Data-Driven Non-Intrusive Speech Intelligibility Prediction using Speech Presence Probability

Pedersen, Mathias; Jensen, Søren Holdt; Tan, Zheng-Hua; Jensen, Jesper

Published in:
IEEE/ACM Transactions on Audio, Speech, and Language Processing

DOI (link to publication from Publisher):
10.1109/TASLP.2023.3321964

Publication date:
2024

Document Version
Accepted author manuscript, peer reviewed version

Data-Driven Non-Intrusive Speech Intelligibility Prediction using Speech Presence Probability

Mathias Bach Pedersen, Søren Holdt Jensen, Zheng-Hua Tan, Jesper Jensen

***Abstract*—Time consuming Speech Intelligibility (SI) listening tests with human subjects can be replaced by algorithmic SI predictors. In recent years, data-driven SI predictors have been showing promising results. A major limiting factor in the advancement of data-driven SI prediction is that there is a scarcity of SI listening test data available to train the data-driven methods. In this paper we propose a data-driven SI predictor that does not require access to an underlying noise-free reference signal, i.e., *non-intrusive*, and which does not require listening test data for training. Instead, the proposed method exploits a hypothesized link between SI and Speech Presence Probability (SPP). We show that a neural network can be trained on easily obtainable speech in additive noise data to estimate SPP, and that a simple post-processing stage can be applied in order to map the estimated SPP to SI predictions with high accuracy. The proposed method is evaluated and compared to other state-of-the art non-intrusive SI predictors, and achieves the highest performance even in the presence of processed noisy speech, which the SPP estimator has not been trained on.**

***Index Terms*—Speech intelligibility prediction, speech presence probability.**

I. Introduction

Speech Intelligibility (SI) is an important concept in many speech processing applications, including noise reduction for speech enhancement, speech dereverberation, speech separation, etc. SI is a measure of the proportion of words in a speech signal that can be understood by humans on average, and considering speech as a conveyor of information between people, SI is a good measure of the efficiency with which the information is conveyed. The SI of a speech signal can be deteriorated by a variety of factors such as background noise, competing speakers, reverberation, digital transmission, processing, etc. The ability to measure SI is valuable when these factors are encountered, for example, during the development of an enhancement system for speech in background noise.

Traditionally, SI is measured via listening tests, where a panel of human listeners evaluate the SI of several instances of speech signals under the noise/processing conditions of interest. The SI is subsequently reported as an average across test participants. To supplement the time consuming process of conducting listening tests it may be desirable to apply SI prediction algorithms, which allow for fast and repeatable, but generally less accurate estimations of SI.

There are several important factors that distinguish existing SI predictors. In particular, *intrusive* and *non-intrusive* predictors require different input [1]. Intrusive SI predictors require either the clean speech signal or the noise signal in isolation, as well as the noisy/processed signal under test. Conversely, non-intrusive SI predictors only use the noisy/processed signal. Obviously, intrusive predictors are generally more accurate than their non-intrusive counterparts because they have access to more information, but the ability to predict SI based only on the noisy/processed signal makes non-intrusive predictors more widely applicable [2]. In portable/wearable devices, e.g., as part of the processing in a hearing assistive device, non-intrusive prediction is perhaps the only practical option.

SI predictors may also be classified as either *classical* or *data-driven*. Classical SI predictors are usually engineered by hand, using classical signal processing methods demonstrated to correlate with measured SI, such as those described in [3], [4] and [5]. Data-driven SI prediction is a more recent approach based on machine learning, where listening test databases are used to train, e.g., a neural network to perform SI prediction. Examples include the networks described in [6], [7] and [8]. As mentioned, the motivation for using or developing SI predictors is that listening tests are time-consuming and in some cases impractical. This fact, however, is also a major limiting factor in the current development of data-driven SI predictors. In particular, the lack of listening test data prevents effective training of large architectures with many layers and parameters. The performance of data-driven SI predictors depends greatly upon the conditions on which they are trained, and we have previously observed significant losses in performance when the trained predictors are applied in unseen conditions [2], [9].

From a machine learning oriented point of view the solution to poor generalization has traditionally been to increase the quantity and diversity of training data until the desired generalizability is achieved. Solving the problem this way, however, is not feasible with the current scarcity of listening test data. Current approaches to mitigate or circumvent this problem include downscaling architectures [10], or hybridizing data-driven and classical approaches [10], [6]. In particular, it is a common approach to replace the listening test measurements of SI used as training labels with approximations produced by classical SI predictors [6], [11]. In this fashion, large training data set of great diversity can be synthesized much faster, at the expense of prediction errors in the SI labels induced by the SI predictor that was used. In fact, any limitations of the chosen classical SI predictor will be inherited by the trained data-driven SI predictor.

In this paper we address the problem of listening test data scarcity by proposing a data-driven non-intrusive SI prediction method that can be trained without access to any labelled data from listening tests. The training data for the proposed method consists of synthetically generated and automatically labelled noisy speech signals, which are easily obtained in abundance. The proposed method relies on the hypothesis that there is a strong relationship between SI and Speech Presence Probability (SPP). SPP can be defined as the probability that a given noisy, potentially processed, speech signal or time-

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

M. B. Pedersen, Z. -H. Tan and J. Jensen are with the Department of Electronic Systems, Aalborg University, Aalborg 9220, Denmark (e-mail: mbp@es.aau.dk).

S. H. Jensen is with the Danish Ministry of Defence Estate Agency, Hjørring 9800, Denmark.

J. Jensen is with Demant A/S, Smørum 2765, Denmark.

frequency region thereof is dominated by speech. SPP is defined for the signal model of speech in additive noise as the probability that the signal-to-noise ratio exceeds a fixed, predefined threshold. We show that a data-driven classifier of time-frequency tile-wise speech dominance, which is trained to minimize the mean squared error, computes SPP. Basically, we propose to train a neural network speech dominance classifier on abundant synthetically generated training data and use it to estimate SPP for the signal under test, subsequently mapping the estimated SPP to SI index predictions via a post processing stage. SPP estimation has been employed for speech recognition and speech enhancement in, e.g., [12, 13]

A major advantage of the proposed SPP estimator is that - as we show - it is easy to generate large amounts of data for training. By sampling from easily obtainable speech and noise data, e.g., from freesound.org, one can generate and label vast amounts of noisy speech signals for training. We demonstrate that despite the training data being constrained to speech in additive noise, the SI of a wide variety of non-linearly processed noisy speech signals can be accurately predicted.

The proposed SI predictor, which we call Deep Speech Presence (DSP), consists of a deep Convolutional Neural Network (CNN), trained to predict SPPs per time-frequency tile in a wide variety of noise conditions and for different speakers, along with a post processing stage, designed to map the time-frequency-wise estimated SPPs into a single scalar SI prediction. The proposed post processing stage encompasses a time-segmented selection and average operation, where the average of the highest p percent of the predicted SPP tiles from short temporal segments is computed. By considering only the highest SPP tiles, errors due to network uncertainty and non speech active time-frequency regions can be mitigated. We show that DSP is able to accurately predict the SI of a variety of listening test conditions outside the scope of the training data, even including several non-linear processing conditions. In particular, DSP compares favorably to existing, non-intrusive SI predictors including the Non-Intrusive Short-Time Objective Intelligibility (NISTOI) [14], the Speech-to-Reverberation Modulation energy Ratio (SRMR) [5] and a data-driven non-intrusive emulator of the Short-Time Objective Intelligibility (STOI) inspired by existing STOI emulators such as the Non-Intrusive Speech Assessment (NISA) [6], the Twin Hidden-Markov-Model based STOI (THMMb-STOI) [11] and STOINET [15].

The paper is structured as follows. Sec. II contains the background and related work. Sec. III details the theory and methodology used in the study, including a description of the architecture of the proposed CNN. Sec. IV describes the data used for network training and for the SI prediction evaluation. Sec. V describes other SI predictors that are evaluated for the purpose of comparison. Sec. VI describes the experiments and results hereof, including a discussion of the results. Finally, Sec. VII concludes the study.

II. Related work

The Articulation Index (AI) [3] is among the earliest attempts to predict SI, and uses long-term Signal-to-Noise Ratios (SNR) in separate frequency bands to quantify the contribution of each band to the overall SI. Finally, these contributions are added together, yielding an index that correlates with SI. Later, the Speech Transmission Index (STI) [16] was designed to predict the SI of speech passed through a given transmission channel. Rather than specific speech signals, the STI uses a bank of probe signals, sent through the transmission channel in pairs, to determine how well the transmission channel is able to preserve sound pressure level differences between probe signal pairs. The average preservation of sound pressure level differences across probe signals can be used as a predictor of the SI of the speech signal after transmission.

Inspired by the AI, the Speech Intelligibility Index (SII) [17] and Extended SII (ESII) [18] are based on linear combinations of contributions to SI from separate frequency bands. In SII, the long term contribution is computed for each frequency band, whereas in the ESII, SII values are computed for short consecutive segments and then averaged across time. This difference allows the ESII to predict SI more accurately than standard SII when the noise contains temporal modulations. The Spectro-Temporal Modulation Index (STMI) [19] takes inspiration from STI, and can similarly be used to predict either the SI of a given noisy/processed speech signal, or the impact on SI of a given transmission channel. For specific speech signals, the STMI works by passing the clean speech signal and the noisy/processed speech signal through an auditory model, designed to extract spectro-temporal modulations, and subsequently measuring the difference in their spectro-temporal modulations. When used to predict the impact of a transmission channel, the clean speech signal is replaced by a set of probe signals in the form of spectro-temporal ripple patterns, and the noisy/processed speech signal is replaced with these same probe signals, passed through the transmission channel under test. The degree to which spectro-temporal modulations are preserved is used to predict SI.

The glimpse proportion model [20] works by computing and counting the number of separate, so-called glimpses that a listener would be able to hear in the noisy/processed signal. In this context, glimpses are defined as isolated time-frequency regions where the local SNR is sufficiently high.

The Short-Time Objective Intelligibility (STOI) [4] and Extended STOI (ESTOI) [21] are based on correlation coefficients between the clean and noisy/processed speech signals. In particular, the correlations are computed between short term (spectro)-temporal envelopes and averaged over time to give a prediction of SI. ESTOI makes use of an extended normalization step to better handle temporally modulated noise than the standard STOI [21].

Speech Intelligibility using Mutual Information (SIMI) [22], Mutual Information Variational Bayes (MI-VB), MI K Nearest Neighbours (MI-KNN), MI Expectation Maximization (MI-EM) [23] and Speech Intelligibility In Bits (SIIB) [24] are all estimators of the mutual information between the clean and noisy/processed speech signals, which can subsequently be mapped to SI.

The Speech to Reverberation Modulation energy Ratio (SRMR) [5] is a non-intrusive SI predictor designed to predict the SI of reverberant speech signals. The SRMR prediction is based on the ratio of energy at low temporal modulation frequencies with that of higher modulation frequencies. The Non-Intrusive STOI (NISTOI) [14], as the name implies, is a non-intrusive front-end for STOI that works by estimating the clean speech from the noisy/processed signal using a statistically based speech enhancement method. The estimated clean speech signal is then used in place of the actual clean speech signal in the original STOI algorithm.

In [25] a relatively small network was trained to map a number of pre-extracted features intrusively to SI. Likewise, in [2] a

network was trained to predict SI intrusively based on the one-third octave band magnitudes of both the clean reference and the noisy/processed signal. In [9] a large CNN was trained to predict SI intrusively, given the clean and noisy/processed waveforms.

In [7] a neural network is trained to non-intrusively predict SI, given the one-third octave band magnitudes of the noisy/processed signal under test. The Multi-Branched Intelligibility Net (MBI-Net) [26] is another non-intrusive, data driven SI predictor based on CNN's. The Non-Intrusive Speech Assessment (NISA) method [6] is a non-intrusive STOI emulator, i.e., it is trained to predict STOI rather than SI, but without using a clean reference signal. In particular, NISA is trained to map a number of pre-extracted features from the noisy/processed signal to SI predictions produced by STOI. Similarly, the twin hidden Markov model based STOI proposed in [11] is another non-intrusive STOI emulator, and the CNN proposed in [27] is a trained non-intrusive STI emulator.

Recently, data-driven Automatic Speech Recognition (ASR) systems have been used as an avenue for SI prediction. In [28] as well as [29] features and metrics are derived from pre-trained ASR hidden Markov models and used as predictors of SI. ASR systems have also been proposed in [30] and [31] as a way of simulating listening tests, successfully predicting SI for hearing impaired listeners. Although some of these ASR based SI predictors do not require the clean reference signal directly, some of the metrics they use rely on ground-truth transcripts. Non-intrusive ASR based SI predictors have been proposed, and include the NO-Reference Intelligibility (NORI) method [8], and the ASR based uncertainty measure [32].

III. Architecture and post processing

In this section we introduce our proposed SI predictor, which we call DSP. First, we derive an optimal time-frequency domain SPP estimator for speech in additive noise. Next, we present the deep learning architecture to be trained for SPP estimation. Then, we present the proposed methodology for converting the SPP estimator into an SI predictor, by appending a simple post processing step to the SPP estimator. Finally, we outline the procedure used to train the proposed architecture.

A. Theoretical motivation

Let us begin by formally defining SPP, and then show how a Deep Neural Network (DNN) can be trained to estimate SPP in the time-frequency domain. We will be using upper case letters to denote random variables, and lower case letters to denote the corresponding realizations. We rely on the following signal model in the Short-Time Fourier Transform (STFT) time-frequency domain,

$$X(k,m)=S(k,m)+V(k,m), \tag{1}$$

where, $S(k,m)$ is a clean speech signal, $V(k,m)$ is a noise signal, and $k=0,...,K-1$ and $m=0,...,M-1$ are discrete time and frequency indices. We assume that $S(k,\,m)$ and $V(k,\,m)$ are statistically independent.

We are interested in whether a given time-frequency tile is speech dominated, meaning that the local SNR of $x(k,m)$ is above a fixed threshold;

$$20\log_{10}\frac{|s(k,m)|}{|v(k,m)|}>\tau \text{ dB}. \tag{2}$$

Let us introduce the indicator variable

$$I(k,m)=\begin{cases}1 & \text{if } x(k,m) \text{ is speech dominated}\\ 0 & \text{otherwise}\end{cases}, \tag{3}$$

which serves as an indicator of the time-frequency tiles where $x(k,m)$ is speech dominated.

Now, choose a context subset, $\mathcal{X}(k,\,m)$, of time-frequency tiles that will be used to estimate the realization $i(k,\,m)$ of $I(k,\,m)$, for example, all tiles $\mathcal{X}(k,\,m)\;=\;\Big\{x(\acute{k},\acute{m})|\acute{k}=0,...,K-1.\,\acute{m}=0,...,M-1\Big\}$. Now, consider the conditional mean,

$$\begin{aligned}\mathrm{E}_I[I(k,m)|\mathcal{X}(k,m)]&=\sum_{i(k,m)} i(k,m)\mathrm{P}(I(k,m)=i(k,m)|\mathcal{X}(k,m))\\ &=\mathrm{P}(I(k,m)=1|\mathcal{X}(k,m)),\end{aligned} \tag{4}$$

where $\mathrm{E}_I[\cdot]$ denotes the expected value with respect to the random variable $I(k,m)$. Note that $\mathrm{P}(I(k,m)=1|\mathcal{X}(k,m))$ is the posterior probability that $x(k,m)$ is speech dominated, i.e., the SPP of $x(k,m)$. In other words, the SPP is given by the conditional mean in (Eq. 4). Recalling that the conditional mean is the Minimum Mean Square Error (MMSE) estimator, $\hat{i}(k,m)$ of $I(k,m)$ [33], it follows that

$$\begin{aligned}\hat{i}(k,m)&=\mathrm{P}(I(k,m)=1|\mathcal{X}(k,m))\\ &=\mathrm{E}_I[I(k,m)|\mathcal{X}(k,m)]\\ &=\underset{\hat{i}(k,m)}{\operatorname{argmin}}\mathrm{E}_I\left[\left(\hat{i}(k,m)-I(k,m)\right)^2\right].\end{aligned} \tag{5}$$

Although many methods have been proposed for estimating the SPP (Eq. 4) for the simple special case where $\mathcal{X}(k,m)=x(k,m)$ using parametric statistical models, e.g., [34, chpt. 5] and the references therein, the fact that the SPP estimation problem can be posed as a MMSE estimation problem, (Eq. 5), opens the possibility for estimating $\mathrm{P}(I(k,m)=1|\mathcal{X}(k,m))$ for larger contexts, $\mathcal{X}(k,m)$, and for using powerful data-driven statistical models. In particular, we propose using a DNN to estimate $\hat{i}(k,m)$ given $\mathcal{X}(k,m)$.

An important practical consideration is that the theoretical sample space of (Eq. 1) is huge, encompassing all combinations of every conceivable speech and noise signal. This means that even though a DNN trained in this way is theoretically an optimal estimator of SPP, in practice the types of speech and noise used to train the SPP estimator can only cover a subset of the signal model (Eq. 1). Optimality is no longer guaranteed when the DNN is applied to new talkers and noise types, and certainly when applied to processed speech, as we will do in Sec. VI. Consequentially, constructing a diverse set of training data is highly important.

Another important consideration is the choice of context subset, $\mathcal{X}(k,m)$. Ideally, all of the time frequency tiles statistically dependent on $x(k,m)$ should be used. Due to causality constraints, system complexity constraints, etc., however, it may be more practical to limit the context subset. In this study, for instance, the architecture has a limited temporal receptive field, which means that the estimation of the SPP in a given time-frequency tile will be based only on the tiles within a certain time window around $x(k,m)$, i.e.,

$$\mathcal{X}(k,m)=\Big\{x(\acute{k},\acute{m})|\acute{k}=k-c,...,k+c.\,\acute{m}=0,...,M-1\Big\}, \tag{6}$$

where c is a constant integer.

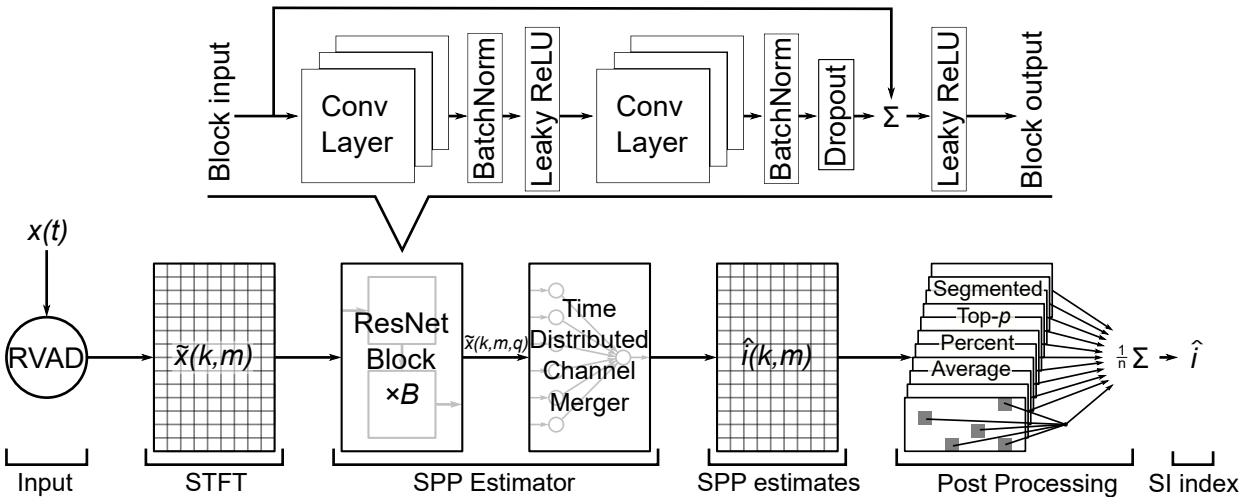


Fig. 1. The proposed SI prediction method. The input $\tilde{x}(k,m)$, pre processed by rVAD, is a potentially noisy/processed speech signal in the STFT-magnitude domain. This signal is processed by B stacked ResNet Blocks, illustrated in the top half of the diagram. Next, the channels, indexed by the variable q, resulting from the ResNet convolutional layers are merged by a time distributed fully connected layer. The output, $\hat{i}(x,m)$, is a tile-by-tile estimation of SPPs. The estimated SPPs are finally mapped to SI predictions via the segmented top-p percent average post processing step.

B. Network design

In this section, the neural architecture used in this study is described. Figure 1 shows a block diagram of the architecture. The input to the SPP estimator is a noisy/processed speech signal that is first processed by a Voice Activity Detector (VAD) to remove any time segments, estimated to not contain any underlying speech. Specifically, we employ the open-source, unsupervised, noise-robust, rVAD [35]. Subsequently, the input signal is represented in the STFT-magnitude domain with M frequency channels, $\tilde{x}(k,m)$. The architecture is inspired by the Residual Network (ResNet) architecture proposed in [36], and the input is given to a stack of B ResNet blocks, each consisting of two convolutional layers with a skip connection into the summation point as shown in the top half of Figure 1. For the convolutional layers, Q kernels, leaky ReLU activations and batch normalization are used.

The dimensions of the output need to match those of the input, $\tilde{x}(k,m)$, but the signal at the output of the ResNet blocks, $\tilde{x}(k,m,q)$, has an extra dimension due to convolution with multiple kernels. To collapse this extra dimension, the final layer of the neural architecture, after the ResNet blocks, is a time distributed channel merger described by the following equation,

$$\hat{i}(k,m)=\sigma\left(\sum_{q}\sum_{\acute{m}}[\tilde{x}(k,\acute{m},q)w(m,\acute{m},q)]+b(\acute{m},q)\right), \quad (7)$$

where w and b represent the weights and biases of the layer respectively. Time distributed means that the layer is designed to work with a single time frame, k, as input, and that the time frames of the total input are passed individually and independently through the layer. Channel merger means that the layer combines all the channels resulting from the the convolutions in the convolutional layers in the preceding ResNet blocks into one, i.e., the three-dimensional intermediate signal, $\tilde{x}(k,m,q)$ where $q=0,1,...,Q-1$ denotes the channel, is condensed into the two-dimensional SPP predictions, $\hat{i}(k,m)$. For each time frame, k, the channel merger is equivalent to a fully connected layer with M nodes, weights w and bias b, each responsible for a single STFT frequency bin in the output. Because this layer is fully connected, it can combine information across all frequency and kernel channels. The final activation function is a sigmoid, σ.

Implicit in this architecture is a choice of context $\mathcal{X}(k,m)$, given by the receptive field of the CNN. In particular, $\mathcal{X}(k,m)$ contains all frequency bands, and all time frames within a certain distance from k depending on B.

The implementation of the architecture along with the trained weights can be found on-line.[1]

The output of the trained SPP estimator is a spectrogram of time-frequency tile-wise estimated SPPs, $\hat{i}$. In order to map the estimated SPPs to a scalar SI prediction, a post processing scheme is required. In preliminary experiments, we observed that a simple average of the estimated SPPs produced by a trained network is already correlated with SI. Starting from this average-operation, two simple modification are proposed in order to increase the correlation with SI.

C. Post processing

1) Top-N percent: We propose a post processing scheme, which takes a fixed percentage of the highest SPP predictions, the Top-p percent, and computes the average of only those SPP tiles. More specifically, take the average of the $\lfloor \frac{N}{100}KM \rfloor$ largest values of $\hat{i}(k,m)$, which are found using the argpartition function in Python's *numpy* library. There are two motivations behind this approach. Firstly, tiles with high estimated SPP may be indicative of glimpses, i.e., isolated time-frequency regions above a fixed SNR threshold. The Glimpse model, proposed by [20], demonstrated that the number of glimpses in a noisy/processed speech signal is correlated with SI. Secondly, the estimated SPPs can also be interpreted as the classification uncertainty of the trained CNN as a classifier of time-frequency tile-wise speech dominance. This uncertainty is maximized when the estimated SPP is equal to 0.5 and minimized when the estimated SPP equals 0 or 1. This means that a tile with an estimated SPP close to 0.5 is not useful for predicting SI. Omitting highly uncertain time-frequency tiles from the average, increases correlation with SI.

[1] https://github.com/Mapede/Deep-Speech-Presence

2) Time-frequency segmentation: While the top-p percent average introduced above improves the correlation with SI compared to a full average operation, in certain situations it can fail. Consider, for example, a completely clean speech signal, and a speech signal, the first half of which has been completely masked by noise. The top-p percent of SPPs would be located in the clean parts of the signals, and in both cases the average would be nearly 1. More generally, the top-p percent average may work less well in fluctuating noise conditions.

In order to improve performance in fluctuating noise conditions, we propose segmenting the SPP spectrogram into short temporal regions of N frames, and subsequently applying the top-p percent selection to each individual window, and finally computing the average across all selected tiles. In the example above, half of the tiles would now be forced to come from the noisy part of the speech signal, correctly reducing the predicted SI. This idea of working on short time segments is common in the field of SI prediction, and is inspired by existing SI predictors such as ESII [18] and ESTOI [21]. A similar segmentation of frequency bands could be performed, however, in preliminary experiments doing so did not increase the correlation with SI. The reason appears to be that reducing the size of the segments beyond a certain point diminishes the effectiveness of the top-p percent averaging strategy.

Overlapping segments are used to avoid boundary effects in the top-p percent averaging. In Sec. VI the robustness to changes in both percentage, p and segment length, N, is investigated.

D. Training procedure

Before training the CNN architecture, the training dataset is constructed. As mentioned, we exclusively use speech in additive noise, which means that every training sample can be generated by adding a clean speech signal to an appropriately scaled noise signal. The speech and noise samples are drawn from two databases, described in detail in Sec. IV-A, which are divided into a range of speakers, and noise types respectively. The generation of training samples is carried out by drawing a sample pair from one speaker and one noise type randomly without replacement, then moving on to a different speaker and noise type for the next sample. This process is repeated until a desired quantity of training data has been generated. For this study, we generated 2^{17} training samples of 1.7 seconds, or ~ 60 hours in total. In addition, 2^{14} samples, or ~ 7.5 hours of validation data is generated.

The training label, $i(k,m)$, of a given input, $x(k,m)$, is computed based on the SNR of the individual time-frequency tile. These SNR values are computed during the training sample generation, where the clean speech and noise signals are available in isolation. The spectrogram of SNR values in dB is computed, cf. (Eq. 2). A threshold, τ, is then used to generate $i(k,m)$ as defined in (Eq. 2) and (Eq. 3).

The network is trained using the MSE loss function, following the derivation in Sec. III-A, with the Adam optimizer [37]. Dropout of 25% is used in the second convolutional layer of each ResNet block. Training was carried out using a decaying learning rate over 72 epochs, and the final weights were chosen to be those minimizing the MSE on the validation data.

IV. Datasets

Two distinct collections of data are used in this paper. First, a training database has been constructed for the purpose of training the SPP estimator. This database consists of a clean speech database that contains a number of speech datasets with different talkers, and a noise database with a large range of additive noise types. These databases are used to generate training, testing and validation samples of speech in noise for the SPP estimator. Secondly, a database of listening tests was used for the purpose of evaluating the performance of the complete SI prediction method.

A. Datasets for training the SPP estimator

1) Speech data: To train the SPP estimator, a variety of talkers from several speech datasets were used. The datasets used are listed in Table I. Amongst these datasets, ADFD provides the majority in terms of speaker variation.

2) Noise data: A variety of recorded and generated noise types are used to train the SPP predictor. The noise data is mixed with the speech data at a range of SNR values from 4 to -30 dB. The focus in constructing the noise dataset for this study was on the variety of noise types. We do not include any noise conditions that contain competing speech, because non-intrusive SI prediction in multi-talker conditions requires some way of identifying the target talker, which is beyond the scope of this study. The advantage of the proposed method of SI prediction is that the training data for the data-driven part, the SPP estimation, is relatively much easier to obtain in large quantities compared to listening test data, because the training labels are generated algorithmically, rather than manually by human listeners. The noise types used are listed in Table II and briefly described in the following. The code used to generate the noise types that we generated is provided on-line[3].

Speech Shaped Noise (SSN) was generated by constructing a prediction filter based on Dantale II, and passing white Gaussian noise sequence through this filter to obtain a noise signal with the long-term spectrum of speech. *High-pass and low-pass filtered Noise* is SSN filtered by high and low-pass filters with cut-off frequencies ranging from 0.15 to 4.5 kHz. *Checkerboard noise* is SSN that has been artificially modified by lowering the intensity in time-frequency checkerboard patterns, with widths ranging from 8×8 to 32×32 tiles in the STFT domain, i.e., time durations from 100 ms to 400 ms and frequency bandwidths from 300 Hz. to 1250 Hz. *Speech thresholded SSN* is generated by applying a time-frequency tile-wise threshold, computed based on speech signals from ADFD. *Modulated harmonic sinusoids* consist of sine wave harmonics with fundamental frequencies ranging from 100 to 250 Hz. and temporal modulations ranging from 0 to 8 Hz. *Speech energy matched noise* is SSN scaled on a time-frequency tile-wise basis to match the energy of speech signals from ADFD. *Speech envelope matched noise* is SSN, where short-time segments are scaled to match the energy of equivalent time segments of speech signals from ADFD. *Artificial musical noise* is generated by setting time-frequency tiles of SSN below a certain magnitude threshold, relative to the average magnitude within each frequency band, to zero. The *CHiME-3* dataset consists of recorded noise of cars on a street [40]. We omitted noise conditions with other talkers from this dataset. The *Clarity-2021* dataset is a collection of the recorded noise of various household appliances, including a dishwasher, microwave, vacuum cleaner, etc. [41]. The *Diverse Environments Multichannel Acoustic Noise Database* (DEMAND) dataset consists of recordings of various crowded

[3]https://github.com/Mapede/Deep-Speech-Presence

TABLE I
OVERVIEW OF THE SPEECH DATASETS USED TO CONSTRUCT THE TRAINING SAMPLES FOR THE PROPOSED SPP PREDICTOR.

Name	Speakers	Sentences	Language
Akustiske Databaser For Dansk[2] (ADFD)	100 Male & Female	26060	Danish
CLUE [38]	1 Male	250	Danish
IEEE dataset [39]	1 Male	720	English

TABLE II
OVERVIEW OF THE NOISE TYPES AND DATASETS USED TO CONSTRUCT THE TRAINING SAMPLES FOR THE PROPOSED SPP PREDICTOR. SUB-TYPES REFERS TO THE NUMBER SUB-TYPES OF NOISE CONTAINED IN EACH DATASET.

Name	Source	Sub-types
Speech Shaped Noise (SSN)	Generated	1
High-pass and low-pass filtered noise	Generated	12
Checkerboard noise	Generated	4
Speech thresholded SSN	Generated	3
Modulated harmonic sinusoids	Generated	12
Speech energy matched noise	Generated	3
Speech envelope matched noise	Generated	3
Artificial musical noise	Generated	3
CHiME-3	Recorded [40]	3
Clarity-2021	Recorded [41]	7
DEMAND	Recorded [42]	6
Ambience and music	Recorded freesound.org	4

locations, including a metro station and public square. Finally, sound files from freesound.org including *forest ambience*[4], *traffic*[5], *instruments*[6], and *drumbeats*[7] are used. Each of the noise types and databases described, contains a number of sub-types given in Table II. An equal amount of training data is constructed using each sub-type.

B. Datasets for evaluating the SI predictor

Stimuli and results from six listening tests, listed in Table III, are used to evaluate the performance of the proposed SI prediction method. These tests contain a variety of additive noise conditions and a large variety of non-linear processing conditions. The datasets are referred to by the four-letter tags given in Table III, which also lists information concerning the speech and noise material, as well as processing types present in the datasets.

A subset of the *Noisex* database [46] is used in *ModulNoise*, specifically the recorded noise of a ship engine and machine gun. The *ICRA* [47] noise database, used in *ModulNoise*, contains artificial noise signals with spectro-temporal modulations, designed to resemble speech on a shorter term than SSN. *Bottle Factory Noise* (BFN), used in *IdealMasked*, contains sounds of clinking bottles on a conveyor, recorded in a bottling factory hall. *Car* refers to noise of a running car recorded inside the cabin. The *Ideal Time-Frequency Segregation* (ITFS), used in *IdealMasked*, refers to enhanced speech signals that have been processed in the time-frequency domain by applying an ideal mask or threshold to remove noise and preserve speech [48]. The *low- and high-pass filtering*, used in *LHPFiltered*, is applied to the speech prior to the addition of SSN. *Pre-noise enhancement*, used in *PreFiltered*, refers to optimal filtering that is applied to clean speech signals, pre-processing them for a known noise environment. The *SSN replacement*, used in *NoiseReplace*, refers to a process where different percentages of short segments of clean speech signals were chosen strategically to be replaced with pure SSN, maximizing the impact on mutual information, cochlea-scaled spectral entropy and sound intensity [45].

[4]https://freesound.org/people/klankbeeld/packs/28209/
[5]https://freesound.org/people/klankbeeld/packs/7274/
[6]https://freesound.org/people/Bansemer/sounds/204745/
[7]https://freesound.org/people/bigjoedrummer/sounds/331665/

V. BASELINE SI PREDICTORS

The proposed SI predictor, DSP, will be compared to a number of existing non-intrusive SI predictors: Non-Intrusive STOI (NISTOI) [14], the Speech-to-Reverberation Modulation energy Ratio (SRMR) [5], and a data-driven, non-intrusive STOI emulator, essentially replicating the functionality of the Non-Intrusive Speech Assessment (NISA) method [6].

A. NISTOI

NISTOI [14] is a non-intrusive extension of STOI [4], which uses a statistical model to estimate a clean speech signal from the noisy/processed signal under test. NISTOI estimates the clean speech signal by projecting the modulation domain representation of the noisy processed speech signal onto a low dimensional subspace of principal components specific to clean speech. The estimated clean speech signal is used as a replacement for the clean reference signal in the conventional STOI method.

B. SRMR

SRMR [5] is a non-intrusive SI predictor designed with a main focus on reverberant and de-reverberated speech. SRMR computes a ratio between the energy at low temporal modulation frequencies and at higher temporal modulation frequencies. The idea is based on the observation that the temporal modulation energy of clean speech is concentrated at low modulation frequencies, and that reverberation increases the energy of higher modulation frequencies. SRMR has been demonstrated to correlate with the SI of reverberant and de-reverberated speech, but has also been applied in the context of additive noise [49].

C. Data-driven non-intrusive STOI emulator

A range of state-of-the-art approaches for data-driven SI prediction have been proposed, which emulate an existing intrusive SI-predictor non-intrusively using, e.g., a neural network [15], a tree based regression model [6], and a twin hidden Markov model [11]. In order to compare the proposed SI predictor to a member of this class of algorithms, we implemented a data-driven, non-intrusive STOI emulator. This STOI emulator is heavily inspired by the NISA [6] method, which is a non-intrusive, data-driven STOI emulator. Since the implementation of NISA is not publicly available, we trained a different architecture, similar to that of the proposed SPP estimator seen in Figure 1, to perform the same task as NISA,

TABLE III
OVERVIEW OF THE DATASETS USED TO EVALUATE THE PERFORMANCE OF THE PROPOSED SI PREDICTOR. THE COLUMNS LABELLED #SUBJ. AND #COND. LIST THE NUMBER OF PARTICIPATING LISTENERS AND THE NUMBER OF DIFFERENT CONDITIONS

Dataset		Size		Content		
Tag	Ref.	#subj.	#cond.	Speech material	Noise & processing types	SNR range (dB)
ModulNoise	[21, Sec. IV-1]	12	60	Dantale II	Noisex, SSN, ICRA	$(-2, -37)$
IdealMasked	[43, Sec. II]	15	126	Dantale II	SSN, BFN, Car, ITFS	$(-7, -60)$
LHPFiltered	[7, Sec. III-D_4]	8	327	ADFD	SSN, Low- and high-pass filtering	$(8, -10)$
PreFiltered	[44, Sec. III-B]	16	18	Dutch Hagerman matrix test	SSN, Pre-noise enhancement	$(-5, -20)$
SCNoiseReduct	[21, Sec. IV-5]	13	20	Dutch Hagerman matrix test	SSN, Single channel noise reduction	$(0, -8)$
NoiseReplace	[45, Chpt. 6.1]	24	13	Dantale II	SSN replacement	N/A

namely predicting the output of STOI given only the noisy/processed signal. The STOI emulator architecture differs from the SPP estimator only in the final layer, where the SPP estimator has M kernels, and the STOI emulator has one. The STOI emulator, as a result, has a single output channel of intermediate STOI predictions, and taking the average across time gives the final STOI prediction.

The NISA-inspired STOI emulator is trained on the same data as the SPP estimator, described in Sec. IV-A. This eliminates any advantage DSP would have over a STOI emulator trained on different data. To verify that the trained STOI emulator achieves the state-of-the-art performance reported in [6], we compare the predictions of the STOI emulator to the true STOI scores on the test set. In this comparison, the trained STOI emulator yields a Spearman correlation of 0.93, a Kendall's tau of 0.79 and root MSE of 0.09. STOI predictions by NISA are reported to give a Spearman correlation coefficient of 0.95 and root MSE of 0.08 with STOI, c.f. [6, Table 5], for a dataset similarly comprised of speech in additive noise. For further comparison, the STOI predictions of the twin Hidden Markov Model (HMM) in [11], is reported to yield a Kendall's tau of 0.74 and root MSE of 0.09, c.f. [11, Table 1], for a dataset similarly comprised of speech in additive noise. The high correlations and low root MSE, combined with the fact that NISA and the twin-HMM report correlations and root MSE's in the same range, all for speech in various additive noise conditions, suggests that the implemented STOI emulator achieves state-of-the-art performance, and can be used as a representative baseline.

VI. RESULTS

To demonstrate the hypothesized link between SPP and SI, and the applicability of DSP for SI prediction, a performance evaluation of the DSP SI predictor is carried out in this section. The architecture, described in Sec. III-B, is trained using the data described in Sec. IV-A to estimate SPP. In order to compute an SI prediction from the estimated SPPs, the post processing stage in Sec. III-C is applied. We analyze the effects of rVAD on the performance of DSP by comparison to an ideal, intrusive VAD (STOIVAD) [4].

A. Parameter values

The following parameters are used in the performance evaluation reported in the following sections. The threshold $\tau = -8$ dB is used to label the training data. The initial time-domain sample rate is 10 kHz, and the STFT uses Hann windows of $W = 2^8$ samples corresponding to a duration of 25.6 ms, with 50% overlap, resulting in $M = 129$ non-negative frequency channels. The SPP estimator architecture uses $B = 8$ ResNet blocks and $Q = 128$, 3×3 kernels. With these settings the SPP predictor network has a total of 4.35 million trainable parameters. The post processing stage uses segments of $N = 30$ frames, corresponding to 384 milliseconds with 25 frames of overlap, i.e., 83.33%. This segment length is chosen to be equivalent to that of STOI [4] and ESTOI [21]. We found that using the top $p = 5$ percent of SPP's in the post processing resulted in good predictions of ESTOI, on the data described in Sec. IV-A. In particular, p=5 was found to give the highest Spearman correlation with ESTOI out of $p = \{5,10,15,20,25\}$.

B. Performance metrics

Three performance metrics relative to the ground truth measured SI are used in the evaluation of SI predictors. These metrics are the Spearman rank order correlation coefficient, the Pearson correlation coefficient and the Root Mean Squared Error (RMSE). The Spearman correlation is a rank order correlation, i.e., it measures the degree to which SI predictions are monotonically related to measured SI. This metric is useful as a gauge of an SI predictors ability to rank different speech signals in order of measured SI, without regard to how close the predictions are to the absolute SI. This is particularly useful, because it reflects the typical non-intrusive SI prediction scenario, where the psychometric function underlying the signal of interest is unknown. The Pearson correlation coefficient is useful as a measure of the linearity between predicted and measured SI. The Pearson correlation can be interpreted as a measure of how well the SI predictions fit the psychometric function of the data in question. The RMSE reflects how well the methods are able to predict absolute SI when the underlying psychometric functions are known. RMSE has the advantage that the unit correspond to the absolute size of prediction errors. Figure 2 displays the SI predictions versus ground truth SI measurements in listening tests. To facilitate a fair comparison between predictors, logistic transforms have been fitted for each predictor-dataset pair. This is standard practice in the evaluation of SI predictors, and serves the purpose of mapping the raw intelligibility index predictions to absolute SI predictions, which can be compared across different predictors. It can be seen in Figure 2 that NISTOI and SRMR work poorly on a few of the datasets. In particular, NISTOI does not work well for *ModulNoise* and *NoiseReplace* and SRMR does not work well for *IdealMasked* and *SCNoiseReduct*, as evidenced by the narrow vertical distribution of points in the scatter plots. This may be explained by the datasets being outside the scope for which these predictors were designed. In particular, NISTOI is based on STOI, which is known to work poorly on temporally modulated noise [21], which happens to be highly prevalent in *ModulNoise* and *NoiseReplace*. SI prediction for speech

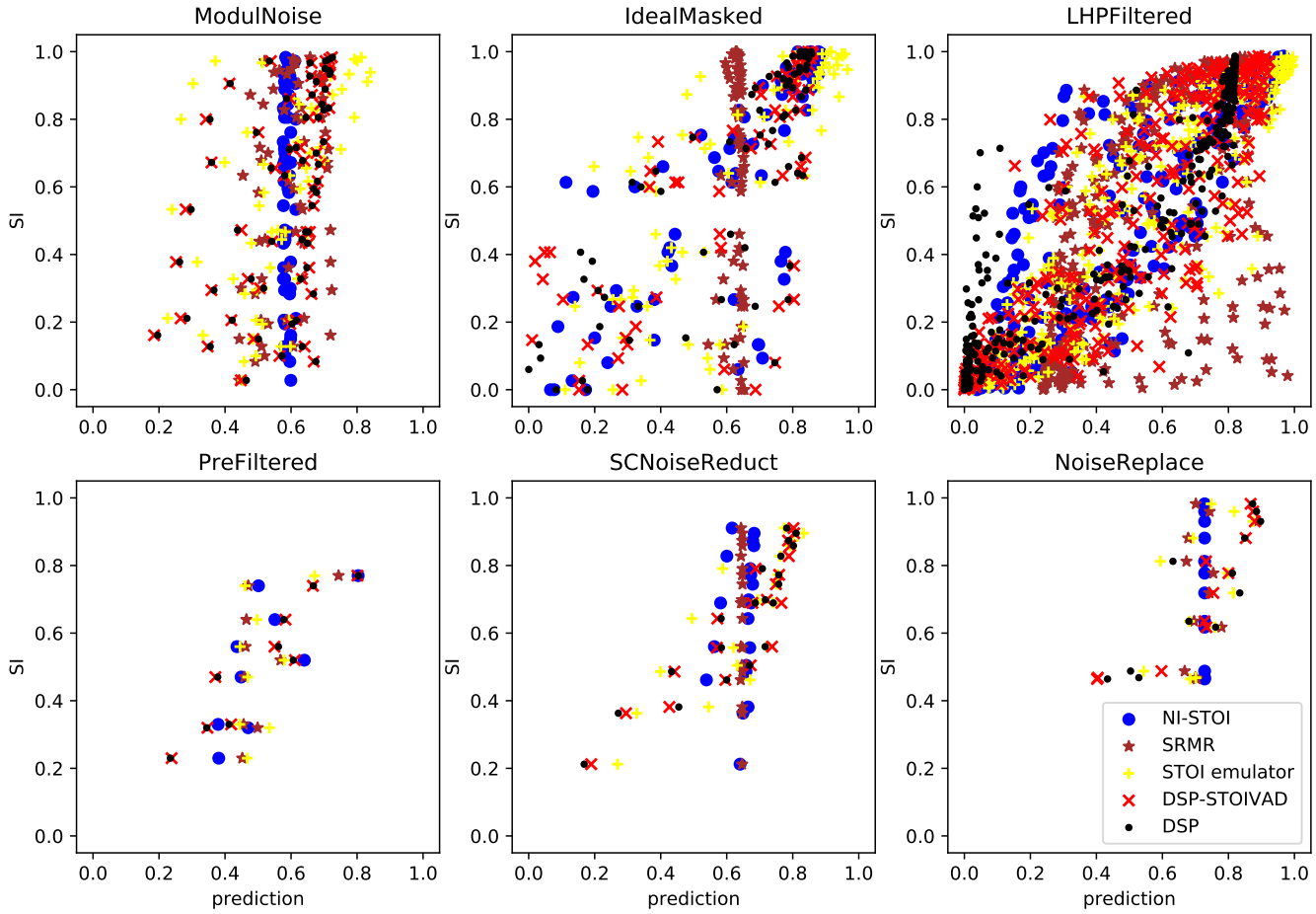


Fig. 2. Absolute measured SI vs. logistically transformed SI predictions.

in temporally modulated noise has been noted to require special consideration during the design of the predictor [21, 28]. It is less obvious to us why SRMR fails on *IdealMasked* and *SCNoiseReduct*. However, these datasets both consist of de-noised speech ad it is possible that de-noising artefacts, or distortions to the underlying speech are what causes SRMR to fail.

C. SI prediction performance evaluation

1) Spearman correlations: Table IV displays the Spearman correlation coefficients specific to each predictor and dataset. The highest score for each dataset is marked in bold face. Among the non-intrusive SI predictors, DSP reaches the highest Spearman correlation four out of six datasets, as well as on average.

2) Pearson correlations: Table V displays the Pearson correlation coefficients. The highest correlation for each dataset is in bold face. Among the non-intrusive SI predictors, the highest Pearson correlation is achieved by DSP on 4 of 6 listening test datasets as well as on average. Williams' t-test [50] was applied and each correlation coefficient, which is not significantly lower than the highest, is marked with an asterisk. The overall significance level is chosen to be $\alpha = 0.05$, and using Bonferroni correction the individual significance level is $0.05/5 = 0.01$. This correction is made to account for the fact that the test must be performed multiple times, i.e., once for each predictor. The correction ensures that the chosen significance level of 0.05 covers the test as a whole. The test shows that there is no significant improvement of DSP's performance when rVAD is replaced with an ideal, intrusive VAD.

3) Root mean squared errors: Table VI displays the RMSE of the logistically transformed predictions. The lowest RMSE for each dataset is marked in bold face. Among the non-intrusive SI predictors, DSP reaches the lowest RMSE on four of the datasets as well as on average, and the STOI emulator achieves the lowest RMSE on three of the datasets.

4) Comparison to baseline SI predictors: It is clear from Tables IV, V and VI that the data-driven SI predictors, i.e., DSP and the STOI emulator, have the highest performance. DSP scores the highest correlations and has the lowest RMSE on four the datasets. This supports the hypothesis that there is indeed a strong link between between SPP and SI, and that a data-driven SPP estimator can be successfully applied to predict SI. Furthermore, despite being trained exclusively on speech in additive noise, DSP performs well in several non-linear processing conditions. This is important, because it suggests that a data-driven SI predictor may be trained using inexpensive, abundant speech signals and additive noise sources, and still be applied to noisy signals subjected to non-linear processing such as speech enhancement algorithms. The main advantage is the ease with which the training data can be expanded to cover new types of noise and processing. The post-processing stage plays an important part in why DSP is able to generalize well

TABLE IV
SPEARMAN CORRELATIONS

	ModulNoise	*IdealMasked*	*LHPFiltered*	*PreFiltered*	*SCNoiseReduct*	*NoiseReplace*	Avg.
DSP	0.56	0.81	**0.93**	**0.93**	**0.93**	0.88	**0.84**
DSP-STOIVAD	0.58	0.81	0.87	**0.93**	**0.93**	**0.92**	**0.84**
STOI emul.	**0.59**	0.80	0.92	0.25	0.87	0.48	0.65
NI-STOI	-0.01	**0.84**	0.88	0.70	0.47	-0.06	0.47
SRMR	0.22	0.27	0.61	0.60	0.21	0.20	0.35

TABLE V
PEARSON CORRELATIONS - ASTERISKS MARK THE BEST PERFORMING SI PREDICTORS, WHICH CAN NOT BE DISTINGUISHED BETWEEN WITH STATISTICAL SIGNIFICANCE

	ModulNoise	*IdealMasked*	*LHPFiltered*	*PreFiltered*	*SCNoiseReduct*	*NoiseReplace*	Avg.
DSP	0.46*	0.77*	**0.90***	**0.94***	**0.90***	0.88*	**0.81**
DSP-STOIVAD	0.48*	0.73*	0.87	0.93*	**0.90***	**0.91***	**0.81**
STOI emul.	**0.51***	**0.78***	**0.90***	0.41	0.83*	0.52*	0.66
NI-STOI	-0.04	0.77*	0.87	0.71*	0.22	0.00	0.42
SRMR	0.27*	-0.06	0.56	0.51	0.01	0.33*	0.27

TABLE VI
ROOT MEAN SQUARED ERRORS

	ModulNoise	*IdealMasked*	*LHPFiltered*	*PreFiltered*	*SCNoiseReduct*	*NoiseReplace*	Avg.
DSP	0.26	**0.21**	0.16	0.07	0.09	0.09	**0.15**
DSP-STOIVAD	0.26	0.23	0.17	**0.06**	**0.08**	**0.08**	**0.15**
STOI emul.	**0.25**	**0.21**	**0.15**	0.16	0.11	0.16	0.17
NI-STOI	0.29	**0.21**	0.17	0.13	0.19	0.18	0.19
SRMR	0.28	0.33	0.29	0.15	0.19	0.17	0.23

to the non-linear processing conditions. This is because tiles with high uncertainty, which can be expected to be more common when applying the network to unseen conditions, are eliminated. This is supported by Figure 4 described in Sec. VI-E. That being said, it is hard to guarantee that this particular trained instance of Deep Speech Presence will work well under any specific acoustic/processing condition outside the scope of our evaluation test set, as is in general the case for data-driven methods. Indeed, none of the SI predictors included in our study, Deep Speech Presence included, perform particularly well on *ModulNoise*. However, we argue as one of the main findings of the paper, that the training set for the proposed SPP based type of SI prediction can be relatively easily extended for a new acoustic scenario, since new listening test data is not required.

5) The effects from VAD: From Figure 2 it is clear that there is a great deal of similarity between DSP using rVAD and the ideal, intrusive STOIVAD, evidenced by the fact that they produce nearly identical predictions for many conditions. However, the difference between using either VAD becomes particularly apparent for *IdealMasked* and *LHPFiltered*, where the predictions differ noticeably. The datasets *IdealMasked* and *LHPFiltered* happen to include longer segments without speech, than the other datasets, which means that rVAD has more opportunities for false positives, which could explain why the SI predictions of the two versions of DSP differ on these datasets in particular.

Note, that there is no statistically significant difference between the results achieved by DSP using rVAD and the ideal, intrusive STOIVAD, cf. Table V. This result is surprising as it would be reasonable to expect an increase in performance when using an ideal VAD.

D. Performance vs. duration of input

The results reported so far, in Tables IV, V and VI, are based on the full duration of available speech in each condition, which varies from listening test to listening test. For certain practical applications of SI predictors, such as an SI prediction stage in the processing of a wearable device, e.g., a hearing aid, SI predictions based on short input signals are desirable. This is so, because it would allow the device to adapt its processing as a function of the predicted SI, even in rapidly changing acoustic environments. In order to give a better idea of the duration of input speech signal that is required for the predictions of DSP to be reliable, an additional simulation experiment is performed. In this experiment, a segment of fixed duration is drawn at random from each available listening test condition, and the Spearman correlation is recomputed based only on the SI predictions made by DSP on these fixed-duration segments. This is done for a range of durations between 384 ms, which is the shortest duration supported by DSP, up to 64 seconds. For each fixed duration, 200 separate random trials are performed, where a fixed-duration segment is drawn with replacement, from the speech available for each individual condition. The SI predictions and correlation with Listening test measurements are computed individually for each trial.

Figure 3 shows the average Spearman correlation coefficient across the 200 trials, as well as the 5'th, 25'th, 75'th and 95'th percentiles as a function of duration for each dataset. It can be noted that the Spearman correlation for the dataset *LHPFiltered* converges and stabilizes very quickly, which may be due to the fact that the distortion in this dataset is high- and low-pass filtering of the speech, which does not vary across time. Conversely, the dataset *NoiseReplace* is contaminated by noise that fluctuates over time, i.e. segments of speech have been replaced with noise. Indeed, the Spearman correlation for this dataset stabilizes more slowly.

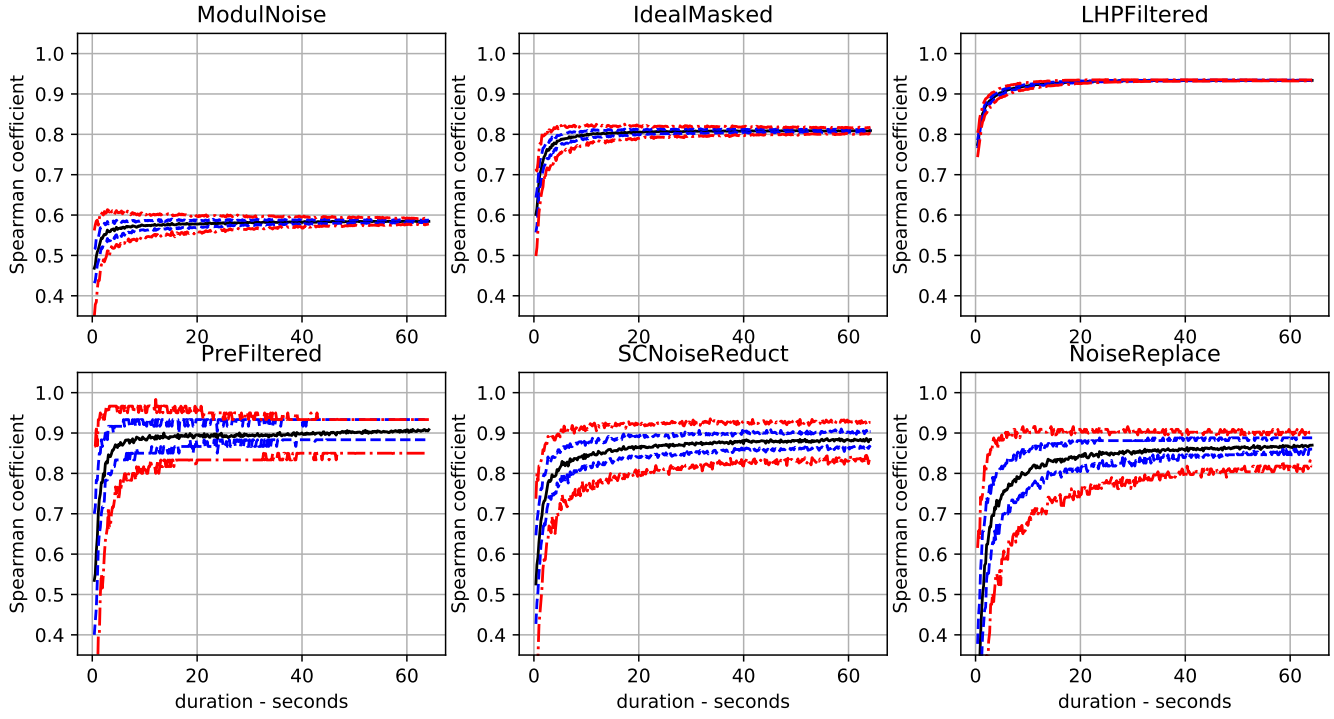


Fig. 3. Performance in terms of Spearman correlation as a function of available input duration. The curves show the average Spearman correlations of 200 trials pr. duration in black, as well as the 25'th and 75'th percentiles in blue and the 5'th and 95'th percentiles in red.

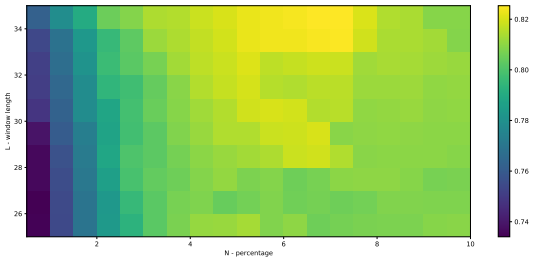


Fig. 4. Average Spearman correlation as a function of post processing parameters, p and N.

From Figure 3 it can be concluded that the minimum required input duration for reliable SI predictions depends on the type of noise and processing or distortion. In the presence of temporally modulated noise, such as that found in *ModulNoise* and *NoiseReplace*, 20 to 30 seconds may be required, whereas speech in stationary noise, such as is found in *LHPFiltered*, can be reliably predicted with as little as 10 seconds of input. A similar conclusion was made in [21] for the intrusive ESTOI, when evaluated with speech signals contaminated by modulated additive noise.

E. Robustness of the Post processing stage

As mentioned in Sec III-C, the performance scores reported in Tables IV, V and VI are achieved with the following values of the post processing parameters: The percentage $p=5$, and the window length $N = 30$. It would be cause for concern if the reported performance was sensitive to these parameters, since this would mean that the post-processing stage would likely not generalize well to new data with the selected parameter values. A sweep of these parameters, using the full listening test dataset summarized in Table III, is performed to ensure that this is not the case. Figure 4 shows the average performance of DSP, in terms of Spearman correlation across datasets as a function of the post processing parameters p and N. This analysis shows that perturbations of the used values, $p=5$ and $N=30$, of these parameters have little impact on the SI prediction performance of DSP on average. From Figure 4 it is apparent that the average Spearman correlation is above 0.8 for $4<p<8$ and depends very little on N. In conclusion, DSP appears to be highly insensitive to perturbations of the post processing parameters p and N.

VII. Conclusion

The scarcity of intelligibility data is a limiting factor for the development of data-driven SI predictors. In this study, we hypothesized a correlation between Speech Presence Probability (SPP) and Speech Intelligibility (SI), and showed that a data-driven SPP estimator can be used for non-intrusive SI prediction. This was done by training a deep neural network to perform SPP estimation, and then applying a simple post processing scheme to map the estimated SPPs to SI predictions. The advantages of this approach to SI prediction is that training data can be generated easily in abundance and variety, without the need to carry out time consuming listening tests. DSP is trained on a dataset constrained to speech in additive noise, but is demonstrated to generalize well to a variety of non-linear processing conditions as well. Specifically, the SI prediction performance of DSP was evaluated on a range of listening test datasets consisting of speech signals contaminated by a variety of noise types and subject to various non-linear processing schemes, and compared

to state-of-the-art non-intrusive SI predictors including NISTOI and SRMR. A data-driven, non-intrusive emulator of the intrusive SI predictor, STOI, was also included in the comparison as a baseline. The proposed DSP SI predictor reached the highest SI prediction performance in terms of Spearman correlation, Pearson correlation and root mean squared error on a majority of listening tests, validating the hypothesized link between SPP and SI, and at the same time showing potential for generalizability to different types of noise and processing than those present in the training set. The required duration of input speech for stable predictions of SI, which is important for potential real-time applications of such an SI predictor, was determined experimentally to be ten to twenty seconds, depending on the temporal properties of the noise.

References

[1] Y. Feng and F. Chen, "Nonintrusive Objective Measurement of Speech Intelligibility: A Review of Methodology," *Biomedical Signal Processing and Control*, vol. 71, p. 103204, 2022.

[2] M. B. Pedersen, A. H. Andersen, S. H. Jensen, and J. Jensen, "A Neural Network for Monaural Intrusive Speech Intelligbility Prediction," *ICASSP*, pp. 336–340, May 2020.

[3] 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, 1947.

[4] 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, Sep. 2011.

[5] T. H. Falk, C. Zheng, and W. Chan, "A Non-Intrusive Quality and Intelligibility Measure of Reverberant and Dereverberated Speech," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 18, no. 7, pp. 1766–1774, Aug. 2010.

[6] 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.

[7] 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, Language Process.*, vol. 26, no. 10, pp. 1925–1939, Oct. 2018.

[8] M. Karbasi, S. Bleeck, and D. Kolossa, "Non-Intrusive Speech Intelligibility Prediction using Automatic Speech Recognition derived Measures," *arXiv preprint arXiv:2010.08574*, 2020.

[9] 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," *INTERSPEECH*, Oct. 2020.

[10] M. B. Pedersen, A. H. Andersen, S. H. Jensen, Z. H. Tan, and J. Jensen, "Training Data-Driven Speech Intelligibility Predictors on Heterogeneous Listening Test Data," *IEEE Access*, vol. 10, pp. 66 175–66 189, Jun. 2022.

[11] M. Karbasi, A. H. Abdelaziz, and D. Kolossa, "Twin-HMM-based Non-Intrusive Speech Intelligibility Prediction," in *ICASSP 2016*, Mar. 2016, pp. 624–628.

[12] Y.-H. Tu, J. Du, and C.-H. Lee, "Speech Enhancement based on Teacher–Student Deep Learning using Improved Speech Presence Probability for Noise-Robust Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2080–2091, 2019.

[13] M. Tammen, D. Fischer, B. T. Meyer, and S. Doclo, "DNN[-based Speech Presence Probability Estimation for Multi-Frame Single-Microphone Speech Enhancement," in *ICASSP 2020-2020 IEEE International conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2020, pp. 191–195.

[14] A. H. Andersen, J. M. de Haan, Z. H. Tan, and J. Jensen, "A Non-Intrusive Short-Time Objective Intelligibility Measure," in *ICASSP 2017*, Mar. 2017, pp. 5085–5089.

[15] R. E. Zezario, S.-W. Fu, C.-S. Fuh, Y. Tsao, and H.-M. Wang, "STOI-Net: A Deep Learning based Non-Intrusive Speech Intelligibility Assessment Model," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, Dec. 2020, pp. 482–486.

[16] T. Houtgast and H. J. Steeneken, "Evaluation of Speech Transmission Channels by using Artificial Signals," *Acta Acust. united Ac.*, vol. 25, no. 6, pp. 355–367, Dec. 1971.

[17] A. N. S. Institute, *American National Standard: Methods for Calculation of the Speech Intelligibility Index*. Acoustical Society of America, 1997.

[18] 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.

[19] M. Elhilali, T. Chi, and S. A. Shamma, "A Spectro-Temporal Modulation Index (STMI) for Assessment of Speech Intelligibility," *Speech Commun.*, vol. 41, no. 2-3, pp. 331–348, Oct. 2003.

[20] M. Cooke, "A Glimpsing Model of Speech Perception in Noise," *J. Acoust. Soc. Amer.*, vol. 119, no. 3, pp. 1562–1573, Feb. 2006.

[21] 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.

[22] ——, "Speech Intelligibility Prediction based on Mutual Information," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 2, pp. 430–440, Jan. 2014.

[23] J. Taghia and R. Martin, "Objective Intelligibility Measures based on Mutual Information for Speech Subjected to Speech Rnhancement Processing," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 1, pp. 6–16, Sep. 2013.

[24] S. V. 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.

[25] K. Kondo, K. Taira, and Y. Kobayashi, "Binaural Speech Intelligibility Estimation Using Deep Neural Networks," *Interspeech*, pp. 1858–1862, Sep. 2018.

[26] R. E. Zezario, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "MBI-Net: A Non-Intrusive Multi-Branched Speech Intelligibility Prediction Model for Hearing Aids," *arXiv preprint arXiv:2204.03305*, 2022.

[27] P. Seetharaman, G. J. Mysore, P. Smaragdis, and B. Pardo, "Blind Estimation of the Speech Transmission Index for Speech Quality Prediction," *ICASSP*, pp. 591–595, Apr. 2018.

[28] M. R. Schädler, A. Warzybok, S. D. Ewert, and B. Kollmeier, "A Simulation Framework for Auditory Discrimination Experiments: Revealing the Importance of Across-Frequency

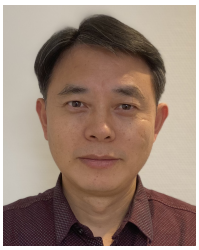
Processing in Speech Perception," *The journal of the acoustical society of America*, vol. 139, no. 5, pp. 2708–2722, 2016.
[29] M. Karbasi and D. Kolossa, "ASR-based Measures for Microscopic Speech Intelligibility Prediction," in *Proc. 1st Int. Workshop Challenges Hear. Assistive Technol.*, Aug. 2017.
[30] L. Fontan, T. Cretin-Maitenaz, and C. Füllgrabe, "Predicting Speech Perception in Older Listeners with Sensorineural Hearing Loss using Automatic Speech Recognition," *Trends in hearing*, vol. 24, 2020.
[31] J. Roßbach, B. Kollmeier, and B. T. Meyer, "A Model of Speech Recognition for Hearing-Impaired Listeners based on Deep Learning," *The Journal of the Acoustical Society of America*, vol. 151, no. 3, pp. 1417–1427, 2022.
[32] Z. Tu, N. Ma, and J. Barker, "Unsupervised Uncertainty Measures of Automatic Speech Recognition for Non-Intrusive Speech Intelligibility Prediction," *arXiv preprint arXiv:2204.04288*, 2022.
[33] S. Kay, *Intuitive Probability and Random Processes using MATLAB®*. Springer Science & Business Media, 2006.
[34] 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*, vol. 9, no. 1, pp. 1–80, 2013.
[35] Z. H. Tan, A. K. Sarkar, N. Dehak *et al.*, "rVAD: An Unsupervised Segment-based Robust Voice Activity Detection Method," *Computer speech & language*, vol. 59, pp. 1–21, Jan. 2020.
[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Jun. 2016, pp. 770–778.
[37] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv:1412.6980*, Dec. 2014.
[38] J. B. Nielsen and T. Dau, "Development of a Danish Speech Intelligibility Test," *Int. J. Audiol.*, vol. 48, no. 10, pp. 729–741, 2009.
[39] N. Li and P. C. Loizou, "Factors Influencing Intelligibility of Ideal Binary-Masked Speech: Implications for Noise Reduction," *J. Acoust. Soc. Am.*, vol. 123, no. 3, pp. 1673–1682, 2008.
[40] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The Third 'CHiME' Speech Separation and Recognition Challenge: Dataset, Task and Baselines," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, Oct. 2015, pp. 504–511.
[41] S. Graetzer, J. Barker, T. J. Cox, M. Akeroyd, J. F. Culling, G. Naylor, E. Porter, and R. e. a. Viveros Munoz, "Clarity-2021 Challenges: Machine Learning Challenges for Advancing Hearing Aid Processing," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2. International Speech Communication Association (ISCA), Aug. 2021, pp. 686–690.
[42] J. Thiemann, N. Ito, and E. Vincent, "DEMAND: a Collection of Multi-Channel Recordings of Acoustic Noise in Diverse Environments," 2013.
[43] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. L. Wang, "Role of Mask Pattern in Intelligibility of Ideal Binary-Masked Noisy Speech," *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1415–1426, Sep. 2009.
[44] 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.
[45] T. P. Knudsen, "Predictability-based Objective Evaluation of Sound," *Aalborg University Denmark*, Jun. 2021.
[46] 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.
[47] 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.
[48] D. L. Wang, "On Ideal Binary Mask as the Computational Goal of Auditory Scene Analysis," in *Speech separation by humans and machines*. Springer, 2005, pp. 181–197.
[49] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie, "Objective Quality and Intelligibility Prediction for Users of Assistive Listening Devices: Advantages and Limitations of Existing Tools," *IEEE signal processing magazine*, vol. 32, no. 2, pp. 114–124, Feb. 2015.
[50] E. J. Williams, "The Comparison of Regression Variables," *J. Royal Stat. Society, Ser. B*, vol. 21, no. 2, pp. 396–399, 1959.

Mathias B. Pedersen (M'18) received the B.Sc. and M.Sc. degree in mathematical engineering from Aalborg University, Aalborg, Denmark, in 2016 and 2018 respectively. He received the Ph.D. degree from Aalborg University, Aalborg, Denmark, in 2023.

Mathias is currently employed at Norlys Energy Trading A/S as a quantitative researcher.

Søren Holdt Jensen (S'87–M'88–SM'00) 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.

Dr. Jensen is currently Principal Research Scientist at CHORA, Aarhus, Denmark, contributing to developments of signal processing solutions for Electronic Warfare Systems. From 1988 to 1990, he was a Member of the Technical Staff with the Telecommunications Laboratory of Telecom Denmark, Ltd, Taastrup (Copenhagen), Denmark, working on specification of the second generation (2G) mobile phone system. From 1990 to 1995 he worked in various research positions within signal processing, numerical linear algebra and algorithms at the Electronics Institute of DTU, the Scientific Computing Group of the Danish Computing Center for Research and Education (UNI·C), Lyngby, and the Electrical Engineering Department of Katholieke Universiteit Leuven (KU Leuven), Leuven, Belgium. From 1995 to 2021 he held positions as Assistant, Associate and Full Professor at the Department of Electronic Systems at AAU in the areas of digital communications, speech and signal processing. From 2021 to 2023 he was a Special Adviser to the Danish Defense and the Danish Ministry of Defense (ministerial department) on noise, acoustics and vibration. He is co-author 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. Dr. Jensen has been Associate Editor for the IEEE Transactions on Signal Processing, IEEE/ACM Transactions on Audio, Speech and Language Processing, Elsevier Signal Processing, and EURASIP Journal on Advances in Signal Processing. He is a recipient of an individual European Community Marie Curie Fellowship, former Chairman of the IEEE Denmark Section and the IEEE Denmark Section's Signal Processing Chapter (founder and first Chairman). From 2011 to 2016 he was member of the Danish Council for Independent Research appointed by Danish Minsters of Higher Education and Science. He is member of the Danish Academy of Technical Sciences (ATV) and since 2006 appointed External Examiner by the Ministry of Higher Education and Science for engineering programmes within Mathematics, Physics, Electronics, IT and Energy.



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 (SJTU), Shanghai, China, in 1999.

He is a Professor in the Department of Electronic Systems and a Co-Head of the Centre for Acoustic Signal Processing Research at Aalborg University, Aalborg, Denmark. He is also a Co-Lead of the Pioneer Centre for AI, Denmark. He was a Visiting Scientist at the Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, USA, an Associate Professor at the Department of Electronic Engineering, SJTU, Shanghai, China, and a postdoctoral fellow at the AI Laboratory, KAIST, Daejeon, Korea. 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 has (co)-authored over 200 refereed publications. He is the Chair of the IEEE Signal Processing Society Machine Learning for Signal Processing Technical Committee (MLSP TC). He is an Associate Editor for the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING. He has served as an Associate/Guest Editor for several other journals. He was the General Chair for IEEE MLSP 2018 and a TPC Co-Chair for IEEE SLT 2016.

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 Assistant Research Professor. From 2000 to 2007, he was a Post-Doctoral Researcher and Assistant Professor with Delft University of Technology, Delft, The Netherlands, and an External Associate Professor with Aalborg University. Currently, he is 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 a Professor with the Section for Artificial Intelligence and Sound (AIS), Department of Electronic Systems, at Aalborg University. He is also a co-founder of the Centre for Acoustic Signal Processing Research (CASPR) at Aalborg University. His main interests are in 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