Non-Intrusive Speech Quality Prediction Using Modulation Energies and LSTM-Network

Benjamin Cauchi, *Student Member, IEEE*, Kai Siedenburg, João F. Santos, Tiago H. Falk, *Senior Member, IEEE*, Simon Doclo, *Senior Member, IEEE*, and Stefan Goetze, *Member, IEEE*

***Abstract*—Many signal processing algorithms have been proposed to improve the quality of speech recorded in the presence of noise and reverberation. Perceptual measures, i.e., listening tests, are usually considered the most reliable way to evaluate the quality of speech processed by such algorithms but are costly and time-consuming. Consequently, speech enhancement algorithms are often evaluated using signal-based measures, which can be either intrusive or non-intrusive. As the computation of intrusive measures requires a reference signal, only non-intrusive measures can be used in applications for which the clean speech signal is not available. However, many existing non-intrusive measures correlate poorly with the perceived speech quality, particularly when applied over a wide range of algorithms or acoustic conditions. In this paper, we propose a novel non-intrusive measure of the quality of processed speech that combines modulation energy features and a recurrent neural network using long short-term memory cells. We collected a dataset of perceptually evaluated signals representing several acoustic conditions and algorithms and used this dataset to train and evaluate the proposed measure. Results show that the proposed measure yields higher correlation with perceptual speech quality than that of benchmark intrusive and non-intrusive measures when considering various categories of algorithms. Although the proposed measure is sensitive to mismatch between training and testing, results show that it is a useful approach to evaluate specific algorithms over a wide range of acoustic conditions and may, thus, become particularly useful for real-time selection of speech enhancement algorithm settings.**

***Index Terms*—Speech quality, non-intrusive prediction, modulation energy, LSTM-network.**

Manuscript received December 21, 2018; revised March 26, 2019; accepted April 9, 2019. Date of publication April 18, 2019; date of current version May 7, 2019. This work was supported in part by the projects Audio-PSS (www.audio-pss.de) and THERESIAH, funded by the Bundesministerium für Bildung und Forschung (BMBF) under Grants 02K16C201 and 13GW0209B, in part by the EU Seventh Framework Programme project DREAMS under Grant ITN-GA-2012-316969, and in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project ID 390895286–EXC 2177/1. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Richard Christian Hendriks. *(Corresponding author: Benjamin Cauchi.)*

B. Cauchi was with the Fraunhofer Institute for Digital Media Technology, 98693 Ilmenau, Germany. He is now with the OFFIS Institute for Information Technology, 26121 Oldenburg, Germany (e-mail: benjamin.cauchi@offis.de).

K. Siedenburg and S. Doclo are with the Department of Medical Physics and Acoustics and Cluster of Excellence Hearing4all, University of Oldenburg, 26129 Oldenburg, Germany (e-mail: kai.siedenburg@uni-oldenburg.de; simon.doclo@uni-oldenburg.de).

J. F. Santos and T. H. Falk are with the Institut National de la Recherche Scientifique (INRS-EMT), University of Québec, Montréal, QC H5A-1K6, Canada (e-mail: jfsantos@emt.inrs.ca; falk@emt.inrs.ca).

S. Goetze is with the Fraunhofer Institute for Digital Media Technology, 98693, Ilmenau, Germany, and also with the Cluster of Excellence Hearing4all, University of Oldenburg, 26129 Oldenburg, Germany (e-mail: s.goetze@idmt.fraunhofer.de).

Digital Object Identifier 10.1109/TASLP.2019.2912123

I. Introduction

In many speech communication applications, such as teleconferencing or hearing-aids, speech of a distant user is recorded by a single or by multiple microphones. Often, the recorded speech signal is corrupted by ambient noise and reverberation, which may severely degrade the perceived speech quality and speech intelligibility. To overcome these effects, many speech enhancement algorithms have been proposed [1]–[4]. Although many algorithms are able to substantially reduce the amount of noise and reverberation in the recorded signal, the choice of the best suited algorithm often depends on the acoustic condition, and processing artefacts may result in a degradation of speech quality [5]. Consequently, speech enhancement algorithms need to be evaluated in terms of speech intelligibility and quality.

Perceptual speech quality evaluation based on listening tests requires a group of human assessors to evaluate the processed speech signals with respect to predefined attributes, such as overall quality, level of reverberation or residual noise, or coloration. Such evaluation is typically performed by grading each attribute on a scale which either consists of a few values, such as for the mean opinion score (MOS) [6], or continuous values, as in the multiple stimuli test with hidden reference and anchor (MUSHRA) [7]. Speech intelligibility can be assessed as the number of speech items, i.e., phonemes or words, identified by assessors in relation to the total number of items present in the signal under test [8]. Speech intelligibility is often reported using the speech reception threshold (SRT), i.e., the level of degradation for which only 50% of the speech items are correctly identified by an assessor [9]. Perceptual measures are generally considered the most reliable way to assess the quality or intelligibility of processed speech signals. However, since these measures are costly and time-consuming, speech enhancement algorithms are often evaluated using signal-based measures.

Signal-based measures, aiming at either speech quality or speech intelligibility prediction, can be categorized as either intrusive or non-intrusive. The computation of intrusive measures requires a (clean) reference signal in addition to the target signal under test, whereas non-intrusive measures can be computed from the target signal only. Among intrusive measures, the articulation index (AI) [10], the speech transmission index (STI) [11], the speech intelligibility index (SII) [12], the short-time objective intelligibility (STOI) [13] and mutual-information-based techniques, such as the algorithm proposed in [14], aim at speech intelligibility prediction, whereas the perceptual evaluation of

speech quality (PESQ) [15], the perceptual objective listening quality assessment (POLQA) [16] or the perception model for quality (PEMO-Q) [17] are used to predict the speech quality. However, in practice, a reference signal is not available, e.g., to evaluate algorithms using realistic corpora or to automatically select the best algorithm for a specific acoustic condition. Consequently, reliable non-intrusive measures are required.

Several measures have been proposed to remove the need for a reference signal. Non-intrusive measures of the speech intelligibility include the recently proposed non-intrusive STOI (NI-STOI) [18], that relies on estimating the amplitude envelope of the clean speech from the input signal, and the use of a trained speech recognizer as proposed in [19]. To evaluate speech quality, non-intrusive measures such as P.563 (P.563) [20] and ANIQUE+ (ANIQUE+) [21] exist, which have not been explicitly developed for the evaluation of speech enhancement algorithms but rather for the evaluation of narrow-band speech codecs. Measures such as the signal to reverberant modulation ratio (SRMR) [22] and its extension, the normalized SRMR (SRMR$_{norm}$) [23], have been developed for both intelligibility and quality prediction and apply a rather simple predicting function to a set of time-averaged modulation energies. Though this measure has shown promising results, e.g., when predicting intelligibility for cochlear implants users in [24], its performance, similarly as for P.563 and ANIQUE+, can be unpredictable when applied to signals processed with different categories of algorithms, as reported in [25]. In [26], twin hidden Markov models (HMMs) have been proposed to generate an estimate of the clean speech before using this estimate and the signal under test as input to an intrusive measure. The reliability of the obtained prediction largely depends on the accuracy of the estimated clean signal such that the method does not outperform the used intrusive measure. Recent approaches have focused on applying machine learning techniques to train a predicting function for speech quality. The approach in [27] uses a classification and regression tree (CART) algorithm while the approach in [28] uses a similar combination of classification and regression with a so-called model-tree [29]. In both [27] and [28], the predicting function does not take into account the time dependencies within the target signal and the evaluation of both approaches was limited by the datasets used for training. Indeed, [27] only uses data labeled using existing signal-based measures and no perceptually evaluated data, whereas [28], used perceptually evaluated data but trained the predicting function separately, in turn, for each acoustic condition.

This paper proposes a novel non-intrusive measure aiming at reliably predicting the speech quality of processed signals across various acoustic conditions and types of processing. For this purpose, we use a predicting function that takes the time dependency of the target signal into account and is trained on a perceptually evaluated dataset.

This paper is structured as follows. In Section II, we present the proposed measure, which combines modulation energy (ME) and a recurrent neural network (RNN) using long short-term memory (LSTM) cells. Using such network as predicting function allows to model the time dependency of the proposed signal and to apply the proposed measure to signals of arbitrary length. In Section III, we describe the perceptually evaluated dataset of speech signals, representing various acoustic conditions, i.e., room impulse responses (RIRs), noise types and signal-to-noise ratios (SNRs), and several categories of algorithms, single- and multichannel, with different settings resulting in various level of interference suppression and processing artefacts. In Section IV, we describe how we used this dataset to train and evaluate the proposed measure and we present our experimental framework and the considered benchmark. The results presented in Section V show that the proposed measure, when trained for a single category of algorithms, outperforms existing non-intrusive measures and yields similar performance as the intrusive measures. When considering several categories of algorithms, the proposed measure outperforms both non-intrusive and intrusive existing measures.

II. Proposed Approach

The time-domain signal $y_m(n)$ recorded in the m-th microphone of $M \geq 1$ available microphones can be modeled as

$$y_m(n) = x_m(n) + v_m(n) = s(n) * h_m(n) + v_m(n), \quad (1)$$

where n denotes the sample index, $s(n)$ denotes the anechoic speech signal, $h_m(n)$ denotes the RIR of length L_h between the source and the m-th microphone and $v_m(n)$ denotes the additive noise component. The reverberant speech component $x_m(n)$ can be written as

$$x_m(n) = d_m(n) + r_m(n), \quad (2)$$

where

$$d_m(n) = s(n) * h_m^d(n), \quad (3)$$

$$r_m(n) = s(n) * h_m^r(n), \quad (4)$$

with $h_m^d(n)$ and $h_m^r(n)$ defined as

$$h_m^d(n) = \begin{cases} h_m(n) \text{ if } n \leq L_d, \\ 0 \text{ otherwise}, \end{cases} \quad (5)$$

$$h_m^r(n) = \begin{cases} h_m(n) \text{ if } n > L_d, \\ 0 \text{ otherwise}, \end{cases} \quad (6)$$

where L_d is set so that $h_m^d(n)$ contains the direct path and a few early reflections while $h_m^r(n)$ contains the late reflections, i.e., the reverberant tail. The output signal $\hat{s}(n)$ of a speech enhancement algorithm is computed from the recorded microphone signals $y_m(n)$, $m \in [1 \cdots M]$, as an estimate of either $s(n)$ or $d_{\mathrm{ref}}(n)$, where ref denotes the index of a reference microphone.

The perceived speech quality $p_{\hat{s}}$ of the processed signal $\hat{s}(n)$ can be obtained from a listening test conducted with several assessors (cf. Section III-B). The measure proposed in this paper aims at non-intrusively predicting $p_{\hat{s}}$, i.e., at computing an estimate $\hat{p}_{\hat{s}}$ of $p_{\hat{s}}$ from the signal $\hat{s}(n)$ while requiring neither a listening test nor or reference signal, i.e., $s(n)$ or $d_{\mathrm{ref}}(n)$. The

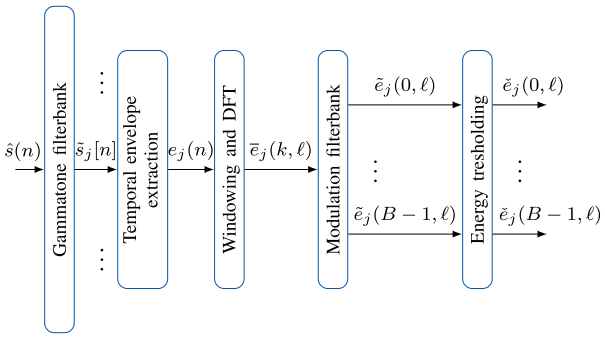

Fig. 1. Overview of the feature extraction for one time frame. Contrary to previous works [23], [28], [30], the proposed measure does not average the features over time before applying the predicting function.

proposed measure relies on extracting a set of time-varying features which are used as input to a predicting function. Section II-A describes the considered features before presenting the RNN used as predicting function in Section II-B.

A. Considered Features

The proposed measure uses modulation energies (MEs) as features, which have already been used in the field of speech quality prediction in [30] and further elaborated in [23], [28]. The computation of these features is depicted in Fig. 1 and can be summarized as follows.

First, the signal under test $\hat{s}(n)$ is filtered by a gammatone filterbank with J channels, resulting in J filtered signals $\tilde{s}_j[n]$, where j denotes the filter index. The temporal envelope $e_j(n)$ is extracted from $\tilde{s}_j[n]$ as

$$e_j(n) = \sqrt{\tilde{s}_j^2[n] + \mathcal{H}\left\{\tilde{s}_j[n]\right\}^2}, \tag{7}$$

where $\mathcal{H}\{\cdot\}$ denotes the Hilbert transform. The temporal envelopes are divided into $L = \lceil N/(W-O) \rceil$ overlapping windowed frames using an overlap of O samples and a window of length W, with N the signal length. The modulation spectral energy $\overline{e}_j(k,\ell)$ is computed as the squared magnitude of the discrete Fourier transform of the ℓ-th frame in the k-th modulation frequency bin.

The modulation spectral energies $\overline{e}_j(k,\ell)$, for frequency bins in the interval $k_{\min}$ to $k_{\max}$, are warped into B overlapping modulation bands whose centre frequencies are set as in [23], resulting in the warped modulation energies $\tilde{e}_j(b,\ell)$, where b denotes the index of the modulation band. As proposed in [23], thresholding is applied to $\tilde{e}_j(b,\ell)$, resulting in $\alpha e_{\text{peak}} \leq \check{e}_j(b,\ell) \in\leq e_{\text{peak}}$, where $0 \leq \alpha < 1$ and where

$$e_{\text{peak}} = \max_{j,b}\left(\frac{1}{L}\sum_{\ell=0}^{L-1}\tilde{e}_j(b,\ell)\right). \tag{8}$$

Finally, a feature vector $\mathbf{e}_\ell$ of length $J \cdot B$ is constructed for each time frame as

$$\begin{aligned}\mathbf{e}_\ell = \big[&\check{e}_0(0,\ell),\ldots,\check{e}_0(B-1,\ell),\ldots,\\ &\check{e}_{J-1}(0,\ell),\ldots,\check{e}_{J-1}(B-1,\ell)\big]^{\mathrm{T}},\end{aligned} \tag{9}$$

where superscript T denotes the transpose operator.

In this paper, the different parameters of the feature extraction have been set as in [23]. The gammatone filterbank is applied to signals downsampled to 8 kHz and uses $J = 23$ channels with center frequencies ranging from 125 Hz to 4 kHz. The modulation frequency bins are grouped into $B = 8$ bands. The temporal envelope $e_j(n)$ is divided in frames using a Hamming window of length W corresponding to 256 ms and an overlap of length O corresponding to 224 ms. The indices $k_{\min}$ and $k_{\max}$ correspond to the range of modulation frequencies between 4 Hz and 40 Hz and α is set to lower bound modulation energies 30 dB below e_{peak}. These values have been shown to reduce the sensitivity of the extracted features to speakers and pitch content [23] compared to the settings initially proposed in [30].

It can be noted that previous use of the ME for speech quality prediction [23], [28], [30] used a single feature vector $\mathbf{e}$ of length B to represent the signal $\hat{s}(n)$, i.e,

$$\mathbf{e} = \big[\mathsf{e}(0)\ \mathsf{e}(1)\ \ldots\ \mathsf{e}(B-1)\big]^{\mathrm{T}}, \tag{10}$$

where

$$\mathsf{e}(b) = \frac{1}{JL}\sum_{j=0}^{J-1}\sum_{\ell=0}^{L-1}\check{e}_j(b,\ell). \tag{11}$$

The SRMR [30] and the normalized SRMR (SRMR$_{norm}$) [23] differ in the extraction of the vector $\mathbf{e}$ but both compute the estimate $\hat{p}_{\hat{s}}$ as the ratio between the lower and higher coefficients of $\mathbf{e}$. The measure proposed by the author in [28] computes the estimate $\hat{p}_{\hat{s}}$ as the output of a model tree, i.e., a combination of classification rules and regression but uses the same features as in [23] and therefore does not take into account time dependencies in the input signal. We propose to compute the estimate $\hat{p}_{\hat{s}}$ using a time ordered sequence of vectors and a trained LSTM-Network.

B. LSTM Network as Predictive Function

Artificial neural networks (ANN) are composed of several layers. Each λ-th layer applies a non linear mapping to an input vector $\mathbf{x}^\lambda$, of length $L^\lambda_{\mathbf{x}}$, in order to compute an output vector $\mathbf{z}^\lambda$, of length $L^\lambda_{\mathbf{z}}$. This mapping is applied by multiplying a weight matrix $\mathbf{W}^\lambda_{\mathbf{x},\mathbf{z}}$ of size $L^\lambda_{\mathbf{z}} \times L^\lambda_{\mathbf{x}}$, where subscripts indicate the connections represented by the matrix, with the input vector $\mathbf{x}^\lambda$ before summing the results with a bias vector $\mathbf{b}^\lambda_{\mathbf{z}}$ of length $L^\lambda_{\mathbf{z}}$ and applying a non-linear activation function $\mathcal{F}(\cdot)$ to the result, i.e.,

$$\mathbf{z}^\lambda = \mathcal{F}\left(\mathbf{W}^\lambda_{\mathbf{x},\mathbf{z}}\mathbf{x}^\lambda + \mathbf{b}^\lambda_{\mathbf{z}}\right). \tag{12}$$

The values of $\mathbf{W}^\lambda_{\mathbf{x},\mathbf{z}}$ and $\mathbf{b}^\lambda_{\mathbf{z}}$ have to be learned during a training phase (cf. Section IV-B) and any number of layers can be used by setting $\mathbf{x}^\lambda = \mathbf{z}^{\lambda-1}$. Layers described by (12) and networks composed exclusively of such layers are commonly qualified as feed-forward.

The use of RNNs is a common extension of (12) to take time dependencies into account. Similarly as feed-forward artificial neural networks (ANNs), RNNs are composed of several layers. However, the input of the λ-th RNN layer is an ordered sequence $\mathbf{X}^\lambda$ of T^λ input vectors $\mathbf{x}^\lambda_t$, where $t \in [0, T^\lambda - 1]$ denotes the

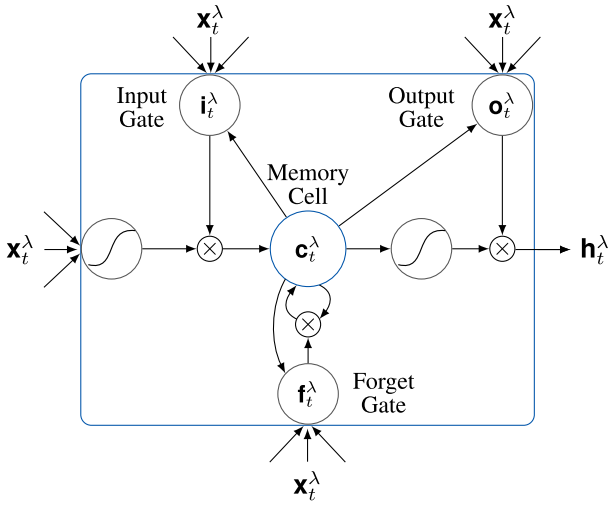

Fig. 2. Overview of the updates applied by an LSTM-layer. Updates are applied along the time-ordered sequence of input vectors resulting in a sequence of hidden vectors used to compute the output of the layer as in (15). The value of each hidden vector depends on the current input as well as on the memory cell and of the weights applied in the input, output and forget gates.

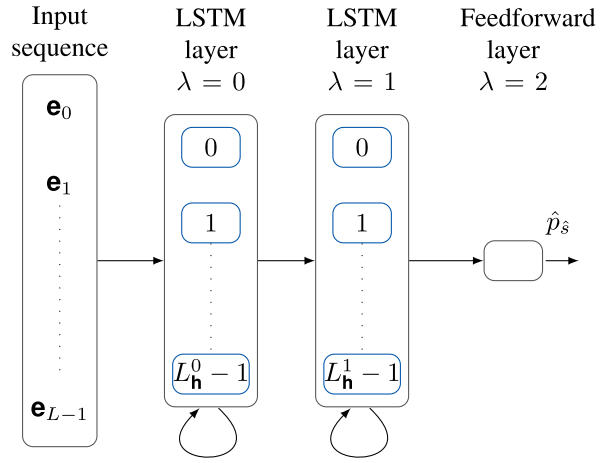

Fig. 3. Overview of the network used as predicting function. The first LSTM-layer computes a sequence of input vectors containing as many vectors as frames available in the target signal. The second LSTM-layer applies updates along this sequence and the last vector of its output sequence is input to the feed-forward layer whose sigmoid activation function results in a prediction bounded between 0 and 1.

sequence index, i.e.,

$$\mathbf{X}^\lambda = \left\{\mathbf{x}_0^\lambda, \mathbf{x}_1^\lambda, \ldots, \mathbf{x}_{T^\lambda-1}^\lambda\right\}. \tag{13}$$

Each layer of an RNN computes a sequence $\mathbf{H}^\lambda$ of hidden vectors $\mathbf{h}_t^\lambda$ of length $L_\mathbf{h}^\lambda$ and a sequence $\mathbf{Z}^\lambda$ of output vectors $\mathbf{z}_t^\lambda$ of length $L_\mathbf{z}^\lambda$, both containing T^λ vectors and defined similarly as in (13). The vectors in these sequences are computed by iteratively applying [31]

$$\mathbf{h}_t^\lambda = \mathcal{F}\left(\mathbf{W}_{\mathbf{x},\mathbf{h}}^\lambda \mathbf{x}_t^\lambda + \mathbf{W}_{\mathbf{h},\mathbf{h}}^\lambda \mathbf{h}_{t-1}^\lambda + \mathbf{b}_\mathbf{h}^\lambda\right), \tag{14}$$

$$\mathbf{z}_t^\lambda = \mathcal{F}\left(\mathbf{W}_{\mathbf{h},\mathbf{z}}^\lambda \mathbf{h}_t^\lambda + \mathbf{b}_\mathbf{z}^\lambda\right), \tag{15}$$

where $\mathbf{W}_{\mathbf{x},\mathbf{h}}^\lambda$, $\mathbf{W}_{\mathbf{h},\mathbf{h}}^\lambda$ and $\mathbf{W}_{\mathbf{h},\mathbf{z}}^\lambda$ denote weight matrices of size $L_\mathbf{h}^\lambda \times L_\mathbf{x}^\lambda$, $L_\mathbf{h}^\lambda \times L_\mathbf{h}^\lambda$ and $L_\mathbf{z}^\lambda \times L_\mathbf{h}^\lambda$, respectively, and where $\mathbf{b}_\mathbf{h}^\lambda$ and $\mathbf{b}_\mathbf{z}^\lambda$ are bias vectors of length $L_\mathbf{h}^\lambda$ and $L_\mathbf{z}^\lambda$, respectively.

For our application, i.e., the prediction of speech quality, RNNs have two main advantages over feed-forward networks. First, the sequence of hidden vectors computed by an RNN allows the prediction to take into account temporal dependencies; second, the iterative updates can be applied to a sequence of arbitrary length. However, the values of the weight matrices and of the bias vectors still have to be learned during a training phase and the formulation in (14) and (15) can cause instability during training, leading to overly long training time or even divergence [32]. In order to avoid these issues, so-called gated units, such as in the LSTM layers used in this paper, are used in practice.

Though in a standard RNN layer, the function $\mathcal{F}(\cdot)$ in (14) is commonly a simple non-linear function such as a sigmoid, in an LSTM layer, this function relies on iterative updates of sequences of vectors, $\mathbf{I}^\lambda$, $\mathbf{O}^\lambda$, $\mathbf{F}^\lambda$ and $\mathbf{C}^\lambda$, the so-called, input gate, output gate, forget gate and cell memory, respectively and their mutual influence on the layer's output is illustrated in Fig. 2. For each step t of the input sequence $\mathbf{X}^\lambda$, the vectors $\mathbf{i}_t^\lambda$ and $\mathbf{f}_t^\lambda$ are computed from the input vector $\mathbf{x}_t^\lambda$ and from the memory cell vector $\mathbf{c}_{t-1}^\lambda$ saved at the previous step, i.e.,

$$\mathbf{i}_t^\lambda = \mathcal{S}\left(\mathbf{W}_{\mathbf{x},\mathbf{i}}^\lambda \mathbf{x}_t^\lambda + \mathbf{W}_{\mathbf{h},\mathbf{i}}^\lambda \mathbf{h}_{t-1}^\lambda + \mathbf{W}_{\mathbf{c},\mathbf{i}}^\lambda \mathbf{c}_{t-1}^\lambda + \mathbf{b}_\mathbf{i}^\lambda\right), \tag{16}$$

$$\mathbf{f}_t^\lambda = \mathcal{S}\left(\mathbf{W}_{\mathbf{x},\mathbf{f}}^\lambda \mathbf{x}_t^\lambda + \mathbf{W}_{\mathbf{h},\mathbf{f}}^\lambda \mathbf{h}_{t-1}^\lambda + \mathbf{W}_{\mathbf{c},\mathbf{f}}^\lambda \mathbf{c}_{t-1}^\lambda + \mathbf{b}_\mathbf{f}^\lambda\right), \tag{17}$$

where $\mathcal{S}(\cdot)$ denotes the logistic sigmoid function. The resulting vectors $\mathbf{i}_t^\lambda$ and $\mathbf{f}_t^\lambda$ weight the influence of the current and previous input, respectively, to the updated vector $\mathbf{c}_t^\lambda$ computed as

$$\mathbf{c}_t^\lambda = \mathbf{f}_t^\lambda \mathbf{c}_{t-1}^\lambda + \mathbf{i}_t^\lambda \tanh\left(\mathbf{W}_{\mathbf{x},\mathbf{c}}^\lambda \mathbf{x}_t^\lambda + \mathbf{W}_{\mathbf{h},\mathbf{c}}^\lambda \mathbf{h}_{t-1}^\lambda + \mathbf{b}_\mathbf{c}^\lambda\right). \tag{18}$$

The influence of this memory cell vector $\mathbf{c}_t^\lambda$ to the layer output is weighted by the ouput gate $\mathbf{o}_t^\lambda$ computed as

$$\mathbf{o}_t^\lambda = \mathcal{S}\left(\mathbf{W}_{\mathbf{x},\mathbf{o}}^\lambda \mathbf{x}_t^\lambda + \mathbf{W}_{\mathbf{h},\mathbf{o}}^\lambda \mathbf{h}_{t-1}^\lambda + \mathbf{W}_{\mathbf{c},\mathbf{o}}^\lambda \mathbf{c}_t^\lambda + \mathbf{b}_\mathbf{o}^\lambda\right), \tag{19}$$

and used to compute the hidden vector $\mathbf{h}_t^\lambda$,

$$\mathbf{h}_t^\lambda = \mathbf{o}_t^\lambda \tanh\left(\mathbf{c}_t^\lambda\right), \tag{20}$$

from which the ouput vector $\mathbf{z}_t^\lambda$ is finally computed as per (15).

We propose to use the stacking of $\Lambda = 3$ layers as the predicting function of the quality of processed speech. Using this network structure, similar to the one used in [33] and depicted in Fig. 3, the speech quality from a signal $\hat{s}(n)$ is predicted by using the sequence of L time ordered frames of features as input to the first LSTM layer, i.e., for $\lambda = 0$ we have $T^0 = L$, and

$$\mathbf{X}^0 = \left\{\mathbf{x}_0^0, \mathbf{x}_1^0, \ldots, \mathbf{x}_{T^0-1}^0\right\}, \text{ with,} \tag{21}$$

$$\mathbf{x}_t^0 = \mathbf{e}_t, t \in [0, T^0 - 1]. \tag{22}$$

The output sequence, obtained after iterating (16)–(20) and (15) along the input sequence, is used as input to a second LSTM layer, i.e., $\mathbf{X}^1 = \mathbf{Z}^0$ and the iterative updates yield the output sequence $\mathbf{Z}^1$. The last vector of this sequence is input to the last, feed-forward, layer, i.e., $\mathbf{x}^2 = \mathbf{z}_{T^1-1}^1$. Aiming at an estimate $\hat{p}_{\hat{s}}$ that is bounded between 0 and 1, we replace $\mathcal{F}(\cdot)$ in (12) by a

TABLE I
EXPECTED BEHAVIOR OF THE CONSIDERED ALGORITHMS

	UN	SC	MVDR	GWPE-MVDR
Denoising	Poor	Good	Fair	Fair
Dereverberation	Poor	Poor	Poor	Good
Speech distortions	Good	Fair	Good	Fair
Noise distortions	Good	Fair	Good	Poor

TABLE II
RIRS USED TO GENERATE THE RECORDINGS ALONG WITH THEIR RESPECTIVE CHARACTERISTICS

Labels	Room	T60 [s]	DRR [dB]
RIR 1	Office 1	0.35	10.45
RIR 2	Building Lobby	0.77	5.13
RIR 3	Lecture Room 2	1.26	3.79

sigmoid and compute the estimate of $p_{\hat{s}}$ as

$$[\hat{p}_{\hat{s}}] = \mathcal{S}\left(\mathbf{W}^2_{\mathbf{x},\mathbf{z}}\mathbf{x}^2 + \mathbf{b}^{\lambda}_{\mathbf{z}}\right). \tag{23}$$

The values of the multiple weight matrices and bias vectors needed for the computation of (23) can be learned during a training phase using perceptually labeled training data. Section III presents the dataset that was collected for this purpose and the training procedure is described in Section IV.

III. DATASET

In order to train and evaluate the proposed measure described in Section II, we collected a database of noisy and reverberant speech signals processed by several categories of algorithms and labeled in terms of perceived speech quality. This section provides short descriptions of the considered algorithms before describing the perceptual evaluation conducted in order to label signals in terms of perceived speech quality.

A. Considered Algorithms

The algorithms considered in this paper process the recorded signal in the short-time Fourier transform (STFT) domain, in which the signal model from (1)–(2) can be expressed as

$$y_m(k,\ell) = x_m(k,\ell) + v_m(k,\ell), \tag{24}$$

$$y_m(k,\ell) = X^d_{m,k}\ell + X^r_{m,k}\ell + v_m(k,\ell), \tag{25}$$

where $y_m(k,\ell)$, $x_m(k,\ell)$, $v_m(k,\ell)$, $X^d_{m,k}\ell$ and $X^r_{m,k}\ell$ denote the STFTs of $y_m(n)$, $x_m(n)$, $v_m(n)$, $d_m(n)$, and $r_m(n)$, respectively.

These algorithms aim at computing the STFT $\hat{s}(k,\ell)$ of $\hat{s}(n)$ from $y_m(k,\ell)$. In addition to the unprocessed (UN) version of the signal, we considered three categories of algorithms, namely, single-channel spectral suppression (SC) [34], the minimum variance distortionless response (MVDR) beamformer and the application of this beamformer to the output of the generalized weighted prediction error (GWPE) [35], denoted by GWPE-MVDR. These algorithms have been chosen for their applicability to realistic scenarios, e.g., real-time applications, as well as to provide a wide range of processing artefacts typically occurring in different reverberation and noise conditions. The expected behavior of the algorithms in terms of interference reduction and introduced distortions is summarized in Table I and the categories of algorithms are briefly described in the next subsections.

All signals have a sampling frequency of f_s =16 kHz. Noisy and reverberant signals have been generated by convolving clean speech extracted from the WSJCAM0 database [36] with RIRs and adding noise to the resulting reverberant speech. We used 3 different RIRs extracted from the ACE challenge dataset [37] recorded using a 42 cm linear array of $M = 8$ equidistant microphones, whose characteristics, summarized in Table II, have been selected to represent a wide range of reverberation levels. We considered two noise types, namely *fan* noise and *babble*, for which noise signals recorded in the same rooms and with the same microphones positions as for the RIRs are available. We consider two SNRs, namely of 5 dB and 15 dB, calculated according to [38]. The 12 resulting combinations of RIRs with noise types and SNRs will be referred to as *acoustic conditions* in the remainder of this paper. When referring to UN as an algorithm, we consider $\hat{s}(n) = y_{\mathrm{ref}}(n)$ with ref arbitrarily set to 1.

1) Single-Channel Spectral Suppression (SC): SC algorithms estimate $\hat{s}(k,\ell)$ by applying a real valued gain to the STFT of one of the input channel, i.e.,

$$\hat{s}(k,\ell) = g(k,\ell)y_{\mathrm{ref}}(k,\ell). \tag{26}$$

The gain $g(k,\ell)$ is computed as

$$g(k,\ell) = \max\left(\tilde{g}(k,\ell), g_{\min}\right), \tag{27}$$

where $g_{\min}$ is a spectral floor introduced to limit possible speech distortion and where $\tilde{g}(k,\ell)$ is in this paper computed as the solution to the minimum mean square error (MMSE) estimator of the speech amplitude proposed in [39], i.e.,

$$\tilde{g}(k,\ell) = \sqrt{\frac{\xi(k,\ell)}{\mu+\xi(k,\ell)}} \cdot \left[\frac{\mathrm{Gam}\left(\mu+\frac{\beta}{2}\right)}{\mathrm{Gam}\left(\mu\right)} \frac{\Phi\left(1-\mu-\frac{\beta}{2},1;-\nu(k,\ell)\right)}{\Phi\left(1-\mu,1;-\nu(k,\ell)\right)}\right]^{1/\beta} \cdot \left(\sqrt{\gamma(k,\ell)}\right)^{-1}, \tag{28}$$

where

$$\nu(k,\ell) = \frac{\gamma(k,\ell)\xi(k,\ell)}{\mu+\xi(k,\ell)}, \tag{29}$$

and where $\Phi(\cdot)$ and $\mathrm{Gam}\,(\cdot)$ denote the confluent hypergeometric function and the complete Gamma function [40], respectively, and where β and μ are parameters of the assumed speech amplitude distribution [34]. Additionally, $\xi(k,\ell)$ denotes the *a priori* signal-to-interference ratio (SIR) defined as

$$\xi(k,\ell) = \frac{\sigma^2_{d,\mathrm{ref}}(k,\ell)}{\sigma^2_{r,\mathrm{ref}}(k,\ell)+\sigma^2_{v,\mathrm{ref}}(k,\ell)}, \tag{30}$$

and $\gamma(k,\ell)$ denotes the *a posteriori* SIR, defined as

$$\gamma(k,\ell) = \frac{|y_{\text{ref}}(k,\ell)|^2}{\sigma^2_{r,\text{ref}}(k,\ell) + \sigma^2_{v,\text{ref}}(k,\ell)}. \tag{31}$$

In (30) and (31), $\sigma^2_{d,\text{ref}}(k,\ell)$, $\sigma^2_{r,\text{ref}}(k,\ell)$ and $\sigma^2_{v,\text{ref}}(k,\ell)$ denote power spectral densities (PSDs), i.e.,

$$\sigma^2_{d,\text{ref}}(k,\ell) = \text{E}\left\{|d_{\text{ref}}(k,\ell)|^2\right\}, \tag{32}$$

where $\text{E}\{\cdot\}$ denotes the expectation operator, and with $\sigma^2_{r,\text{ref}}(k,\ell)$ and $\sigma^2_{v,\text{ref}}(k,\ell)$ defined similarly. In practice, these power spectral densitys (PSDs) are unknown and their estimates, $\hat{\sigma}^2_{d,\text{ref}}(k,\ell)$, $\hat{\sigma}^2_{r,\text{ref}}(k,\ell)$ and $\hat{\sigma}^2_{v,\text{ref}}(k,\ell)$ have to be used instead.

The choice of the PSD estimators can greatly influence the performance of SC algorithms. In this paper, we denote by 'SCa' the combination described in [34], which has been shown effective in improving speech qualtiy in reverberant scenarios [41], [42]. The estimates of the PSDs, $\sigma^2_{v,\text{ref}}(k,\ell)$, $\sigma^2_{r,\text{ref}}(k,\ell)$ and $\sigma^2_{d,\text{ref}}(k,\ell)$, are estimated using a modified version of the well known minimum statistics (MS) estimator [43], the Lebart approach [44] and cepstral smoothing [45], respectively, $\tilde{g}(k,\ell)$ is computed using (28) and g_{min} is set to a minimum gain of -10 dB. A detailed description of the approach is available in [34].

Aiming at measuring the effect of distortions that a poorly tuned SC algorithm could introduce, we used a modified version of this scheme denoted by 'SCb'. In this case, we estimate $\sigma^2_{v,\text{ref}}(k,\ell)$ and $\sigma^2_{d,\text{ref}}(k,\ell)$ using the estimators proposed in [46] and in [47], respectively, and set g_{min} to a minimum gain of -30 dB.

2) MVDR Beamformer: The MVDR beamformer considered in this paper estimates $\hat{s}(k,\ell)$ by filtering and summing the STFT coefficients of the multichannel input, i.e., with superscript H denoting hermitian conjugation,

$$\hat{s}(k,\ell) = \mathbf{w}^{\text{H}}_{\hat{\theta}}(k)\mathbf{y}(k,\ell) \tag{33}$$

where $\mathbf{w}_{\hat{\theta}}(k)$ denotes the stacked filter coefficient vector of the beamformer steered towards the estimate $\hat{\theta}$ of the direction of arrival (DOA), θ, of the target speech and where $\mathbf{y}(k,\ell)$ denotes the M-dimensional stacked vector of the received microphone signals

$$\mathbf{y}(k,\ell) = [y_1(k,\ell)\ y_2(k,\ell)\ \ldots\ y_M(k,\ell)]^{\text{T}}. \tag{34}$$

Aiming at minimizing the noise power while providing a unity gain in the direction of the target speech, the filter coefficients of the MVDR beamformer are computed as [48]

$$\mathbf{w}_{\hat{\theta}}(k) = \frac{\hat{\mathbf{\Gamma}}^{-1}(k)\mathbf{d}_{\hat{\theta}}(k)}{\mathbf{d}^H_{\hat{\theta}}(k)\hat{\mathbf{\Gamma}}^{-1}(k)\mathbf{d}_{\hat{\theta}}(k)}, \tag{35}$$

where $\mathbf{d}_{\theta}(k)$ and $\hat{\mathbf{\Gamma}}(k)$ denote the steering vector of the target speaker and the noise coherence matrix, respectively.

In this paper, the estimate $\hat{\mathbf{\Gamma}}(k)$ is computed as

$$\hat{\mathbf{\Gamma}}(k) = \overline{\mathbf{\Gamma}}(k) + \varrho(k)\mathsf{I}_M, \tag{36}$$

where $\overline{\mathbf{\Gamma}}(k)$ denotes the coherence matrix of a diffuse noise field [48], I_M denotes the $M \times M$-dimensional identity matrix and $\varrho(k)$ denotes a frequency-dependent regularization parameter used to limit potential amplification of uncorrelated noise, especially at low frequencies. This regularization parameter is computed iteratively such that

$$\mathbf{w}^{\text{H}}_{\hat{\theta}}(k)\mathbf{w}_{\hat{\theta}}(k) \leq \text{WNG}_{\max}, \tag{37}$$

where $\text{WNG}_{\max}$ denotes the so-called white noise gain constraint [49]. In this paper we set this constraint to -10 dB, compute the steering vector $\mathbf{d}_{\hat{\theta}}(k)$ using a far-field assumption and measure the true θ from the main peaks of the used RIRs. In order to evaluate the impact of steering error on the performance of the beamformer we consider perfectly steered, i.e., $\hat{\theta} = \theta$, denoted by 'MVDRa', and missteered beamformer, i.e., $\hat{\theta} = \theta + \epsilon_{\hat{\theta}}$ with $\epsilon_{\hat{\theta}} = \pi/4$ denoted by 'MVDRb'.

3) GWPE-MVDR: The combination of GWPE and MVDR beamformer (GWPE-MVDR) considered in this paper estimates $\hat{s}(k,\ell)$ as

$$\hat{s}(k,\ell) = \mathbf{w}^{\text{H}}_{\hat{\theta}}(k)\left(\mathbf{y}(k,\ell) - \hat{\mathbf{r}}(k,\ell)\right), \tag{38}$$

where $\mathbf{w}_{\hat{\theta}}(k)$ is computed as in (35) and

$$\hat{\mathbf{r}}(k,\ell) = [\hat{r}_1(k,\ell)\ \hat{r}_2(k,\ell)\ \ldots\ \hat{r}_M(k,\ell)]^{\text{T}}, \tag{39}$$

where $\hat{r}_m(k,\ell)$ denotes an estimate of $X^r_{m,k}\ell$, i.e., $\hat{s}(k,\ell)$ is estimated by subtracting a complex valued estimate of the late reverberation from the multichannel input signal before applying an MVDR beamformer.

In this paper, the estimate $\hat{\mathbf{r}}(k,\ell)$ is computed using the approach described in [35], i.e., as

$$\hat{\mathbf{r}}(k,\ell) = \mathbf{P}^{\text{H}}(k,\ell)\tilde{\mathbf{y}}(k,\ell-\Delta), \tag{40}$$

where Δ denotes a delay introduced to preserve the early reflections, and

$$\mathbf{P}(k,\ell) = [\mathbf{p}_1(k,\ell)\ \cdots \mathbf{p}_M(k,\ell)] \in \mathbb{C}^{ML_P \times M}, \tag{41}$$

where $\mathbf{p}_m(k,\ell) \in \mathbb{C}^{ML_P}$ denotes a multichannel prediction filter, and

$$\begin{aligned}\tilde{\mathbf{y}}(k,\ell) = [&y_1(k,\ell)\ \cdots y_1(k,\ell-L_P+1)\ \cdots \\ &y_M(k,\ell)\ \cdots y_M(k,\ell-L_P+1)]^{\text{T}},\end{aligned} \tag{42}$$

denotes a vector of STFT coefficients of length $M \cdot L_P$.

For each time-frequency bin, the matrix $\mathbf{P}(k,\ell)$ is computed by applying γ iterative updates aiming at solving the optimization problem [35]

$$\begin{aligned}&\underset{\mathbf{P}(k,\ell)}{\text{argmin}}\ \text{tr}\left\{\mathbf{P}^{\text{H}}(k,\ell)\hat{\mathbf{A}}(k,\ell)\mathbf{P}(k,\ell)\right\} \\ &-2\Re\left\{\text{tr}\left\{\mathbf{P}^{\text{H}}(k,\ell)\hat{\mathbf{B}}(k,\ell)\right\}\right\} \\ &\text{subject to } |\mathbf{P}^{\text{H}}(k,\ell)\tilde{\mathbf{y}}(k,\ell-\Delta)|^2 \leq \hat{\boldsymbol{\sigma}}^2_r(k,\ell),\end{aligned} \tag{43}$$

where

$$\hat{\boldsymbol{\sigma}}^2_r(k,\ell) = [\hat{\sigma}^2_{r,1}(k,\ell)\ \cdots \hat{\sigma}^2_{r,M}(k,\ell)]^{\text{T}}, \tag{44}$$

and $\hat{\sigma}^2_{r,m}(k,\ell)$ is computed similarly as in Section III-A1 and with

$$\hat{\mathbf{A}}(k,\ell) = \sum_{i=1}^{\ell} \delta^{\ell-i} \hat{w}_P(k,i) \tilde{\mathbf{y}}(k,i-\Delta) \tilde{\mathbf{y}}^{\mathrm{H}}(k,i-\Delta), \quad (45)$$

$$\hat{\mathbf{B}}(k,\ell) = \sum_{i=1}^{\ell} \delta^{\ell-i} \hat{w}_P(k,i) \tilde{\mathbf{y}}(k,i-\Delta) \mathbf{y}^{\mathrm{H}}(k,i), \quad (46)$$

where $\delta \in [0,1]$ denotes a smoothing constant, and $\hat{w}_P(k,\ell)$ denotes the weight used to emphasize frames where the the signal to be preserved is expected to have low power, computed as

$$\hat{w}_P(k,\ell) = \left(\frac{1}{M} \| \hat{\boldsymbol{\sigma}}^2_d(k,\ell) \|_2^2 + \epsilon \right)^{-1}, \quad (47)$$

where ϵ denotes a small regularization constant and where

$$\hat{\boldsymbol{\sigma}}^2_d(k,\ell) = [\hat{\sigma}^2_{d,1}(k,\ell) \ \cdots \hat{\sigma}^2_{d,M}(k,\ell)]^{\mathrm{T}}, \quad (48)$$

where $\hat{\sigma}^2_{d,m}(k,\ell)$ is an estimate of $\sigma^2_{d,m}(k,\ell)$ computed from $\hat{\sigma}^2_{r,m}(k,\ell)$ and $\hat{\sigma}^2_{y,m}(k,\ell)$ using recursive temporal smoothing.

It can be noted that the optimization problem in (43) does not take noise into account as the approach presented in [35] has been designed aiming at dereverberation in noise-free scenarios. The filtered noise signal resulting from (38) might have different spatial properties than the noise signal recorded by the microphones and might result in lower noise reduction achieved by GWPE-MVDR compared to MVDR alone. In practice, GWPE-MVDR could be combined with spectral suppression to overcome such drawbacks. Such combination has not been considered in this paper in order to obtain processed signals containing a wide range of processing artefacts. We used prediction filters of length $L_P = 5$ and a smoothing constant $\delta = 0.95$. Other parameters have been set as in [35]. Similarly as for MVDR, we consider both perfect steering and steering error and refer to the corresponding settings as 'GWPE-MVDRa' and 'GWPE-MVDRb', respectively.

All STFTs have been computed using a Hamming window. In the case of SC and MVDR, we used a window of 32 ms and an overlap of 16 ms while in the case of GWPE-MVDR, we used a window of 64 ms and an overlap of 48 ms, in order to replicate the implementation from [34] and [35].

B. Perceptual Evaluation

In order to obtain a dataset of processed signals labeled in terms of *overall quality*, we conducted a MUSHRA test [7] involving 20 self-reported normal-hearing assessors. All assessors evaluated all combinations of the algorithm categories SC, MVDR and GWPE-MVDR, with two settings being considered for each, e.g., SCa and SCb, with the 12 acoustic conditions described in Section III-A. For each assessor, this total of 72 combinations was divided into two equally sized groups assigned to two sessions of listening tests. For each session, the UN algorithm was added to the group of combinations to be evaluated resulting in a total of 48 combinations per session and per assessor. The 48 combinations were randomly divided into six partitions and three clean male speech and three clean female speech utterances were randomly extracted from the WSJCAM0 database [36], with a sanity check insuring that no utterance would have been previously assigned to another session or assessor. One clean speech utterance was randomly assigned to each partition and used to generate signals, for the corresponding combinations as described in Section III-A. Each partition was appended with the clean signal, used as reference signal, and two anchors, differing in the type of considered noise signal, either *babble* or *fan* noise. These anchors were generated by convolving the clean signal with the first channel of RIR 3 (cf. Table I), and adding noise with a SNR of 5 dB measured according to [38] and band-pass filtering the resulting noisy and reverberant signal according to [50]. Once all signals were generated, the test procedure for each assessor was conducted as follows.

TABLE III
OVERVIEW OF THE DATABASE OF PERCEPTUALLY EVALUATED DATA, EXCLUDING REFERENCES AND ANCHORS. NUMBERS DENOTE THE DURATION IN MINUTES, FOR EACH COMBINATION OF ACOUSTIC CONDITION AND ALGORITHM, FOR A TOTAL OF 5.23 HOURS AND 1920 PERCEPTUALLY EVALUATED SIGNALS

			UN	SC		MVDR		GWPE-MVDR	
	Noise	SNR		a	b	a	b	a	b
RIR 1	Fan	5 dB	6.27	3.55	3.03	3.41	3.31	3.35	3.45
		15 dB	6.32	3.04	3.14	3.41	3.49	3.14	3.20
	Babble	5 dB	6.52	3.38	3.13	3.27	3.30	3.14	3.67
		15 dB	6.36	3.50	2.96	3.00	3.44	3.04	3.20
RIR 2	Fan	5 dB	6.53	3.22	2.92	3.06	3.40	2.87	3.17
		15 dB	6.63	2.94	3.19	3.05	3.45	3.21	3.52
	Babble	5 dB	6.52	3.43	3.37	3.30	3.51	3.35	3.26
		15 dB	6.62	3.58	3.27	3.30	3.30	3.25	3.32
RIR 3	Fan	5 dB	6.88	3.40	3.26	3.34	3.33	3.55	3.31
		15 dB	6.64	3.29	3.27	3.09	3.44	3.06	3.12
	Babble	5 dB	6.44	3.41	3.24	3.18	3.04	3.12	3.31
		15 dB	6.48	3.28	3.46	3.20	3.32	3.50	3.51

The signals under test were normalized to have their maximum level, calculated over segments of 500 ms, equal to 65 dB SPL. Stimuli were diotically presented over headphones (Senheiser HD200) in a soundproof booth. The speech material presented to the assessor was first presented in a training phase during which the assessor could listen to all stimuli in order to become familiar with the material. Subsequently, the signals corresponding to each partition of test conditions were presented simultaneously on a screen. For each signal, there was a corresponding slider that the assessor was prompted to use in order to grade the overall quality of the material on an integer-valued scale between 0 (poor) and 100 (excellent). The scores assigned to the reference and anchor conditions were used to ensure that the assessors conducted the taskreliably, i.e., that they assigned the highest score to the hidden reference and a low score to the anchor. The significance of differences between the 12 acoustic conditions was assessed using a repeated-measures analysis of variance (ANOVA) and post-hoc analysis, whose details are presented in Appendix A. An overview of the collected dataset is presented in Table III and the scores for all combinations are summarized in Fig. 4. In the remainder of this paper, we consider

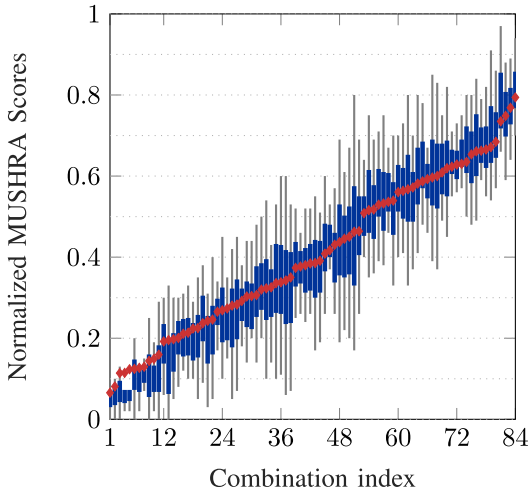

Fig. 4. MUSHRA scores after removing references and anchors, for all combinations of acoustic conditions, algorithms and settings. The mean, represented by a red dot, is considered the ground truth.

the true perceived speech quality of a signal to be the mean of the scores assigned by the assessors to all signals of the same acoustic condition, algorithm and algorithm setting and refer to it as the *ground truth*.

IV. Experiments

A. Benchmark and Figures of Merit

The results presented in Section V compare the performance of the proposed measure with several measures of the literature. Though aiming at non-intrusive prediction of the speech quality, our benchmark includes three intrusive measures, namely PESQ [15], POLQA [16], PEMO-Q [17] as they are commonly used to evaluate speech enhancement algorithms. It should however be emphasized that, as the computation of these measures requires the clean reference signal, they are not applicable in all scenarios and have an advantage in terms of performance compared to non-intrusive measures. Our benchmark includes four non-intrusive measures, namely P.563 [20], ANIQUE+ [21], SRMR$_{norm}$ [23] and the combination of modulation energies and model tree proposed by the author in [28] and denoted by 'Tree'. It should be noted that both Tree and the proposed measure rely on predicting functions trained using machine-learning techniques and that, contrary to the other considered measures, their performance depends on the data included in the training set.

We assess the performance of the proposed measure and of the benchmark measures using four figures of merit. For each measure, the linear relationship between the predicted quality and the ground truth is quantified using the Pearson correlation coefficient ρ, the ranking capability of each measure is quantified by the Spearman rank correlation coefficient ρ_{spear} and the correlation coefficient ρ_{sig} is computed similarly as ρ after applying a sigmoidal mapping, whose parameters are computed from the training set, to the predicted values. Finally, $\epsilon-\text{RMSE}$ is used to represent the error between the predicted value and the ground truth. This figure of merit is similar to the conventional RMSE but takes the uncertainty of the subjective ratings into account, i.e., $\epsilon-\text{RMSE}$ will be lower if the variance of the subjective ratings is high. An ideal measure should yield correlation values close to one and an $\epsilon-\text{RMSE}$ close to zero. Details on the computation of ρ_{sig} and $\epsilon-\text{RMSE}$ can be found in [51].

B. Training Framework

The training of the predicting functions used by the proposed approach and of Tree, as well as the linear mapping used in the computation of ρ_{sig} require a training set of signals for which the ground truth value of $p\hat{s}$ is known. Additionally, a testing set is needed to assess the performance of these trained measures and of the benchmark measures listed above. The network described in Section II was set with $L^0_{\mathbf{h}} = L^1_{\mathbf{h}} = 128$ and was trained using Keras [52] and the Adam algorithm [53]. During training, zero padding was applied to ensure that all sequences had same length and a masking layer was added before the first LSTM layer to ignore time frames containing only zeros. Each training epoch computed as many iterations as needed to take the entire training set into account using a batch size of 128 sequences. In our implementation Dropout [54] was applied both to the input of the network and to the output of each LSTM layer, i.e., 30% of the values input to the network and output by each LSTM layer were randomly selected and replaced by zeros at each iteration. At each iteration, the model, i.e., weight matrices and bias vectors, was updated to minimize the mean squared error (MSE) between the predicted and ground truth value of the speech quality assigned to each file of the training set. In order to avoid overfitting, 10% of the training set was set aside prior to each training phase to be used as a validation set. The training algorithm computed 500 epochs and testing was done using the model that yielded the lowest MSE over the validation set.

We conducted three experiments that differ in the training and testing sets constructed from the dataset presented in Section III. In all experiments, anchors and reference signals were discarded before training and testing. The first experiment aims at assessing the ability of the proposed measure to predict the speech quality from signals processed using a single category of algorithms, e.g., SC, but different settings, e.g., SCa and SCb. For this purpose, the dataset was divided into 4 subsets containing only files processed with the same category of algorithm (UN, SC, MVDR and GWPE-MVDR). For each subset, we used 5-folds cross validation, proceeding as follows. The 20 assessors have been randomly divided into 5 equally-sized disjoint groups. For each fold, the signals in one of these groups were considered as the test set, while the data corresponding to the remaining groups were used for training. Using this folding, assessors, speech stimuli and noise segments always differ between training and testing. The second experiment aims at assessing the ability of the proposed measure to predict the speech quality from signals processed with a variety of algorithms categories. For this purpose, we use the same 5-folds validation procedure but apply it to the entire collected dataset. We used the same training sets and folds for the proposed approach, Tree, and learning of the parameters of the sigmoid used in the computation of ρ_{sig}.

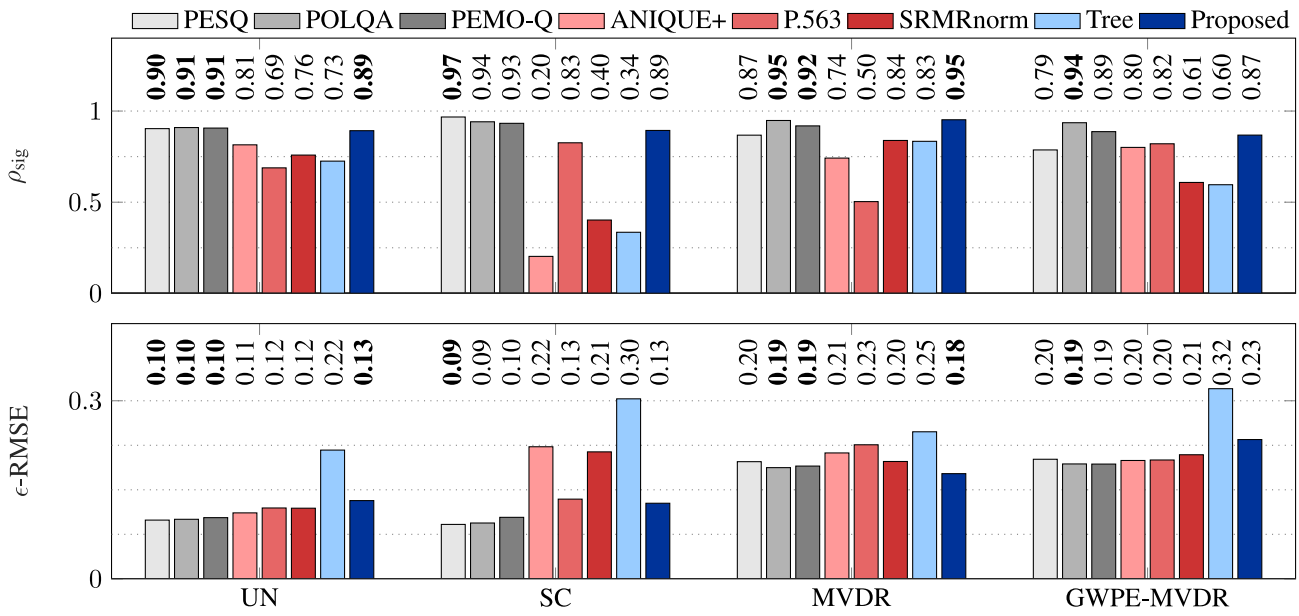

Fig. 5. Performance of the considered measures in terms of ρ_{sig} (top) and $\epsilon - \mathrm{RMSE}$ (bottom). The labels along the x-axis denote the category of algorithms included in training and testing sets. Numbers in bold typeface denote the best attained performance (statistically indifferent) per considered category of algorithm.

The third experiment examines the behavior of the proposed approach in case of mismatch between the algorithms included in the training and the testing set. For this purpose, all signals processed with a single category of algorithms are included in the testing set while all others are included in the training set. It should be noted that using such partition yields a larger training set than for the previous experiments.

The figures of merit reported for the first and second experiments in Section V are averaged over all folds. In the case of the correlation measures, a Fisher Z-test, at a significance level of 0.05, has been conducted before averaging to ensure that the values did not differ significantly between folds [55]. A similar Fisher Z-test has been used to determine if the difference between the correlations measures yielded by the considered measures were significant. In the case of $\epsilon-\mathrm{RMSE}$, significance was determined using the F-measure criteria suggested by ITU-T in [56] and detailed, e.g., in [51].

V. Results

This section reports the results obtained considering the different training and testing sets previously described. As the three measures of correlation showed similar behavior at all of the considered measures we only report ρ_{sig} and $\epsilon - \mathrm{RMSE}$. The performance obtained when training and testing the proposed measure for a single category of algorithms at a time are depicted in Fig. 5 along with the performance of the considered benchmark measures on the same testing sets. With the exception of the proposed measure, non-intrusive measures are consistently outperformed by intrusive measures, as could be expected. The proposed measure, however, yields similar performance as the intrusive measures for all considered categories of algorithms and, when training and testing on either unprocessed signals (UN) or signals processed using the MVDR beamformer, there is no significant difference between the proposed measure and the intrusive measures (indicated by bold typeface in Fig. 5). Although for SC and GWPE-MVDR the proposed measure yields a slightly poorer performance than the intrusive measures, it outperforms all non-intrusive measures in terms of both ρ_{sig} and $\epsilon - \mathrm{RMSE}$, except for GWPE-MVDR where the proposed measure yields a slightly higher $\epsilon - \mathrm{RMSE}$ than the benchmark measures. The non-intrusive benchmark measures yield similar performance for unprocessed signals but perform inconsistently across the other categories of algorithms. Notably, ANIQUE+, SRMR_{norm} and Tree yield low ρ_{sig} (0.2 to 0.4) and high $\epsilon - \mathrm{RMSE}$ (0.2 to 0.3) in the case of SC. As both SRMR_{norm} and Tree use ME features which are, contrary to the case of the proposed measure, averaged over time, this suggests that taking into account the time-dependency is beneficial. It can be noted than the difference in performance along different categories of algorithms is coherent with the results from previous works such as [25], in which it appeared that existing quality measures are often reliable when considering only one category of algorithms.

The performance obtained when training and testing the proposed measure for all categories of algorithms are depicted in Fig. 6 along with the performance of the considered benchmark measures on the same testing sets. Unsurprisingly, for all measures, correlations are lower than when considering a single category at a time and, still with the exception of the proposed measure, non-intrusive measures are consistently outperformed by intrusive measures. The measure Tree performs poorly suggesting that, though it had been shown to yield high correlations in [28], it is not suitable to predict the speech quality in various acoustic conditions or with algorithms that might produce high

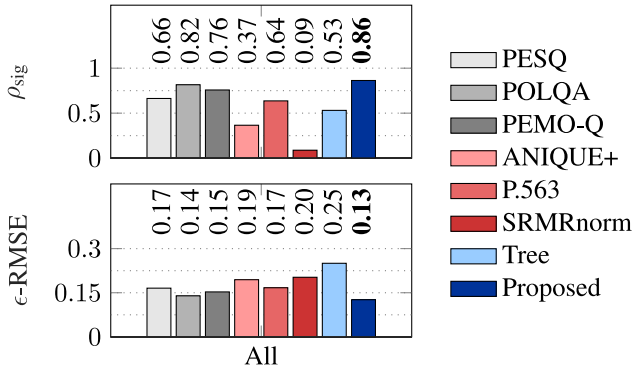

Fig. 6. Performance of the considered measures in terms of ρ_{sig} (top) and $\epsilon - \text{RMSE}$ (bottom). All considered categories of algorithms were included in both training and testing sets. Numbers in bold typeface denote the best attained performance (statistically indifferent).

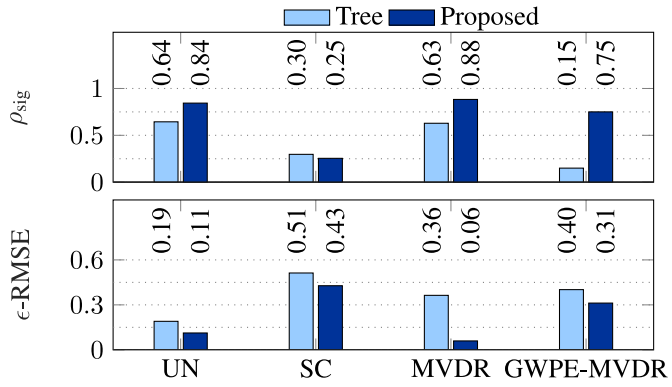

Fig. 8. Performance of the considered measures in terms of ρ_{sig} (top) and $\epsilon - \text{RMSE}$ (bottom). The labels along the x-axis denote the category of algorithms included in testing sets while other categories were included in training.

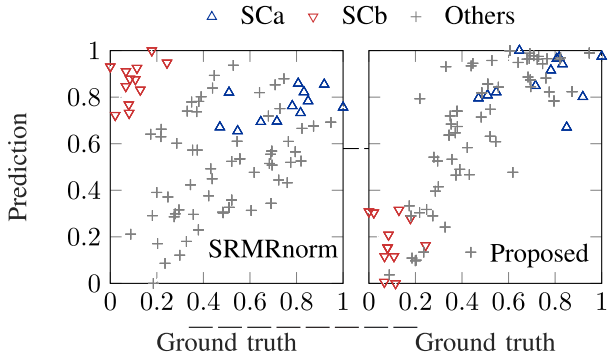

Fig. 7. Scatter plot of the predicted speech quality over ground truth data for SRMR_{norm} and the proposed measure when trained and tested for all considered algorithm categories, for all signals. Values corresponding to signals processed using SCa and SCb are highlighted, for readability, scores are normalized to range between 0 and 1.

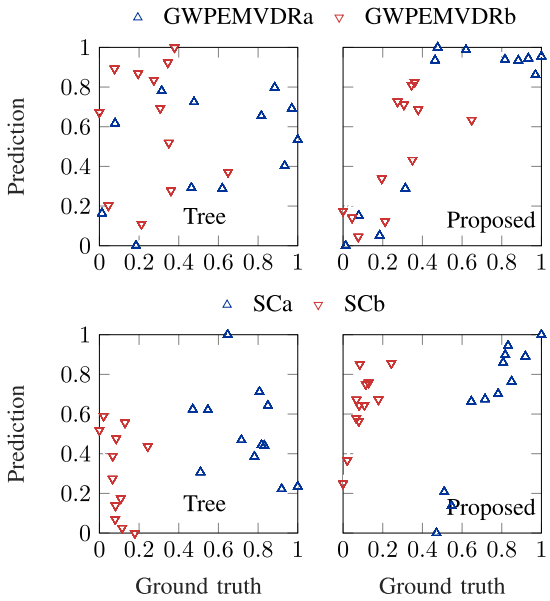

Fig. 9. Scatter plot of the predicted speech quality over ground truth data for Tree (left) and the proposed measure (right) when using a testing set of signals processed using either generalized weighted prediction error and MVDR beamformer (GWPE-MVDR) (top) or SC (bottom). Scores are normalized to range between 0 and 1.

levels of distortions. Similarly as in the case of SC showed in Fig. 5, SRMR_{norm} performs poorly with $\rho_{\text{sig}} = 0.09$, while the proposed measure outperforms all measures, including the intrusive ones, in terms of both ρ_{sig} and $\epsilon - \text{RMSE}$. This behavior is better illustrated in Fig. 7. In terms of ground truth, a clear divide appears between SCa and SCb, illustrating the clear preference of assessors for the well tuned single-channel scheme (SCa) over the purposefully poorly tuned (SCb). It appears as well that SRMR_{norm} largely overestimates the speech quality for signals processed with SCb while the proposed measure is able to adequately reflect the difference in performance between the two settings. This behavior is to be expected as, by averaging features over time, SRMR_{norm} effectively discards information about time-varying distortions (such as musical noise) while the proposed measure is designed to model such time-dependent effects.

As the RNN used as predicting function for the proposed measure is dependent on a training phase, one might want to consider the performance obtained in case of mismatch between the algorithms included in the training and the testing set. The performance of Tree and of the proposed measure in the presence of such mismatch is depicted in Fig. 8. It appears that when using a testing set composed of signals processed using either UN or MVDR, both Tree and the proposed measure yield similar performance as in the previous experiments in terms of ρ_{sig} and an even lower $\epsilon - \text{RMSE}$. Such behavior can be explained by the fact that MVDR, even misteered, does not introduce large amount of distortions and that the predicting functions were trained on a larger training set. However, mismatch between algorithms included in training and testing greatly deteriorates performance of both Tree and the proposed measure when the testing set is composed of signals processed using either GWPE-MVDR or SC.

In the case of GWPE-MVDR, for the proposed measure, correlation decreases only slightly in comparison with previous experiments, $\rho_{\text{sig}} = 0.75$. However, the variance is high,

$\epsilon - \mathrm{RMSE} = 0.31$. In the case of SC, both Tree and the proposed measure fail in their prediction, with low correlations $\rho_{\mathrm{sig}} = 0.30$ and $\rho_{\mathrm{sig}} = 0.25$ for Tree and the proposed measure, respectively. This difference in performance between the two measures and algorithms considered for testing is better illustrated in Fig. 9.

It appears that in the case of GWPE-MVDR, Tree does not yield an accurate prediction for any of the settings, i.e., GWPE-MVDRa and GWPE-MVDRb, while the proposed measure seems to overestimate the quality of signals processed with GWPE-MVDRa. It the case of SC, Tree fails similarly as in the GWPE-MVDR case but the poor performance of the proposed measure seems to come from an overestimation of the quality of signals processed using SCb. This behavior is unsurprising considering that the distortions introduced by spectral suppression, e.g., musical noise, differ greatly from the ones introduced by the other algorithms and that using this mismatch training set the predicting functions could not take them into account. Consequently, the proposed measure cannot be reliably used if the algorithm category under test is not included in the training set.

VI. Conclusion

Aiming at non-intrusively predicting the quality of processed signals, in this paper, we proposed a combination of modulation energies and of an RNN with LSTM cells that takes the time-dependency of the target signal into account. For this purpose, we collected a large dataset of perceptually evaluated signals representing a wide range of acoustic conditions and various categories of algorithms with different settings. We conducted several experiments, differing in terms of training and testing sets used to train and evaluate the proposed measure. The aim of these experiments was to evaluate the reliability of the proposed measure when trained and tested for either a single category of algorithms or several categories, and to investigate the performance of the measure in case of a mismatch between the algorithms included in the training and the testing sets.

Experimental results show that when trained and tested for a single category of algorithms, the proposed measure outperforms the considered non-intrusive benchmark measures and yields a similar performance as the intrusive benchmark measures. When trained and tested for several categories of algorithms, the proposed measure outperforms both intrusive and non-intrusive benchmark measures. However, as could be expected, the proposed measure can be unreliable in case of a mismatch in terms of algorithms between the training and testing sets. Consequently, the proposed measure might not be suitable to assess the performance of completely new categories of algorithms, but could be a useful approach for the real-time selection of algorithms or algorithm parameters.

Appendix A
Statistical Analysis of Perceptual Scores

A repeated-measures analysis of variance (ANOVA) and post-hoc analysis was applied to the perceptual scores presented in Subsection III-B in order to assess the significance of differences between the 12 acoustic conditions. This analysis indicated that all three acoustic factors, i.e., RIR, noise type and SNR, significantly affected the rating scores.

First, there was a significant effect of RIR ($F(2,38) = 11.4, p = 0.0013, \eta_p^2 = 0.376$) which was mainly due to significantly lower scores for RIR 2 (mean $M = 37.2$) compared to RIR 1 ($M = 42.8$) and RIR 3 ($M = 41.6$, paired $t(19) > 3.6, p < 0.002$) but no significant differences between RIR 1 and RIR 3 ($t(19) = 0.82, p = 0.42$). Second, there was a significant effect of noise type ($F(1,19) = 23.2, p < 0.001, \eta_p^2 = 0.55$) and fan noise ($M = 43.2$) was rated significantly higher compared to babble noise ($M = 37.9$). Third, there was a significant of SNR ($F(1,19) = 204.4, p < 0.001, \eta_p^2 = 0.91$) and the 5 dB condition ($M = 33.0$) was rated significantly lower compared to 15 dB condition ($M = 48.0$).

In addition, there was a significant interaction between RIR and SNR ($F(2,38) = 40.9, p < 0.001, \eta_p^2 = 0.68$). At 5 dB this was associated with significantly lower scores of RIR 1 ($M = 32.6$) and RIR 2 ($M = 28.6$) compared to RIR 3 ($M = 38.0$, paired $t(19) > 3.4, p < 0.018$, Bonferroni-corrected for multiple comparisons, n = 6) and only marginal differences between RIR 1 and RIR 2 ($t(19) = 2.8, p = 0.0624$). At 15 dB this was associated with significantly higher scores of RIR 1 ($M = 53.1$) compared to RIR 2 ($M = 45.8$) and RIR 3 ($M = 45.2$, paired $t(19) > 4.0, p < .001$) but no significant differences between RIR 2 and RIR 3 ($t(19) = 0.4, p = 0.72$). The above two-way interaction appears to be partially driven by the significant three-way interaction of RIR, noise type, and SNR ($F(2,38) = 5.0, p = 0.0119, \eta_p^2 = 0.21$). This interaction was due to insignificant differences of RIR 1 at 5 dB fan noise ($M = 33.2$) compared to babble noise ($M = 32.0$, paired $t(19) = 0.53, p = 0.59$) and insignificant differences of RIR 3 at 15 dB fan noise ($M = 47.6$) compared to babble noise ($M = 42.9$, $t(19) = 2.2, p = 0.23$) but at least marginally significant differences between the two noise types in all other combinations of conditions ($t(19) > 2.8, p < 0.065$, using Bonferroni-correction for multiple comparisons, n = 6).

References

[1] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer-Verlag, 2008.

[2] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*. Berlin, Germany: Springer, 2010.

[3] S. Doclo, W. Kellermann, S. Makino, and S. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 18–30, Mar. 2015.

[4] E. Vincent, T. Virtanen, and S. Gannot, *Audio Source Separation and Speech Enhancement*. Hoboken, NJ, USA: Wiley, 2018.

[5] K. Kinoshita *et al.*, "A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP J. Adv. Signal Process.*, vol. 2016, pp. 1–19, Jan. 2015.

[6] *Subjective Test Methodology for Evaluating Speech Communication Systems That Include Noise Suppression Algorithms*, ITU-T Standard P.835, Nov. 2003.

[7] *Method for the Subjective Assessment of Intermediate Quality Levels of Coding Systems*, ITU-T Standard BS.1534–3, Nov. 2003.

[8] P. Flipsen, "Measuring the intelligibility of conversational speech in children," *Clin. Linguistics Phonetics*, vol. 20, no. 4, pp. 303–312, 2006.

[9] C. S. J. Doire, M. Brookes, and P. A. Naylor, "Robust and efficient Bayesian adaptive psychometric function estimation," *J. Acoust. Soc. Amer.*, vol. 141, no. 4, pp. 2501–2512, 2017.

[10] 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.

[11] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Amer.*, vol. 67, no. 1, pp. 318–326, Jan. 1980.

[12] *Methods for the Calculation of the Speech Intelligibility Index*, ANSI Standard S3.5–1997 (R2007), 1997.

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

[14] J. Taghia and R. Martin, "Objective intelligibility measures based on mutual information for speech subjected to speech enhancement processing," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 1, pp. 6–16, Jan. 2014.

[15] *Perceptual Evaluation of Speech Quality (PESQ), An Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs*, ITU-T Recommendation P.862, Feb. 2001.

[16] *Perceptual Objective Listening Quality Assessment: An Advanced Objective Perceptual Method for End-to-End Listening Speech Quality Evaluation of Fixed, Mobile, and IP-Based Networks and Speech Codecs Covering Narrowband, Wideband, and Super-Wideband Signals*, ITU-T, Standard P.863, Jan. 2011.

[17] R. Huber and B. Kollmeier, "PEMO-Q—A new method for objective audio quality assessment using a model of auditory perception," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 1902–1911, Nov. 2006.

[18] A. H. Andersen, J. M. de Haan, Z. Tan, and J. Jensen, "A non-intrusive short-time objective intelligibility measure," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, New Orleans, LA, USA, Mar. 2017, pp. 5085–5089.

[19] C. Spille, S. D. Ewert, B. Kollmeier, and B. T. Meyer, "Predicting speech intelligibility with deep neural networks," *Comput. Speech Lang.*, vol. 48, pp. 51–66, 2018.

[20] L. Malfait, J. Berger, and M. Kastner, "P.563—The ITU-T standard for single-ended speech quality assessment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 1924–1934, Nov. 2006.

[21] D. S. Kim and A. Tarraf, "ANIQUE+: A new American national standard for non-intrusive estimation of narrowband speech quality," *Bell Labs Tech. J.*, vol. 12, pp. 221–236, 2007.

[22] T. H. Falk and W.-Y. Chan, "Temporal dynamics for blind measurement of room acoustical parameters," *IEEE Trans. Instrum. Meas.*, vol. 59, no. 4, pp. 978–989, Apr. 2010.

[23] J. F. Santos, M. Senoussaoui, and T. H. Falk, "An improved non-intrusive intelligibility metric for noisy and reverberant speech," in *Proc. Int. Workshop Acoust. Signal Enhancement*, Sep. 2014, pp. 55–59.

[24] J. Santos and T. Falk, "Updating the SRMR-CI metric for improved intelligibility prediction for cochlear implant users," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 2197–2206, Dec. 2014.

[25] S. Goetze *et al.*, "A study on speech quality and speech intelligibility measures for quality assessment of single-channel dereverberation algorithms," in *Proc. Int. Workshop Acoust. Signal Enhancement*, Sep. 2014, pp. 233–237.

[26] M. Karbasi, A. Abdelhaziz, H. Meutzner, and D. Kolossa, "Blind non-intrusive speech intelligibility prediction using twin-HMMs," in *Proc. INTERSPEECH*, San Francisco, CA, USA, Sep. 2016, pp. 625–629.

[27] D. Sharma, Y. Wang, P. Naylor, and M. Brookes, "A data-driven non-intrusive measure of speech quality and intelligibility," *Speech Commun.*, vol. 80, no. C, pp. 84–94, Jun. 2016.

[28] B. Cauchi *et al.*, "Predicting the quality of processed speech by combining modulation-based features and model trees," in *Proc. ITG Conf. Speech Commun.*, Paderborn, Germany, Oct. 2016, pp. 180–184.

[29] E. Frank, Y. Wang, S. Ingliss, G. Holmes, and I. W, "Using model trees for classification," *Mach. Learn.*, vol. 32, no. 1, pp. 63–76, 1998.

[30] 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, Lang. Process.*, vol. 18, no. 7, pp. 1766–1774, Sep. 2010.

[31] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Vancouver, BC, Canada, May 2013, pp. 6645–6649.

[32] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1310–1318.

[33] J. Santos and T. Falk, "Blind room acoustics characterization using recurrent neural networks and modulation spectrum dynamics," in *Proc. AES 60th Int. Conf.*, Leuven, Belgium, Feb. 2016, pp. 1–8.

[34] B. Cauchi *et al.*, "Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech," *EURASIP J. Adv. Signal Process.*, vol. 2015, pp. 1–12, Jul. 2015.

[35] A. Jukić, T. van Waterschoot, and S. Doclo, "Adaptive speech dereverberation using constrained sparse multichannel linear prediction," *IEEE Signal Process. Lett.*, vol. 24, no. 1, pp. 101–105, Jan. 2017.

[36] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAMO: A British English speech corpus for large vocabulary continuous speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Detroit, MI, USA, May 1995, pp. 81–84.

[37] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "The ACE challenge—Corpus description and performance evaluation," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, NY, USA, 2015, pp. 1–5.

[38] *Objective Measurement of Active Speech Level*, ITU-T Recommendation P.56, Mar. 1993.

[39] C. Breithaupt, M. Krawczyk, and R. Martin, "Parameterized MMSE spectral magnitude estimation for the enhancement of noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2008, pp. 4037–4040.

[40] I. Gradshteyn and I. Ryzhik, *Table Integrals, Series, Products*. Boston, MA, USA: Academic, 1994.

[41] K. Kinoshita *et al.*, "Summary of the reverb challenge," Sep. 2015. [Online]. Available: http://reverb2014.dereverberation.com/workshop/reverb_summary.pdf

[42] C. S. J. Doire *et al.*, "Single-channel online enhancement of speech corrupted by reverberation and noise," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 3, pp. 572–587, Mar. 2017.

[43] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[44] K. Lebart, J. M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech de-reverberation," *Acta Acoustica*, vol. 87, pp. 359–366, 2001.

[45] C. Breithaupt, T. Gerkmann, and R. Martin, "A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Las Vegas, NV, USA, Apr. 2008, pp. 4897–4900.

[46] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[47] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[48] J. Bitzer, K. U. Simmer, and K.-D. Kammeyer, "Multi-microphone noise reduction techniques as front-end devices for speech recognition," *Speech Commun.*, vol. 34, pp. 3–12, 2001.

[49] H. Cox, R. M. Zeskind, and M. M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-35, no. 10, pp. 1365–1376, Oct. 1987.

[50] *General Performance Objectives Applicable to All Modern International Circuits and National Extension Circuits*, ITU-T Recommendation G.151, Mar. 1980.

[51] T. H. Falk *et al.*, "Objective quality and intelligibility prediction for users of assistive listening devices," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, no. 2, pp. 114–124, Mar. 2015.

[52] F. Chollet *et al.*, "Keras," 2015. [Online]. Available: https://github.com/fchollet/keras

[53] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[54] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, 2014.

[55] R. C. Sprinthall and S. T. Fisk, *Basic Statistical Analysis*. Englewood Cliffs, NJ, USA: Prentice-Hall, 2013.

[56] *Statistical Evaluation Procedure for POLQA*, ITU-T Standard TD 12rev1 (WP 2/12), Mar. 2009.

Benjamin Cauchi received the M.Sc. degree in acoustic and signal processing applied to music from the University Pierre and Marie Curie, Paris VI, Paris, France. He worked toward his Ph.D. degree as an ESR Marie Curie Fellow in the project group Hearing, Speech, and Audio Technology, Fraunhofer Institute for Digital Media Technology (IDMT), Oldenburg, Germany, and the Signal Processing Group, University of Oldenburg, Oldenburg, Germany. Before starting his Ph.D. degree, he was a Research Assistant with Fraunhofer IDMT, and Imperial College London, U.K. He is currently researcher in the OFFIS Institute for Information Technology, where his research focuses on machine learning for speech applications and on the development of technologies for the hearing aid industry.

Kai Siedenburg studied mathematics and musicology at Humboldt University Berlin and as a Fulbright visiting student at the University of California, Berkeley. He received the Ph.D. degree in Music Technology from McGill University, Montreal. He is currently a Marie Sklodowska-Curie Individual Postdoctoral Fellow with the University of Oldenburg, Oldenburg, Germany. He is a Co-Editor of the upcoming Springer handbook entitled *Timbre: Acoustics, Perception, and Cognition*. He received the best paper award at the 2017 International Conference on Digital Audio Effects (DAFX) in Edinburgh, U.K.

João F. Santos received the bachelor's degree in electrical engineering from the Federal University of Santa Catarina, Florianópolis, Brazil, in 2011 and the M.Sc. degree in telecommunications in 2014 from the Institut National de la Recherche Scientifique, Montréal, QC, Canada, where he entered the Dean's honour list and was awarded the Best M.Sc. Thesis Award and is currently working towards the Ph.D. degree in telecommunications. His main research interest is in the applications of deep learning to speech and audio signal processing applications (speech enhancement, speech synthesis, speech recognition, and audio scene classification).

Tiago H. Falk (SM'14) received the B.Sc. degree from the Federal University of Pernambuco, Recife, Brazil, in 2002, and the M.Sc. and Ph.D. degrees from Queens University, Kingston, ON, Canada, in 2005 and 2008, respectively, all in electrical engineering. From 2009 to 2010, he was an NSERC Postdoctoral Fellow with Holland-Bloorview Kids Re-habilitation Hospital, affiliated with the University of Toronto. Since 2010, he has been with the Institut National de la Recherche Scientifique, Montréal, QC, Canada, where he heads the Multimedia/Multimodal Signal Analysis and Enhancement Laboratory. His research interests lie at the intersection of multimedia and biomedical signal processing and their interplay in the development of anthropomorphic technologies.

Simon Doclo (S'95–M'03–SM'13) received the M.Sc. degree in electrical engineering and the Ph.D. degree in applied sciences from the Katholieke Universiteit Leuven, Leuven, Belgium, in 1997 and 2003, respectively. From 2003 to 2007, he was a Postdoctoral Fellow with the Research Foundation Flanders, Electrical Engineering Department, Katholieke Universiteit Leuven, and the Cognitive Systems Laboratory, McMaster University, Canada. From 2007 to 2009, he was a Principal Scientist with NXP Semiconductors at the Sound and Acoustics Group, Leuven, Belgium. Since 2009, he has been a Full Professor with the University of Oldenburg, Oldenburg, Germany, and Scientific Advisor for the project group Hearing, Speech and Audio Technology, Fraunhofer Institute for Digital Media Technology, Ilmenau, Germany. His research activities center around signal processing for acoustical and biomedical applications, more specifically microphone array processing, speech enhancement, active noise control, acoustic sensor networks, and hearing aid processing. He received the Master Thesis Award of the Royal Flemish Society of Engineers in 1997, the Best Student Paper Award at the International Workshop on Acoustic Echo and Noise Control in 2001, the EURASIP Signal Processing Best Paper Award in 2003 (with Marc Moonen), and the IEEE Signal Processing Society 2008 Best Paper Award (with Jingdong Chen, Jacob Benesty, Arden Huang). He is member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing, the EURASIP Special Area Team on Acoustic, Speech and Music Signal Processing, and the EAA Technical Committee on Audio Signal Processing. He was and is involved in several large-scale national and European research projects (ITN DREAMS, Cluster of Excellence Hearing4All, CRC Hearing Acoustics). He was Technical Program Chair of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) in 2013 and Chair of the ITG Conference on Speech Communication in 2018. In addition, he was a Guest Editor for several special issues (IEEE SIGNAL PROCESSING MAGAZINE, *Elsevier Signal Processing*) and is an Associate Editor for IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING and *EURASIP Journal on Advances in Signal Processing*.

Stefan Goetze received the Ph.D. degree in 2013 from the University of Bremen, Bremen, Germany, where he was Research Engineer from 2004 to 2008. He is the Head of the Group Automatic Speech Recognition and Department Head of the Department Hearing, Speech, and Audio Technology, Fraunhofer Institute for Digital Media Technology IDMT, Oldenburg, Germany. He has been a Lecturer with the University of Bremen since 2007 and Project Leader of national and international research projects in the field of acoustic signal enhancement and recognition technologies. His research interests are sound pick/up, processing, and enhancement such as noise reduction, acoustic echo cancellation, and dereverberation, as well as assistive technologies, human–machine interaction, detection and classification of acoustic events, and automatic speech recognition.