the generalizability of the predictor. Using the dataset-specific mapping functions for training and validation improved the test performance of the network. A cross validation experiment, where each dataset was excluded from the training set one by one, demonstrated that the network generalized well to new listening conditions and test paradigms, with performance on par with state of the art classical speech intelligibility predictors, for datasets that were not seen during training, and improved performance for seen datasets.

REFERENCES

[1] A. MacPherson and M. A. Akeroyd, "Variations in the slope of the psychometric functions for speech intelligibility: A systematic survey," *Trends Hearing*, vol. 18, Jun. 2014, Art. no. 2331216514537722.

[2] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, Dec. 2011.

[3] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "An instrumental intelligibility metric based on information theory," *IEEE Signal Process. Lett.*, vol. 25, no. 1, pp. 115–119, Jan. 2018.

[4] A. N. S. Institute, *Methods for Calculation of the Speech Intelligibility Index*, S3.5-1997. New York, NY, USA: ANSI, 1997.

[5] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Amer.*, vol. 19, no. 1, pp. 90–119, Jan. 1947.

[6] K. S. Rhebergen and N. J. Versfeld, "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *J. Acoust. Soc. Amer.*, vol. 117, no. 4, pp. 2181–2192, Apr. 2005.

[7] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 7, pp. 1766–1774, Sep. 2010.

[8] M. Elhilali, T. Chi, and S. A. Shamma, "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," *Speech Commun.*, vol. 41, nos. 2–3, pp. 331–348, Oct. 2003.

[9] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.

[10] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (HASPI)," *Speech Commun.*, vol. 65, pp. 75–93, Nov. 2014.

[11] D. Sharma, Y. Wang, P. A. Naylor, and M. Brookes, "A data-driven non-intrusive measure of speech quality and intelligibility," *Speech Commun.*, vol. 80, pp. 84–94, Jun. 2016.

[12] M. Karbasi, A. H. Abdelaziz, and D. Kolossa, "Twin-HMM-based non-intrusive speech intelligibility prediction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 624–628.

[13] P. Seetharaman, G. J. Mysore, P. Smaragdis, and B. Pardo, "Blind estimation of the speech transmission index for speech quality prediction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 591–595.

[14] K. Kondo, K. Taira, and Y. Kobayashi, "Binaural speech intelligibility estimation using deep neural networks," in *Proc. Interspeech*, Sep. 2018, pp. 1858–1862.

[15] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, "Nonintrusive speech intelligibility prediction using convolutional neural networks," *IEEE/ACM Trans. Audio, Speech, Langauge Process.*, vol. 26, no. 10, pp. 1925–1939, Oct. 2018.
[16] M. B. Pedersen, A. Heidemann Andersen, S. H. Jensen, and J. Jensen, "A neural network for monaural intrusive speech intelligibility prediction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 336–340.
[17] T. Houtgast and H. J. Steeneken, "Evaluation of speech transmission channels by using artificial signals," *Acta Acustica United Acustica*, vol. 25, no. 6, pp. 355–367, Dec. 1971.
[18] J. Jensen and C. H. Taal, "Speech intelligibility prediction based on mutual information," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 2, pp. 430–440, Feb. 2014.
[19] J. B. Allen, "The articulation index is a Shannon channel capacity," in *Auditory Signal Processing*. Cham, Switzerland: Springer, 2005, pp. 313–319.
[20] A. Leijon, "Articulation index and Shannon mutual information," in *Hearing–From Sensory Processing to Perception*. Cham, Switzerland: Springer, 2007, pp. 525–532.
[21] J. Taghia and R. Martin, "Objective intelligibility measures based on mutual information for speech subjected to speech enhancement processing," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 1, pp. 6–16, Jan. 2013.
[22] M. B. Pedersen, M. Kolbæk, A. H. Andersen, S. H. Jensen, and J. Jensen, "End-to-end speech intelligibility prediction using time-domain fully convolutional neural networks," in *Proc. INTERSPEECH*, Oct. 2020, pp. 1151–1155.
[23] A. Pandey and D. Wang, "A new framework for CNN-based speech enhancement in the time domain," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 7, pp. 1179–1188, Jul. 2019.
[24] S. R. Park and J. W. Lee, "A fully convolutional neural network for speech enhancement," in *Proc. Interspeech*, Aug. 2017, pp. 1993–1997.
[25] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," in *Proc. Interspeech*, Aug. 2017, pp. 3642–3646.
[26] A. H. Moore, J. M. de Haan, M. S. Pedersen, P. A. Naylor, M. Brookes, and J. Jensen, "Personalized signal-independent beamforming for binaural hearing aids," *J. Acoust. Soc. Amer.*, vol. 145, no. 5, pp. 2971–2981, May 2019.
[27] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, "Predicting the intelligibility of noisy and nonlinearly processed binaural speech," *IEEE/ACM Trans. Audio, Speech, Langauge Process.*, vol. 24, no. 11, pp. 1908–1920, Nov. 2016.
[28] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. Wang, "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *J. Acoust. Soc. Amer.*, vol. 126, no. 3, pp. 1415–1426, Sep. 2009.
[29] W. B. Kleijn and R. C. Hendriks, "A simple model of speech communication and its application to intelligibility enhancement," *IEEE Signal Process. Lett.*, vol. 22, no. 3, pp. 303–307, Mar. 2014.
[30] T. Bentsen, A. A. Kressner, T. Dau, and T. May, "The impact of exploiting spectro-temporal context in computational speech segregation," *J. Acoust. Soc. Amer.*, vol. 143, no. 1, pp. 248–259, Jan. 2018.
[31] C. H. Taal, J. Jensen, and A. Leijon, "On optimal linear filtering of speech for near-end listening enhancement," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 225–228, Mar. 2013.
[32] K. Wagener, J. L. Josvassen, and R. Ardenkjær, "Design, optimization and evaluation of a Danish sentence test in noise: Diseño, optimización y evaluación de la prueba Danesa de frases en ruido," *Int. J. Audiol.*, vol. 42, no. 1, pp. 10–17, Jan. 2003.
[33] J. B. Nielsen and T. Dau, "Development of a Danish speech intelligibility test," *Int. J. Audiol.*, vol. 48, no. 10, pp. 729–741, Jan. 2009.
[34] J. Koopman, R. Houben, W. A. Dreschler, and J. Verschuure, "Development of a speech in noise test (matrix)," in *Proc. 8th EFAS Congr., 10th DGA Congr.*, 2007.
[35] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, Jul. 1993.
[36] W. A. Dreschler, H. Verschuure, C. Ludvigsen, and S. Westermann, "ICRA noises: Arificial noises with speech-like spectral and temporal properties for hearing instrument assessment," *Audiology*, vol. 40, no. 3, pp. 148–157, 2001.
[37] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*. Cham, Switzerland: Springer, 2005, pp. 181–197.
[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Dec. 2014, *arXiv:1412.6980*.
[39] M. Abadi *et al.* (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. [Online]. Available: https://www.tensorflow.org/
[40] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "An evaluation of intrusive instrumental intelligibility metrics," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 11, pp. 2153–2166, Nov. 2018.
[41] M. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Amer.*, vol. 119, no. 3, pp. 1562–1573, Mar. 2006.
[42] E. J. Williams, "The comparison of regression variables," *J. Roy. Stat. Soc. B, Methodol.*, vol. 21, no. 2, pp. 396–399, 1959.

MATHIAS BACH PEDERSEN (Member, IEEE) received the B.Sc. and M.Sc. degrees in mathematical engineering from Aalborg University, Aalborg, Denmark, in 2016 and 2018, respectively.

He is currently employed with Aalborg University as a Ph.D. Fellow. His research interests include machine learning for speech intelligibility prediction, speech and audio processing, and signal representation.

ASGER H. ANDERSEN received the B.Sc. degree in electronics & IT, the M.Sc. degree in wireless communication, and the Ph.D. degree from Aalborg University, Aalborg, Denmark, in 2012, 2014, and 2017, respectively.

While pursuing the Ph.D. degree, he was employed at Demant A/S and associated with the signal and information processing section with Aalborg University. From 2017 to 2022, he worked at Demant A/S as a DSP Specialist. He is currently employed at WS Audiology A/S as a Machine Learning Engineer. His research interests include prediction and measurement of speech intelligibility and speech enhancement, and the applications of both within the development of hearing aids.

SØREN HOLDT JENSEN (Senior Member, IEEE) received the M.Sc. degree in electrical engineering from Aalborg University (AAU), Aalborg, Denmark, in 1988, and the Ph.D. degree in signal processing from the Technical University of Denmark (DTU), Lyngby, Denmark, in 1995.

He is currently a Special Adviser to the Danish Defense and The Danish Ministry of Defense (Ministerial Department) on noise, acoustics, and vibration. From 1988 to 1990, he was a member of the Technical Staff with the Telecommunications Laboratory, Telecom Denmark Ltd., Taastrup (Copenhagen), Denmark. From 1990 to 1995, he worked in various research positions within signal processing, numerical linear algebra, and algorithms with the Electronics Institute, DTU, the Scientific Computing Group of the Danish Computing Center for Research and Education (UNI-C), Lyngby, and the Electrical Engineering Department, Katholieke Universiteit Leuven (KU Leuven), Leuven, Belgium. From 1995 to 2021, he was an Assistant Professor, an Associate Professor, and a Full Professor with the Department of Electronic Systems, AAU, in the areas of digital communications, speech, and signal processing. He is the coauthor of the textbook *Software-Defined GPS and Galileo Receiver—A Single-Frequency Approach* (Birkhäuser), Boston, USA, also translated to Chinese: National Defence Industry Press, China. He was a recipient

of an Individual European Community Marie Curie Fellowship, the Former Chairperson of the IEEE Denmark Section and the IEEE Denmark Section's Signal Processing Chapter (the Founder and the first Chairperson). He is a member of the Danish Academy of Technical Sciences (ATV). He was a member of the Danish Council for Independent Research (2011–2016) appointed by Danish Ministers of Higher Education and Science. He has been an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING, IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING, *Signal Processing* (Elsevier), and *EURASIP Journal on Advances in Signal Processing*.

ZHENG-HUA TAN (Senior Member, IEEE) 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 (SJTU), Shanghai, China, in 1999.

He was a Visiting Scientist with the Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, USA, an Associate Professor with the Department of Electronic Engineering, SJTU, Shanghai, and a Postdoctoral Fellow with the AI Laboratory, KAIST, Daejeon, South Korea. He is currently a Professor with the Department of Electronic Systems and a the Co-Head of the Centre for Acoustic Signal Processing Research, Aalborg University, Aalborg, Denmark. He is also a Co-Lead of the Pioneer Centre for AI, Denmark. He has coauthored over 200 refereed publications. His research interests include machine learning, deep learning, pattern recognition, speech and speaker recognition, noise-robust speech processing, multimodal signal processing, and social robotics. He was the General Chair of IEEE MLSP 2018 and the TPC Co-Chair of IEEE SLT 2016. He is the Chair of the IEEE Signal Processing Society Machine Learning for Signal Processing Technical Committee (MLSP TC). He is an Associate Editor of the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING. He has served as an associate/a guest editor for several other journals.

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 (CPK), Aalborg University, as a Ph.D. student and an Assistant Research Professor. From 2000 to 2007, he was a Postdoctoral Researcher and an Assistant Professor with the Delft University of Technology, Delft, The Netherlands, and an External Associate Professor with Aalborg University. He is currently a Senior Principal Scientist 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 also a Professor with the Department of Electronic Systems, Section for Artificial Intelligence and Sound (AIS), Aalborg University, where he is also the Co-Founder of the Centre for Acoustic Signal Processing Research (CASPR). His research 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.