On the Predictive Power of Objective Intelligibility Metrics for the Subjective Performance of Deep Complex Convolutional Recurrent Speech Enhancement Networks

Femke B. Gelderblom, Tron V. Tronstad, Torbjørn Svendsen, Tor Andre Myrvoll

***Abstract*—Speech enhancement (SE) systems aim to improve the quality and intelligibility of degraded speech signals obtained from far-field microphones. Subjective evaluation of the intelligibility performance of these SE systems is uncommon. Instead, objective intelligibility measures (OIMs) are generally used to predict subjective performance increases. Many recent deep learning (DL) based SE systems, are expected to improve the intelligibility of degraded speech as measured by OIMs.**

However, validation of the ability of these OIMs to predict subjective intelligibility when enhancing a speech signal using DL-based systems is lacking. Therefore, in this study, we evaluate the predictive performance of five popular OIMs. We compare the metrics' predictions with subjective results. For this purpose, we recruited 50 human listeners, and subjectively tested both single channel and multi-channel Deep Complex Convolutional Recurrent Network (DCCRN) based speech enhancement systems.

We found that none of the OIMs gave reliable predictions, and that all OIMs overestimated the intelligibility of 'enhanced' speech signals.

***Index Terms*—Speech enhancement, intelligibility, objective metrics, subjective evaluation**

I. Introduction

BUSINESSES have embraced online meetings at a never-before-seen rate during the Covid-19 pandemic. As societies are opening up again, many organizations are adopting combinations of remote and on-location work. So-called 'hybrid' meetings, with both in-office and remote participants, are becoming increasingly common.

The quality and intelligibility of the audio is crucial to the meeting experience, but those on the remote end often find themselves straining to hear what is being said by in-office participants that do not use near-mouth microphones. Both noise and reverberance degrade the intelligibility and quality of speech [1], [2]. Far-field microphones, such as those embedded into a webcam, ceiling-mounted conference systems, or table-top speakerphones, inevitably pick up noise and reverberation, hence reducing both the quality and intelligibility of the transmitted speech signal.

F.B. Gelderblom and T.V. Tronstad are with the Acoustics group of the Department of Sustainable Communication Technologies, SINTEF Digital, Trondheim, Norway. Torbjørn Svendsen, Tor Andre Myrvoll and F.B. Gelderblom (additionally) are with the Signal Processing group at the Department of Electronic Systems, Norwegian University of Science and Technology, Trondheim, Norway

Manuscript received XXXX; revised XXXX

As such, speech enhancement (SE) of far-field microphone recordings for online meetings is more relevant than ever. Hand in hand comes the need to ensure that we have reliable tools for measuring the performance of SE systems; this is the topic of our study.

Microsoft has organised multiple challenges to stimulate research on improving the *quality* of noisy and reverberant speech signals and simultaneously open-sourced a subjective evaluation framework for this purpose [3]–[5]. This has resulted in several State-Of-The-Art-Systems that significantly improve the quality of single channel speech signals. For the Interspeech 2020 Deep Noise Suppression (DNS) challenge [3], Hu *et al.* proposed the deep complex convolution recurrent network (DCCRN) [6], which won the real-time-track. For the ICASSP 2021 DNS challenge [4], it was Li *et al.* who proposed the winning system: a two-stage complex network with a low-complexity post-processing scheme (TSCN-PP) [7]. The authors later extended this network into the simultaneous speech denoising and dereverberation network (SDDNet) [8], which became the winner of the third DNS challenge [5].

All of these networks (and many other competitors) improved the *subjective quality* of speech: human listeners rated the output of these SE systems as having higher quality than the noisy input speech. As such, the challenges had two (arguably equally) important outcomes: Not only did they stimulate the development of better SE systems, they also led to a far more widespread reliance on subjective evaluation of system performance with respect to quality. Evidence for the significance of the latter was, for example, provided by Li *et al.* who found that including the proposed post-processing step of their winning system was consistently preferred by listeners, even though the objective measures had predicted the opposite effect [7].

Reducing noise, distortion and reverberance, should not only be beneficial for quality (how comfortable or annoying the sound is to listen to), but also for *intelligibility*. Intelligibility can be assessed both subjectively (with listening tests) and objectively (with mathematical metrics). Subjective intelligibility can be evaluated using either qualitative or quantitative approaches. Qualitative assessments ask the test subjects to rate the intelligibility of a word/sentence (e.g. rating it on a ordinal scale with help text such as *fully intelligible*, *slightly raised effort to understand*, *difficult to understand*, and *impossible to* *understand*, etc.) and are an indirect measure of intelligibility. Speech recognition threshold (SRT), on the other hand, is an example of a quantitative measure that gives a number related to the ratio of words/sentences correctly perceived and is a direct measure of intelligibility. It has been found that qualitative measures often are biased, and care should be taken when using them [9].

Since subjective testing is costly and time consuming, objective intelligibility measures (OIMs) are the most common method for evaluating the intelligibility performance of speech enhancement systems. These metrics can be either 'intrusive' or 'non-intrusive'. 'Intrusive' means they require the clean reference/target speech in addition to the noisy/distorted/processed signal to be evaluated. Generally speaking, the intelligibility score is then based on some measure of mathematically defined (human hearing inspired) closeness between the signals. Their non-intrusive counterparts usually have less predictive power [10], and during the training of supervised speech enhancement systems, the clean reference signal is readily available. As such, intrusive measures of intelligibility are logical choices for the evaluation of speech enhancement systems.

While the previously mentioned DNS challenges focused on *subjective quality*, many of the participants also provided objective performance scores of their systems with respect to intelligibility, recognizing the importance of the latter. Most of the studies (i.e. [11]–[21]) provided short-time objective intelligibility (STOI [22], [23]) scores, while a few (i.e. [16], [24], [25]) presented extended STOI (ESTOI [26]) scores.

However, OIMs have their limitations and do not necessarily work well for complex nonlinear DNN-based processing, or for the more realistic degradations of speech signals that include reverberation and distortion [27]–[31]. This means that the intelligibility performance of SE systems should be checked with subjective testing. Yet, it is rare to see SE systems being evaluated *subjectively* for intelligibility. Notable exceptions to this observation come from the field of SE for hearing impaired users, where a limited number of research groups have put considerable effort into systematically testing their denoising or speech separation systems subjectively. Over the years, they have published single channel models that improve subjective intelligibility for different levels of generalization (for example from known speakers to complete language mismatch, and from overlapping noise samples to completely unseen noise types) and from simpler to more complex degradations (including reverberation and non-stationary noises) [28], [31]–[39]. The difficulty of improving subjective intelligibility is evident from the fact that we were unable to find any studies demonstrating subjectively improved intelligibility of noisy reverberant single channel speech, under combined novel noise and unseen speaker/speech conditions.

While the studies of [28], [31]–[39] mostly focus on applications for the hearing impaired, their results are also highly relevant for the setting of online meetings. One general conclusion we can draw from these studies, is that it seems to be easier to provide benefit to those that struggle the most. Subjective intelligibility is quantitatively measured by means of the speech recognition threshold (SRT) of a subject: the SRT is the level where the subject can repeat 50 % of the speech material correctly. Hearing impaired listeners have elevated SRTs, meaning their intelligibility scores are lower at relatively high signal to noise ratios (SNRs). At these higher SNRs, SE systems have to remove less noise to recover the clean speech, than at the lower SNRs where people with normal hearing start to struggle.

For the meeting experience, the intelligibility *should* be (close to) 100 %, which requires SNRs well above the SRTs of normal hearing subjects. If SNRs are that high, quality would be the more important factor. However, it is all too common to see poorly placed equipment and sub-optimal sound absorption in meeting rooms, which often leads to problematic SNRs and reduced intelligibility. Humans are also extremely apt at 'guessing' content from context, and report higher intelligibility when the topic of conversation is familiar. This happens at the cost of increased listening effort, making such meetings more tiring than they would have been if the speech signals had been clearer. Additionally, retirement age is increasing, and international cooperation is well-established, so many meeting participants do have elevated SRTs due to (mild) hearing loss and/or unfamiliarity with the language, which reduces their ability to guess from context. Therefore, we argue that intelligibility is highly relevant also at the higher SNRs that one may expect for hybrid meetings from a decent conference room.

Subjective evaluation is currently the only way to determine how a particular SE system actually affects intelligibility, but objective metrics are much faster and simpler to use. Relying solely on subjective testing would impede progress, but we do need to validate the use of OIMs on modern SE systems.

The latest comprehensive review of intrusive objective intelligibility measures was published by Kuyk *et al.* in 2018 and evaluates a total of 12 such metrics across a large range of subjective listening test results on different datasets of degraded and enhanced audio. The novelty of our study is in its focus on the enhancement by DL-based models, as none of the datasets in Kuyk *et al.* 's review included deep learning based enhancement models. In that sense, our study is closer to the recent studies of Keshavarzi *et al.* and Naithani *et al.*, who evaluated the effect of their respective DL-based SE systems on subjective evaluation [40], [41]. However, these studies relied on a qualitative method that measures whether the listener *believes* that the intelligibility of the speech signal has been improved.

Our study is similar in aim to our previous studies from 2017/2018 [29], [42], but now conducted with more complicated neural networks (with convolution and recurrent layers, instead of only feed-forward layers) that estimate a mask for the degraded signal (instead of the enhanced signal itself). Apart from being technically different, the OIMs also predict much higher speech enhancement performance for the SE systems of this study, thus giving higher expectations for the subjective performance. Additionally, the current study evaluates the predictive performance of five objective intelligibility metrics, instead of just one. Recently, López-Espejo *et al.* also published results of a study with an aim similar to ours. Important differences between our and their study are

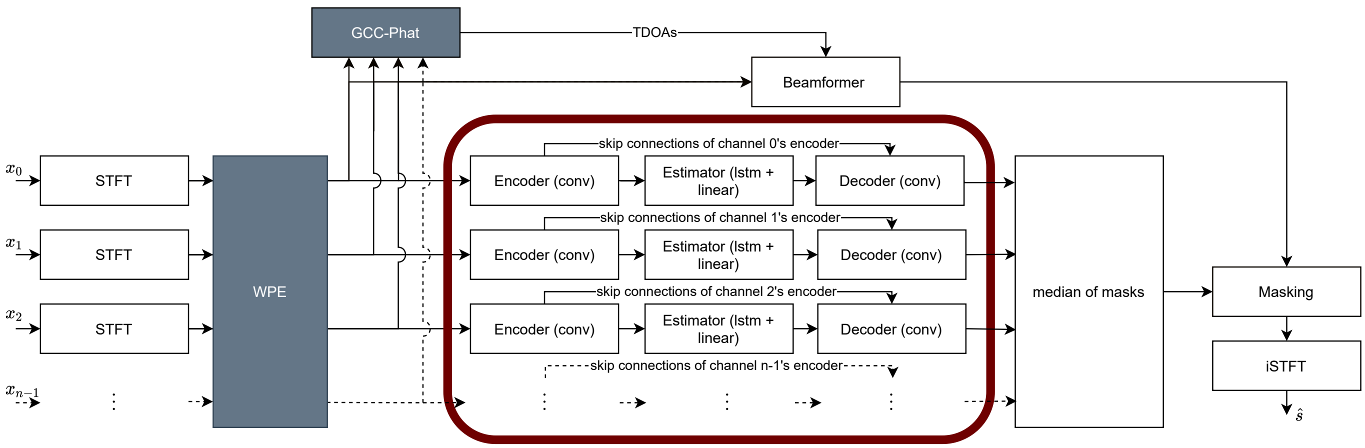

Fig. 1. Overview of the proposed speech enhancement and dereverberation system. The highlighted WPE and GCC-Phat boxes are only employed during inference. The red frame contains all blocks with trainable parameters, where each Encoder-Estimator-Decoder structure represents a single channel DCCRN. The input of the system is the noisy reverberant multichannel speech signal x, and the output is the estimate of the clean speech signal $\hat{s}$. Figure taken from [44] (©2021 IEEE).

the different types of networks used, and the focus of their approach on using different OIMs as training loss, rather than as evaluation tools [43].

In this study, we compare objective predictions to subjective results of both the multi-channel and single channel DCCRN speech enhancement systems from [44] and [6]. We have taken particular care to create a challenging and realistic test set, where speech is made reverberant with room impulse responses (RIRs) *recorded* in the same meeting room as where the noise was recorded. Speakers do not necessarily look at the microphone (array), which leads to a weaker direct path to the microphone, and more reverberant input. Furthermore, there is a speaker and language mismatch as training data did not include Norwegian, the language used for the subjective evaluation. Subjective intelligibility was evaluated by obtaining speech recognition thresholds for 50 participants, representing both native and non native office workers with or without self reported normal hearing.

We would like to emphasize here that this study does *not* focus on the evaluation of state-of-the-art DL-based SE systems, nor on whether such systems can provide benefit to its users. Instead, we aim to investigate the reliability of commonly used OIMs to evaluate such systems. This less common focus is motivated by the ample use of OIMs for this exact purpose, despite there being no clear evidence of the suitability of these OIMs for DL-based SE systems. Instead, the literature provides us with some evidence of the opposite [28]–[31]. We argue that the systems evaluated in this study are representative enough of modern SE systems to provide evidence concerning the predictive power of objective intelligibility metrics.

II. Speech Enhancement Systems

A. Problem formulation

A speech signal from a single stationary speaker, recorded by a single microphone at a fixed position in stationary room conditions can be expressed as:

$$x = h * s + v, \quad (1)$$

where x, s and v are the noisy, clean and noise time domain signals, respectively. Furthermore, h denotes the frequency response of the reverberation filter, which is time invariant, as long as relative positions between the speaker, the microphone and the reflective surfaces in the room do not change. The noise signal may come from one or more sources, and each of these will have their own reverberance, but all of these signal components are here collected in the definition of v.

As a microphone array is nothing more than a collection of microphones (each located at a unique location), the problem can be expanded to a multi-channel problem using index i for each microphone element:

$$x_i = h_i * s + v_i, \qquad i = 1 \ldots N, \quad (2)$$

where N is the number of microphone elements in the array.

Both the noise, v, and the reverberance, h, degrade the intelligibility and quality of the speech, s. The ultimate goal of speech enhancement is therefore to recover the speech signal s from the single or multi-channel noisy signal x.

B. SE models

From a machine learning perspective, it is natural to formulate the speech enhancement problem in terms of supervised regression. This requires two matching datasets containing corrupted input samples of x and their respective clean speech targets s. A model is then trained to minimize the difference between these two, using a suitable loss function and output formulation (often defined as a mask) that ideally puts extra weight on differences that are particularly important for human perception.

Figure 1 shows an overview of our multi-channel system first proposed in [44]. Specifically, the system denoted "MPDR (oracle TDOAs) + Single channel DCCRN" was selected for the subjective evaluation of this study, without any changes to its implementation.

This system builds upon the challenge winning single channel DCCRN system proposed in [6]. At its input, the multi-channel corrupted speech signal is taken to the Fourier domain by a short-time Fourier transform (STFT) operation. The STFTs for each channel are then passed through a weighted prediction error (WPE) block for dereverberation [45]. Single channel DCCRN blocks estimate N masks (one for each channel), all of which are then collapsed into a single mask using the median operator. Finally, this mask is applied to a beamformed version of the output of the WPE blocks, before the enhanced signal is converted back to a time-signal using an inverse STFT block.

During beamforming, the channels of a multi-channel signal are delayed, weighted, and then combined into a single signal that is steered towards a specific source/direction. This so-called steering vector requires time difference of arrival (TDOA) values. During training, the system knows the true speaker direction. During evaluation, we can either estimate TDOAs by performing generalized cross correlation with phase transform (GCC-Phat) [46] on the dereverberated WPE output, or set them to the true TDOAs.

For the beamformer, we rely on the minimum power distortionless response (MPDR) beamformer. This beamformer is also often referred to as a specific implementation of the popular minimum variance distortionless response (MVDR) beamformer, where the implementation differentiates itself from the general MVDR beamformer, by deriving the distortionless filter for a specified steering direction that minimizes the mean square output *power*, and as such, it requires only the corrupted input signal. To avoid ambiguity, we have chosen to comply with Van Trees' practice of referring to it as the MPDR beamformer [47]. While the MPDR is known to be more sensitive to the correctness of the steering vector, it has the advantage that it allows for separation of the steering vector and mask estimation processes as it does not need estimates of the noise statistics. Further implementation details of the MPDR-based multi-channel system and its superior objective performance over several baseline systems are given in [44].

We evaluate two variants of this multi-channel system (with oracle and unknown TDOAs), in addition to the single-channel DCCRN it is based on. All processing conditions were implemented in Python and the neural models were implemented in PyTorch. To ensure we obtain the change caused by the DCCRN component of the systems over the results that we would have obtained just with beamforming and WPE dereverberation, we also define a relevant baseline for each of the three systems. This gives us a total of six processing conditions:

- **Baseline 1, Noisy**: Single channel noisy and reverberant speech.
- **Baseline 2, MPDR (estimated TDOAs)**: Multi-channel noisy and reverberant speech that has been dereverbed with WPE and beamformed with the MPDR beamformer, where TDOAs were estimated using GCC-Phat on the noisy reverberant input.
- **Baseline 3, MPDR (oracle TDOAs)**: Multi-channel noisy and reverberant speech that has been dereverbed with WPE and beamformed with the MPDR beamformer, using oracle TDOAs.
- **SE system 1, DCCRN**: Single channel noisy and reverberant speech passed through a WPE block and a single channel DCCRN SE model.
- **SE system 2, MPDR (estimated TDOAs) + DCCRN**: Multi-channel noisy and reverberant speech that has been passed through the complete multi-channel system shown in 1. Here TDOAs are estimated from the dereverberated output of the WPE blocks using GCC-Phat.
- **SE system 3, MPDR (oracle TDOAs) + DCCRN**: Multi-channel noisy and reverberant speech that has been passed through the complete multi-channel system shown in 1. Here oracle TDOAs are used.

C. Training Data

The performance of deep learning based SE models is highly dependent on the data that these models are trained on. Training data needs to be varied enough to cover all possible use cases, and realistic enough to avoid mismatch during later use. Supervised training also requires that the desired target is available. Therefore, we have taken the common approach of corrupting clean speech with suitable noise and reverberance.

We relied on the DNS Challenge 2021 speech and noise data, as it is a high quality database that covers multiple languages and many different types of noises. For the RIRs, we used the ISM-dir dataset described in [30]. These RIRs are simulated using the image source method with the addition that all speaker sources are modelled as directive sources with an average male/female speaker pattern directivity.

Training input samples were generated in an 'online' manner, meaning that new samples were generated during training from convolving random samples of speech with random RIRs and then adding (non-reverberant) random noise. In 20 % of the cases the speech was also left non-reverberant. We experimented both with reverberant and non-reverberant speech as target samples during training, and found the reverberant speech to work best, as objective testing showed that the DCCRN network was not able to remove reverberance.

III. Evaluation

A. Evaluation Data

In order to directly compare results, we used the same dataset for the objective and the subjective evaluations.

We chose a highly common type of meeting noise for the evaluation, with transient components produced by typing on a keyboard on a background of mostly stationary noise from the air conditioning system. More than an hour of this type of noise was *recorded* with a 9-channel circular microphone array (planar) with 4 cm radius, positioned on a table approximately in the middle of a typical rectangular meeting room with dimensions 4.5 x 3.8 x 2.6 m, and a reverberation time ($RT60_{1kHz}$) of 0.3 s.

RIRs were then recorded with the same microphone array in the same room, at various speaker positions and orientations. More details on how these RIRs were obtained can be found in [30]. We included both the RIR recordings for speakers facing the array, and the RIRs for speakers facing away at a 90 degree angle.

We used these noise and RIR recordings to corrupt the clean speech material from the Norwegian speech-in-noise test (see

Section III-C). For each sentence and SNR, a random clip of noise and a random RIR was selected to corrupt the clean speech to all SNRs ranging from -36 dB to 10 dB, with a 2 dB stepsize.

As such, we obtained a challenging evaluation dataset with an unknown and unseen noise type, recorded RIRs that 'faced' towards or away from the array, and speech material from an unknown speaker in a language that was not present in the training material.

B. Objective evaluation

Being able to objectively determine the intelligibility of a speech signal has been relevant since the invention of telephony, over a hundred years ago. This eventually lead to the Articulation Index (AI) [48], which was standardized in 1969 and revised in 1997, into an updated metric called the 'speech intelligibility index' (SII) [49]. In 1980, the speech transmission index (STI) was proposed [50], which can account for some simple nonlinear degradations such as clipping. All these metrics are still in use today.

However, these metrics are based on long-term signal statistics, which make them unsuitable for non-stationary noise and enhancement algorithms that introduce distortions. Several metrics have been proposed to improve upon these important limitations and we evaluate five of these metrics that are commonly used when testing speech enhancement systems. All of the tested metrics have been validated to monotonically relate to subjective intelligibility under the conditions for which they were designed [22], [23], [26], [51]–[54].

All of these metrics are intrusive, which means they require both the corrupted signal and a corruption free reference signal, as input. The metrics then estimate intelligibility based on a mathematical measure of similarity between these two signals. Intrusive measures generally perform better than their non-intrusive counterparts, making intrusive testing the obvious choice in cases like ours where the reference signal is readily available [10].

For objective testing, we have obtained predictions for each metric for the entire evaluation dataset. For this study, we have only included metrics that aim to estimate intelligibility, as it is also the intelligibility of speech that will be estimated subjectively. Therefore, commonly used metrics to evaluate the quality of speech (such as PESQ [55] and HASQI [56]) are omitted. The evaluated metrics are:

1) NCM (normalized covariance metric): The normalized covariance measure (originally proposed in [51]) is closely related to the STI. First, both the corrupted signal and the clean reference are band-pass filtered into different frequency bands. Then the normalized covariance (the Pearson correlation coefficient) is calculated for all the temporal envelopes of the reference and corrupted frequency bands. The normalized covariances are converted to apparent SNRs for each frequency, which are clipped and averaged into a single score using frequency dependent weights. We relied on the implementation from [52] for the calculation of NCM scores, using the updated signal dependent weights proposed in [57]. Van Kuyk *et al.* found that this NCM implementation works well for datasets where a speech enhancement system has post-processed degraded speech, but had less correlation with subjective results for datasets where speech was only degraded, or where enhancement had been added as a pre-processing step (before the speech was corrupted) [27].

2) CSII (coherence speech intelligibility index): The CSII metric attempts to extend the SII metric by making it applicable to a wider range of distortions, where SII was designed specifically for additive noise. Instead of finding the SNR of each frequency band, the signal-to-distortion ratio (SDR) is estimated for each band, based on the coherence between the corrupted speech, and the clean reference signal. Speech segments are also divided into three energy level based categories, and different weights determine the contribution of low-, mid- and high-level speech segment scores (CSII_{Mid},CSII_{Mid} and CSII_{Mid}, respectively) to the total CSII score [53]. We have relied on the implementation from [52]. Van Kuyk *et al.* found that a slightly different implementation of the CSII score had acceptable predictive power on most datasets (in terms of improved correlation coefficients), but notably struggled with datasets where speech enhancement was applied as a post processing step [27].

3) STOI (short-time objective intelligibility): STOI has been specifically designed to deal with noisy speech processed with time-frequency (TF) weighting techniques. To ensure that the effect of local TF degradation is taken into account, signals are segmented into short-time windows, and the overall score is obtained by averaging the scores of all segments. These scores themselves depend on the Pearson correlation coefficient between the temporal envelopes of the corrupted and clean reference speech signals [22], [23]. We relied on the implementation provided by the original authors. STOI is possibly the most popular OIM within the field of speech enhancement, but multiple studies have noted its limitations for evaluating performance of DNN-based SE systems [28], [29], [31], [42].

4) ESTOI (extended STOI): ESTOI is similar to STOI, but does not assume mutual independence between frequency bands and incorporates spectral correlation, to improve its performance on modulated noise sources [26]. We relied on the implementation provided by the original authors. Van Kuyk *et al.* found that ESTOI was one of the higher performing metrics, but noted that 'its usefulness is limited to situations where noise is the main source of degradation'. Zhao *et al.* found that ESTOI especially underestimated intelligibility of unprocessed noisy-reverberant speech [31].

5) HASPI (hearing-aid speech perception index): HASPI was first introduced in [54], and later updated to better predict the intelligibility of reverberant speech (HASPI version 2) [58]. We relied on the implementation of version 2 that we obtained from the original authors through direct communication. HASPI allows for intelligibility predictions based on the subject's hearing loss, but we assumed normal hearing conditions for all calculations. This means that during calculation, both the corrupted signal and its reference were passed through the same auditory model, giving two sets of envelope modulation features. These outputs are then passed through an ensemble of neural networks that have been fit to subjective intelligibility data. HASPI has the most complicated auditory model of the tested metrics, and Van Kuyk found HASPI (version 1) to be the overall top performing intrusive metric [27].

Subjective intelligibility is not just dependent on the speech degradation, but also on the test setup. As such, it is common to map predicted scores to subjective results for a given test setup [26]. In order to obtain intelligibility predictions for our specific subjective evaluation setup, we have mapped the OIM scores to the subjective results of our single channel noisy and reverberant baseline. Crucial concepts here are that the mapping is monotonic, and kept equal for all the six processing conditions defined in Section II-B.

For STOI and ESTOI, we relied on the mapping proposed in their respective papers [22], [23], [26]

$$\hat{I} = \frac{100}{1 + \exp\left(a\tilde{I} + b\right)}, \tag{3}$$

where $\hat{I}$ is the predicted intelligibility, $\tilde{I}$ the predicted score, and a and b are the coefficients to be determined with the non-linear least squares method. This mapping was empirically found to also work well for NCM scores. For HASPI (which has already been fit to subjective data), we found that a simple translation along the SNR-axis leads to a closer match. For the CSII metric, we used non-linear least squares to fit our data to the mapping proposed by the original authors [53],

$$c = a_1 + a_2\text{CSII}_{\text{Low}} + a_3\text{CSII}_{\text{Mid}} + a_4\text{CSII}_{\text{High}}, \tag{4}$$

$$\hat{I} = \frac{100}{1 + \exp\left(-c\right)}, \tag{5}$$

where the tunable parameters are the coefficients a.

We relied on the paired Wilcoxon rank sum test (also called the Mann-Whitney U test), which is a non-parametric test for paired observations that does not assume normality of distributions, for testing whether results obtained for the different processing conditions are significantly different. First we obtain the SNR where the metric predicts 50 % intelligibility for the noisy single channel processing condition. Then we test, pairwise, for equality of the population medians of the scores obtained at this SNR, for the noisy single channel baseline condition, and the 5 remaining processing conditions defined in Section II-B.

C. Subjective evaluation

To compare the different SE models subjectively, we determined the speech recognition threshold (SRT) of each subject, for each processing condition. To find the SRTs, we relied on a Norwegian implementation of a Hagerman test [59]. In this test, each sentence is built up as follows: [Name], [Verb], [Numeral], [Adjective], [Noun]. There are 10 possible options for each class of words, giving 10^5 possible unique sentences, but for practical purposes we relied on a subset of 500 unique sentences from this database.

When evaluating subjective intelligibility, it is important to control for the so-called ‘learning effect’, as listeners are known to become better at speech-in-noise tests with practice [59], possibly biasing the evaluation to the order in which models are tested. This is one of the advantages of Hagerman tests; that there are only ten possible words for each of the five word categories means that the listener quickly becomes familiar with all the possible answers, speeding up the learning effect [60]. Once the learning effect has been established, the listeners’ answers directly represent the effects of the processing condition we wish to evaluate. Therefore, each subject was asked to complete a training round of the speech-in-noise test, before completing the six different processing conditions in an order that was randomized for each individual. For the subjective evaluation of the different SE models, we recruited 50 (25 male and 25 female) office workers. Our recruitment process was intentionally inclusive also to those who may struggle more in such meetings, either because they suspect/know their hearing is not optimal, or because they are not native speakers of Norwegian. None of the participants had participated in any form of speech-in-noise test in the past year. We were not required to notify the Norwegian Centre for Research Data (NSD) about our study as we collected only anonymous data.

Three subgroups of volunteers were recruited: 1) 15 native listeners with self-reported normal hearing, 2) 16 native listeners with self-reported known/suspected hearing loss, and 3) 19 non-native listeners with self-reported normal hearing. These different subgroups were recruited to ensure a wide spread in SRTs. This allowed us to evaluate whether models provide a different benefit to those who already struggle with speech understanding at higher SNRs (i.e. the non-native listeners or those with hearing loss) than to normal hearing listeners that can understand speech at extremely low SNRs.

After completing the evaluation, all subjects were reassigned to three new subgroups based on the SRTs they obtained for the noisy baseline condition: low ($n = 16$), medium ($n = 17$) or high ($n = 16$). This ensured minimal intelligibility performance spread within each subgroup, as it compensates for the fact that self-reported hearing loss was found to be a rather poor indicator of SRTs and that, as expected, the results of the non-native volunteers was highly dependent on the number of years of their experience with Norwegian, and on the closeness of Norwegian to their mother tongue. By reassigning candidates we obtained three approximately equally sized subgroups that are maximally homogeneous in SRT. Results from one subject were discarded as this subject’s complete unfamiliarity with the language caused intelligibility scores to be lower than the SRT threshold (50 %) even at the highest test SNRs.

The chosen Norwegian speech-in-noise test was implemented in Matlab, and allowed subjects to complete the procedure independent of an operator. The program presented the subject with a graphical user interface that showed ten possible words for each of the five word categories. Each noisy/processed 5-word sentence was presented only once, and the subject was asked to click on all the words he/she had recognized. Guessing was allowed, but the test was not forced choice.

Responses were recorded and scored automatically and used as input to an adaptive psychometric function estimation procedure called the Ψ method [61]. Using this procedure,

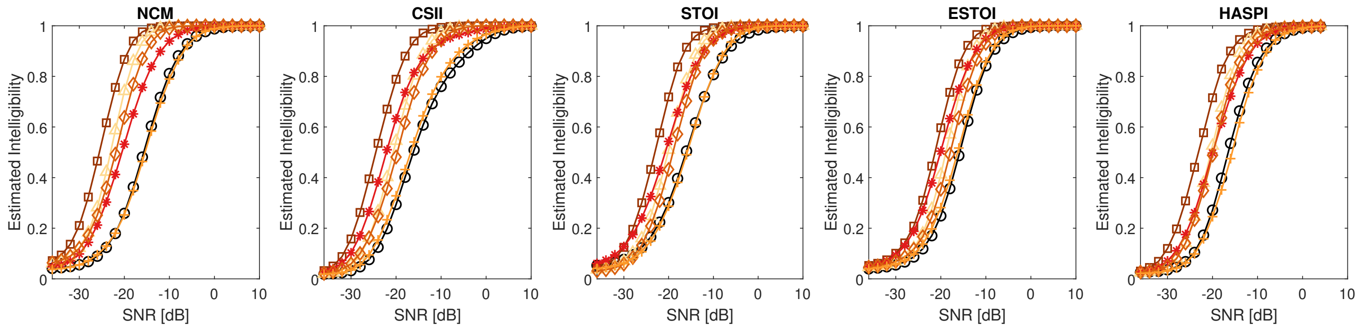

Fig. 2. Intelligibility versus SNR predicted by each metric, for the following conditions: noisy, Single channel DCCRN, MPDR (estimated TDOAs), , MPDR (estimated TDOAs) + single channel DCCRN, MPDR (oracle TDOAs), MPDR (oracle TDOAs) + single channel DCCRN.

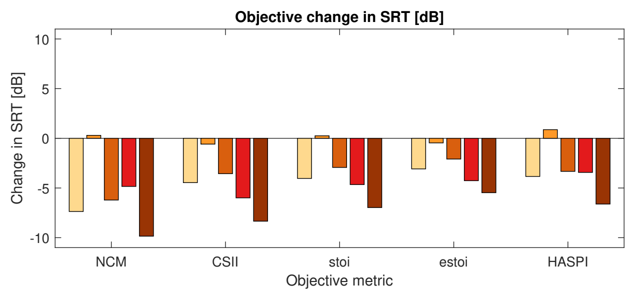

Fig. 3. Change in SRT predicted by each metric, when the following systems are compared to the single channel noisy condition: Single channel DCCRN, MPDR (estimated TDOAs) only, MPDR (estimated TDOAs) + single channel DCCRN, MPDR (oracle TDOAs) only, MPDR (oracle TDOAs) + single channel DCCRN. Negative numbers indicate improvement in speech intelligibility.

the routine continuously estimated the SRT and slope of the psychometric function during the test. The final threshold estimate was obtained after 20 sentences.

Subjects were encouraged to take small breaks in between models, were allowed to repeat the training round (though none did), and could adjust the volume of the test to their own preferred setting. All participants received a 150 NOK ($\approx$ 15 EUR) voucher for their effort.

Experiments were conducted in the sound insulated lab of SINTEF's acoustics group. Sentences were presented binaurally through a Sennheiser HD 600 type headphone.

We again relied on the paired Wilcoxon rank sum test for testing our null hypothesis. Here we tested pairwise for equality of the population medians of the SRT scores (obtained for each subject) for the single channel noisy condition, versus the 5 remaining processing conditions defined in Section II-B.

IV. Results

A. *Objective results*

Figure 2 shows the predicted psychometric functions of intelligibility for the six different processing conditions defined in Section II-B, and the five different objective metrics.

Figure 3 summarises these objective predictions by presenting the change in SRT predicted by each metric when five of these processing conditions are compared to the remaining noisy single channel condition. Here, negative values indicate a reduced SRT (i.e. increased intelligibility).

The changes in SRTs shown in Figure 3 were found to be highly significant ($p \ll 0.01$) for all but one of the systems. Namely, for the MPDR on its own and with estimated TDOAs, only ESTOI and HASPI predicted (small but) significant changes ($p < 0.05$), while all other metrics predicted insignificant changes ($p > 0.05$). All metrics therefore predict increased intelligibility for all our DL-based SE systems.

We see similar trends across metrics in the predictions. The objective measures do not necessarily agree on how much improvement the systems provide, but performance gain is nonetheless predicted whenever we compare a DCCRN-based system to its appropriate baseline, or the noisy single channel condition. Additionally, all metrics predict that beamforming on its own (without DCCRN involvement) gives increased intelligibility over the single channel noisy condition, but only for oracle TDOAs. When the TDOAs are unknown, beamforming is predicted to have little to no effect at all. The noisy (unprocessed) input is expected to give the lowest intelligibility independent of the predictive measure chosen, and all metrics predict that the combined MPDR + DCCRN (with oracle DOAs) will give the highest intelligibility.

B. *Subjective results*

Figure 4 shows the subjective results for all processing conditions, together with their respective objective predictions for each different metric. The results are averaged over the 16 respondents with SRTs below -15 dB on the single channel noisy baseline: our best hearing subjects.

Objective scores for the single channel noisy condition are mapped to the corresponding subjective results as described in Section II-B. The mappings work equally well for all metrics, as evident from the overlap of all the lines in the top-left plot (noisy) in Figure 4. All mappings slightly underestimate the slope of the psychometric function, but even at the extreme ends, the differences between objective scores and subjective answers are minor. The same mapping also works reasonably well for the other baseline systems (Figure 4, top row), although there seems to be a slight systematic overestimation of intelligibility performance of the MPDR with oracle TDOAs.

When we move our attention to the DCCRN-based systems (Figure 4, bottom row), we see that all metrics overestimate intelligibility since all the colored lines are above the black

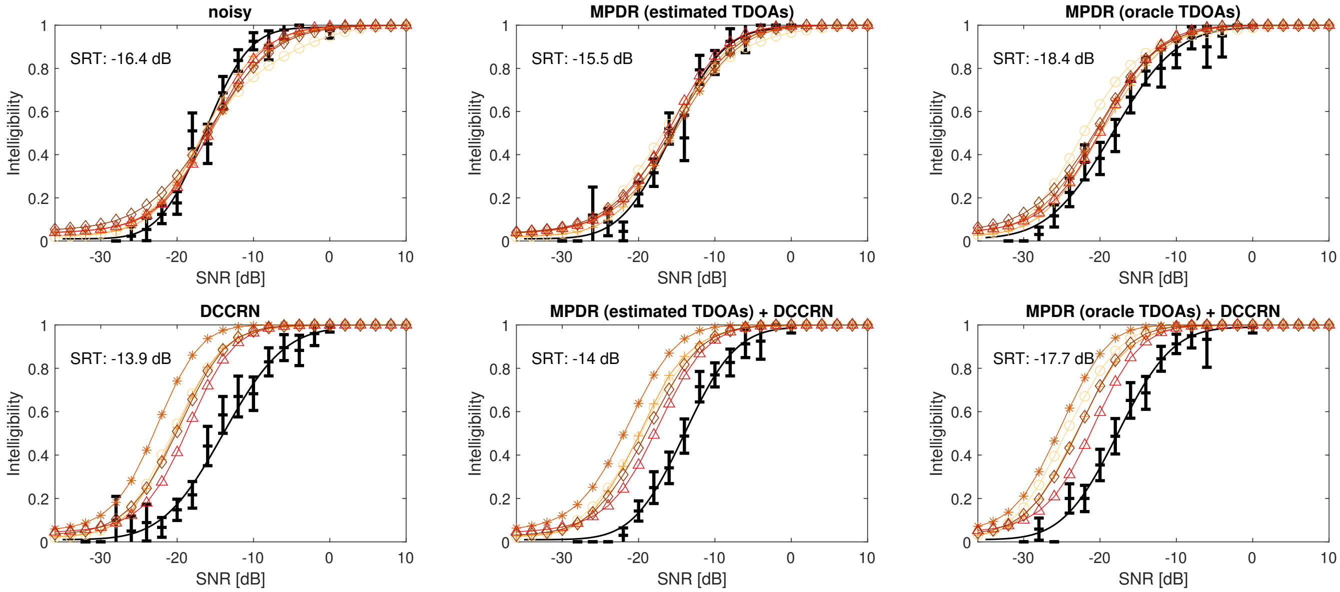

Fig. 4. Psychometric functions obtained from normal hearing native speakers (n=14) for different processing pipelines. The subjective responses (error bars indicating confidence intervals) and their logistic fits are shown in black (——), together with the corresponding predictions from the objective metrics: CSII, HASPI, NCM, ESTOI, and STOI.

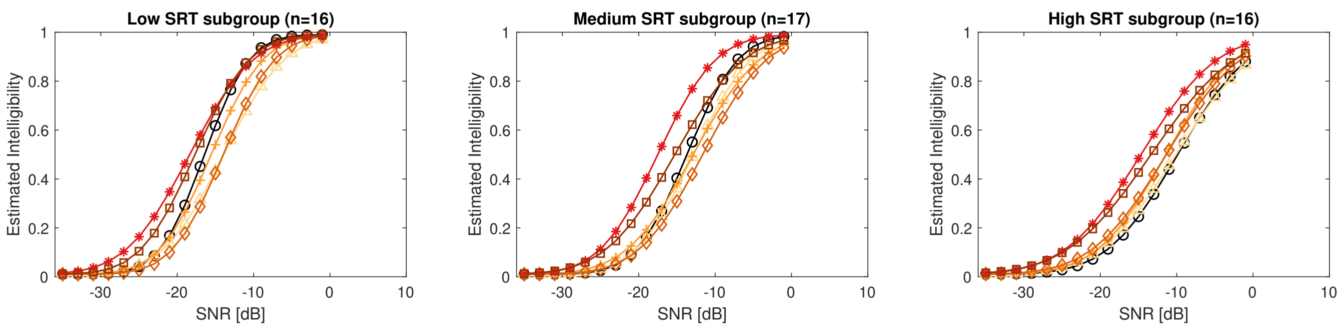

Fig. 5. Subjective Intelligibility versus SNR for each subgroup of subjects: noisy, Single channel DCCRN, MPDR (estimated TDOAs), MPDR (estimated TDOAs) + single channel DCCRN, MPDR (oracle TDOAs), MPDR (oracle TDOAs) + single channel DCCRN.

lines. This is not only true close to the SRT (SNR at intelligibility 50 %), but across the entire intelligibility range.

Looking at the subjective results, the two systems based on an MPDR supplied with oracle TDOAs (Figure 4, rightmost column) are the only ones that lead to lower SRT scores when compared to the noisy input (indicating improved intelligibility). All other forms of processing make the noisy input less intelligible. Here it is important to note that the MPDR (oracle TDOAs) system *without* a DCCRN outperforms the system *with* a DCCRN.

Figure 5 shows the subjective results for all three subject groups and the six processing conditions. Here we can see that all systems with a DCCRN have comparable performance or do worse than their respective baselines, over the entire range of test SNRs. Only the systems with a MPDR that knows where the speaker is (with or without DCCRN), clearly outperform the single channel noisy baseline.

This observation is also apparent in Figure 6. Here statistically significant changes are marked with an asterisk. For the low SRT group (our best hearing subjects), processing with a single channel DCCRN significantly reduces intelligibility, while for the other groups, the change in SRT is insignificant. An MPDR that needs to estimate the direction of speech (the MPDR with unknown TDOAs) neither degrades nor improves the signal for any of the groups. When a DCCRN is added to this type of MPDR, we see a degradation of speech intelligibility for both the low and medium SRT groups. In contrast, the MPDR-based systems where the TDOAs are known, do significantly improve intelligibility. Here it is important to note that the system *without* a DCCRN consistently outperforms the combined system.

V. Discussion

The objective of this study has been to assess different OIMs that are commonly used to evaluate DL-based speech enhancement systems.

All OIMs indicated that the intelligibility of noisy speech signals should be improved by the tested DCCRN-based

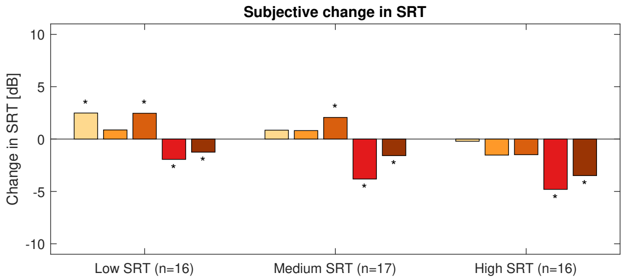

Fig. 6. Change in SRT for each group of subjects, when the following systems are compared to the single channel noisy condition: Single channel DCCRN, MPDR (estimated TDOAs) only, MPDR (estimated TDOAs) + single channel DCCRN, MPDR (oracle TDOAs) only, MPDR (oracle TDOAs) + single channel DCCRN. Positive numbers indicate a degradation in speech intelligibility. Statistically significant changes are marked with an *.

systems. However, the subjective evaluation of these systems shows the opposite: the DCCRN-based systems degrade the intelligibility of the noisy speech. Therefore, the SE systems fail to deliver their expected performance, proving that the OIMs are unreliable.

For our subgroup with the highest SRTs (those that tolerate the least amount of noise) the SE systems seem to be doing little harm. So for them, the systems may have merit if the focus is only on quality, but this is not true for the other two subgroups (the better-hearing subjects with higher language familiarity).

Additionally, it is important to note that the metrics predicted significant increases in intelligibility for all SNRs, making the OIMs unreliable across the range. All tested metrics wrongly predicted that the DL-based SE systems would improve the SRTs of the listeners.

Therefore, no correlation scores are presented since this could be taken as a pointer to which metric is the 'least wrong'. It is not possible to conclude that any of the tested metrics is *more* suitable to be used in the development of DL-based SE systems than the others.

Note also that even though all metrics correctly predicted approximately the right shape of the psychometric function, this is merely an effect of using an appropriate mapping function – it is the SRT that provides the developer with information. The fact that the psychometric functions of the top row of Figure 4 match reasonably well indicates that the reliability of the metrics specifically break down when the DCCRN network becomes part of the processing pipeline, whereas beamforming on its own does not.

There is an ever-growing number of DL-based SE systems proposed in the literature, and the exact results presented here do not directly carry over to other set-ups. However, our findings are alarming since they reveal that the performance results estimated by the commonly used OIMs were not reliable for these systems and the observed trends have strong implications also for he expected performance of OIMs for similar systems.

In earlier studies, we namely also obtained similar results for very different DNN-based systems [29], [42]. The current study expands on the earlier findings by studying more advanced types of networks (where the OIMs also predict much higher increases in performance), and over a larger range of metrics. That the findings remain the same despite these differences further strengthens the conclusion that OIMs cannot be assumed to be reliable predictors of subjective speech intelligibility for DNN-based speech enhancement systems. As such, we conclude that DL-based SE systems should be evaluated subjectively at least until the reliability of the OIM has been validated for the chosen SE system.

As subjective evaluation seldom is used to evaluate the intelligibility improvement of DL-based SE systems, we have not found any evidence that OIMs have ever been reported to reliably predict the quantitative subjective intelligibility performance of DL-based SE systems. Instead, recently López-Espejo *et al.* also found no strong positive monotonic relationship between objective and subjective intelligibility for their systems, which in contrast to ours are based on a *fully* convolutional neural network (FCNN) architecture and trained with a range of different loss functions. Like us, López-Espejo *et al.* conclude that subjective evaluation of DNN-based SE systems cannot systematically be replaced by objective intelligibility evaluation as of today [43].

That the tested OIMs do not provide reliable predictions for the system's tested here, naturally does not mean they are necessarily unreliable for all DL-based systems. Further work is required to identify exactly where the limitations of the metrics lie and what causes their lack of reliability. A major challenge for such a study is the black-box nature of DL-based systems. From the current study, it is not possible to determine what causes the OIMs to be off in their predictions. This is essential for the development of new OIMs, and should be looked into in future research. However, as the OIMs were unreliable over the entire range of SNRs tested, we are able to conclude that they are unreliable, not only for extremely low SNRs, or vice versa. The fact that the tested OIMs (despite their different approaches) all struggle with predicting the intelligibility of the samples processed with the DNN-based SE systems, while being reasonably accurate for the beamformed samples, suggests it is something in the highly non-linear nature of DNN processing that causes their failure. However, from the current study, it is impossible to include how or why this occurs. Recently, a new type of OIMs in the form of neural prediction networks has emerged [62]–[64]. Further study is required to evaluate whether such metrics are better suited to evaluate DNN-based SE networks, than the more commonly used OIMs tested in this study.

Finally, we note the potential of an MPDR beamformer that knows the speaker location. That beamforming works (as long as you know where to steer the beam), is not new knowledge, but it does provide us with a clear opportunity to avoid the issue of objective intelligibility predictions altogether. Instead of relying on OIMs to develop and evaluate multi-channel SE systems, focus could be moved to speaker localization. For this study we used a TDOA estimation algorithm that can easily be improved upon; even so, the results shown in Figure 6 already suggest it was close to starting to provide benefit to those subjects with the lowest SRTs.

Most importantly, the advantage of direction of arrival estimation is that the error between target and estimate is mathematically speaking well defined, and in no way dependent on human hearing and perception.

Absolute intelligibility scores (i.e. the exact ratio of speech understood at a specific SNR) are highly dependent on test conditions: the type of noise, type of test, presence of context, lengths of sentences, etc. From the SE system developer's point of view, these absolute intelligibility scores obtained for a specific processing condition are therefore not that crucial. Instead, we need tools to reliably predict whether a specific type of processing enhances or reduces speech intelligibility. Until such tools are available, we suggest that developers of DL-based SE systems incorporate early subjective intelligibility testing on human listeners. While this step requires more resources, it will ensure that design choices are based on actual intelligibility improvements, instead of predictions from metrics that very well may be unreliable.

VI. Conclusion

We have evaluated the predictive power of five popular OIMs (i.e.: NCM, CSII, STOI, ESTOI and HASPI) by comparing objective prediction to subjective results for single-channel and multi-channel DCCRN-based SE systems. All metrics predicted increased intelligibility across the entire range of relevant SNRs. The results from the subjective tests tell a different story: performance is either worse, or insignificantly different. Predictions were unreliable across the entire range of SNRs, including the higher SNRs that are the most relevant for the online meeting scenario.

Therefore we conclude that there are severe limitations to the usefulness of these OIMs for the purpose of developing DNN-based SE systems.

Acknowledgments

We thank all test subjects for their contribution.

References

[1] K. S. Helfer and L. A. Wilber, "Hearing Loss, Aging, and Speech Perception in Reverberation and Noise," *Journal of Speech, Language, and Hearing Research*, vol. 33, no. 1, pp. 149–155, Mar. 1990.

[2] A. A. de Lima, S. L. Netto, L. W. P. Biscainho, F. P. Freeland, B. C. Bispo, R. A. de Jesus, R. Schafer, A. Said, B. Lee, and T. Kalker, "Quality Evaluation of Reverberation in Audioband Speech Signals," in *E-Business and Telecommunications*, J. Filipe and M. S. Obaidat, Eds. Berlin, Germany: Springer, 2009, vol. 48, pp. 384–396.

[3] C. K. A. Reddy, E. Beyrami, H. Dubey, V. Gopal, R. Cheng, R. Cutler, S. Matusevych, R. Aichner, A. Aazami, S. Braun, S. Srinivasan, and J. Gehrke, "The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Speech Quality and Testing Framework," in *INTERSPEECH*. Shanghai, China: ISCA, 2020, pp. 2492–2496.

[4] C. K. A. Reddy, H. Dubey, K. Koishida, A. Nair, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "Interspeech 2021 Deep Noise Suppression Challenge," in *INTERSPEECH*. Brno, Czechia: ISCA, 2021, pp. 2796–2800.

[5] C. K. A. Reddy, H. Dubey, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "ICASSP 2021 Deep Noise Suppression Challenge," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Toronto, Canada: IEEE, Jun. 2021, pp. 6623–6627.

[6] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement," in *INTERSPEECH*. Shanghai, China: ISCA, 2020, pp. 2472–2476.

[7] A. Li, W. Liu, X. Luo, C. Zheng, and X. Li, "ICASSP 2021 Deep Noise Suppression Challenge: Decoupling Magnitude and Phase Optimization with a Two-Stage Deep Network," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. Toronto, ON, Canada: IEEE, Jun. 2021, pp. 6628–6632.

[8] A. Li, W. Liu, X. Luo, G. Yu, C. Zheng, and X. Li, "A Simultaneous Denoising and Dereverberation Framework with Target Decoupling," in *INTERSPEECH*. Brno, Czechia: ISCA, Jun. 2021, pp. 2801–2805.

[9] E. C. Poulton, "Quantitative subjective assessments are almost always biased, sometimes completely misleading*," *British Journal of Psychology*, vol. 68, no. 4, pp. 409–425, 1977.

[10] 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, Mar. 2015.

[11] X. Hao, X. Su, R. Horaud, and X. Li, "Fullsubnet: A Full-Band and Sub-Band Fusion Model for Real-Time Single-Channel Speech Enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. Toronto, ON, Canada: IEEE, Jun. 2021, pp. 6633–6637.

[12] T. Vuong, Y. Xia, and R. M. Stern, "A Modulation-Domain Loss for Neural-Network-based Real-time Speech Enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. Toronto, Canada: IEEE, Feb. 2021, pp. 6643–6647.

[13] S. Zhao, T. H. Nguyen, and B. Ma, "Monaural Speech Enhancement with Complex Convolutional Block Attention Module and Joint Time Frequency Losses," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. Toronto, Canada: IEEE, Feb. 2021, pp. 6648–6652.

[14] X. Le, H. Chen, K. Chen, and J. Lu, "DPCRN: Dual-Path Convolution Recurrent Network for Single Channel Speech Enhancement," in *INTERSPEECH*. Brno, Czechia: ISCA, 2021, pp. 2811–2815.

[15] X. Li and R. Horaud, "Online Monaural Speech Enhancement Using Delayed Subband LSTM," in *INTERSPEECH*. Shanghai, China: ISCA, Oct. 2020, pp. 2462–2466.

[16] K. Oostermeijer, Q. Wang, and J. Du, "Lightweight Causal Transformer with Local Self-Attention for Real-Time Speech Enhancement," in *INTERSPEECH*. Brno, Czechia: ISCA, 2021, pp. 2831–2835.

[17] M. Strake, B. Defraene, K. Fluyt, W. Tirry, and T. Fingscheidt, "INTERSPEECH 2020 Deep Noise Suppression Challenge: A Fully Convolutional Recurrent Network (FCRN) for Joint Dereverberation and Denoising," in *INTERSPEECH*. Shanghai, China: ISCA, Oct. 2020, pp. 2467–2471.

[18] N. L. Westhausen and B. T. Meyer, "Dual-Signal Transformation LSTM Network for Real-Time Noise Suppression," in *INTERSPEECH*. Shanghai, China: ISCA, Oct. 2020, pp. 2477–2481.

[19] Z. Xu, M. Strake, and T. Fingscheidt, "Deep Noise Suppression With Non-Intrusive PESQNet Supervision Enabling the Use of Real Training Data," in *INTERSPEECH*. Brno, Czechia: ISCA, 2021, pp. 2806–2810.

[20] K. Zhang, S. He, H. Li, and X. Zhang, "DBNet: A Dual-Branch Network Architecture Processing on Spectrum and Waveform for Single-Channel Speech Enhancement," in *INTERSPEECH*. Brno, Czechia: ISCA, Aug. 2021, pp. 2821–2825.

[21] X. Zhang, X. Ren, X. Zheng, L. Chen, C. Zhang, L. Guo, and B. Yu, "Low-Delay Speech Enhancement Using Perceptually Motivated Target and Loss," in *INTERSPEECH*. Brno, Czechia: ISCA, 2021, pp. 2826–2830.

[22] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Dallas, TX, USA: IEEE, 2010, pp. 4214–4217.

[23] ——, "An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[24] A. Li, W. Liu, X. Luo, C. Zheng, and X. Li, "ICASSP 2021 Deep Noise Suppression Challenge: Decoupling Magnitude and Phase Optimization with a Two-Stage Deep Network," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. Toronto, ON, Canada: IEEE, Jun. 2021, pp. 6628–6632.

[25] A. Li, W. Liu, X. Luo, G. Yu, C. Zheng, and X. Li, "A Simultaneous Denoising and Dereverberation Framework with Target Decoupling," in *INTERSPEECH*. Brno, Czechia: ISCA, 2021, pp. 2801–2805.

[26] J. Jensen and C. H. Taal, "An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[27] S. V. Kuyk, W. B. Kleijn, and R. C. Hendriks, "An Evaluation of Intrusive Instrumental Intelligibility Metrics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2153–2166, 2018.

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

[29] F. B. Gelderblom, T. V. Tronstad, and E. M. Viggen, "Subjective Intelligibility of Deep Neural Network-Based Speech Enhancement," in *INTERSPEECH*. Stockholm, Sweden: ISCA, Aug. 2017, pp. 1968–1972.

[30] F. B. Gelderblom, Y. Liu, J. Kvam, and T. A. Myrvoll, "Synthetic Data For DNN-Based DOA Estimation of Indoor Speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Toronto, Canada: IEEE, Jun. 2021, pp. 4390–4394.

[31] Y. Zhao, D. Wang, E. M. Johnson, and E. W. Healy, "A deep learning based segregation algorithm to increase speech intelligibility for hearing-impaired listeners in reverberant-noisy conditions," *The Journal of the Acoustical Society of America*, vol. 144, no. 3, pp. 1627–1637, Sep. 2018.

[32] E. W. Healy, S. E. Yoho, Y. Wang, and D. Wang, "An algorithm to improve speech recognition in noise for hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 134, no. 4, pp. 3029–3038, Oct. 2013.

[33] J. Chen, Y. Wang, S. E. Yoho, D. Wang, and E. W. Healy, "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *The Journal of the Acoustical Society of America*, vol. 139, no. 5, pp. 2604–2612, May 2016.

[34] J. J. M. Monaghan, T. Goehring, X. Yang, F. Bolner, S. Wang, M. C. M. Wright, and S. Bleeck, "Auditory inspired machine learning techniques can improve speech intelligibility and quality for hearing-impaired listenersa)," *The Journal of the Acoustical Society of America*, vol. 141, no. 3, pp. 1985–1998, Mar. 2017.

[35] T. Bentsen, T. May, A. A. Kressner, and T. Dau, "The benefit of combining a deep neural network architecture with ideal ratio mask estimation in computational speech segregation to improve speech intelligibility," *PLOS ONE*, vol. 13, no. 5, p. 13, May 2018.

[36] L. Bramsløw, G. Naithani, A. Hafez, T. Barker, N. H. Pontoppidan, and T. Virtanen, "Improving competing voices segregation for hearing impaired listeners using a low-latency deep neural network algorithm," *The Journal of the Acoustical Society of America*, vol. 144, no. 1, pp. 172–185, Jul. 2018.

[37] E. W. Healy, H. Taherian, E. M. Johnson, and D. Wang, "A causal and talker-independent speaker separation/dereverberation deep learning algorithm: Cost associated with conversion to real-time capable operation," *The Journal of the Acoustical Society of America*, vol. 150, no. 5, pp. 3976–3986, Nov. 2021.

[38] E. W. Healy, E. M. Johnson, M. Delfarah, D. S. Krishnagiri, V. A. Sevich, H. Taherian, and D. Wang, "Deep learning based speaker separation and dereverberation can generalize across different languages to improve intelligibility," *The Journal of the Acoustical Society of America*, vol. 150, no. 4, pp. 2526–2538, Oct. 2021.

[39] E. W. Healy, K. Tan, E. M. Johnson, and D. Wang, "An effectively causal deep learning algorithm to increase intelligibility in untrained noises for hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 149, no. 6, pp. 3943–3953, Jun. 2021.

[40] M. Keshavarzi, T. Reichenbach, and B. C. J. Moore, "Transient Noise Reduction Using a Deep Recurrent Neural Network: Effects on Subjective Speech Intelligibility and Listening Comfort," *Trends in Hearing*, vol. 25, p. 233121652110414, Jan. 2021.

[41] G. Naithani, K. Pietila, R. Niemisto, E. Paajanen, T. Takala, and T. Virtanen, "Subjective Evaluation of Deep Neural Network Based Speech Enhancement Systems in Real-World Conditions," in *2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP)*. Shanghai, China: IEEE, Sep. 2022, pp. 1–6.

[42] F. B. Gelderblom, T. V. Tronstad, and E. M. Viggen, "Subjective evaluation of a noise-reduced training target for deep neural network-based speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 3, pp. 583–594, Mar. 2019.

[43] I. López-Espejo, A. Edraki, W.-Y. Chan, Z.-H. Tan, and J. Jensen, "On the deficiency of intelligibility metrics as proxies for subjective intelligibility," *Speech Communication*, vol. 150, pp. 9–22, May 2023.

[44] F. B. Gelderblom and T. A. Myrvoll, "Deep Complex Convolutional Recurrent Network for Multi-Channel Speech Enhancement and Dereverberation," in *IEEE International Workshop on Machine Learning for Signal Processing*. Gold Coast, Australia: IEEE, Oct. 2021, pp. 1–6.

[45] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and Biing-Hwang Juang, "Speech Dereverberation Based on Variance-Normalized Delayed Linear Prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, Sep. 2010.

[46] M. Brandstein and H. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1. Munich, Germany: IEEE, 1997, pp. 375–378.

[47] H. L. Van Trees, *Optimim Array Processing*. Wiley, 2002.

[48] N. R. French and J. C. Steinberg, "Factors Governing the Intelligibility of Speech Sounds," *The Journal of the Acoustical Society of America*, vol. 19, no. 1, pp. 90–119, Jan. 1947.

[49] American National Standard, "ANSI/ASA S3.5-1997 (R2017): Methods For Calculation Of The Speech Intelligibility Index," 1997.

[50] T. Houtgast and H. J. M. Steeneken, "Evaluation of Speech Transmission Channels by Using Artificial Signals," *Acta Acustica united with Acustica*, vol. 25, no. 6, pp. 355–367, Dec. 1971.

[51] I. Holube and B. Kollmeier, "Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model," *The Journal of the Acoustical Society of America*, vol. 100, no. 3, pp. 1703–1716, 1996.

[52] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. CRC Press, 2013.

[53] J. M. Kates and K. H. Arehart, "Coherence and the speech intelligibility index," *The Journal of the Acoustical Society of America*, vol. 117, no. 4, pp. 2224–2237, 2005.

[54] ——, "The Hearing-Aid Speech Perception Index (HASPI)," *Speech Communication*, vol. 65, pp. 75–93, Nov. 2014.

[55] ITU, "Recommendation P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," 2001.

[56] j. m. kates and k. h. arehart, "The Hearing-Aid Speech Quality Index (HASQI) Version 2," *Journal of the Audio Engineering Society*, vol. 62, no. 3, pp. 99–117, Mar. 2014.

[57] J. Ma, Y. Hu, and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *The Journal of the Acoustical Society of America*, vol. 125, no. 5, pp. 3387–3405, 2009.

[58] J. M. Kates, "The Hearing-Aid Speech Perception Index (HASPI) Version 2," *Speech Communication*, pp. 35–46, 2021.

[59] J. Øygarden, "Norwegian speech audiometry," Ph.D. dissertation, Norwegian University of Science and Technology (NTNU), Trondheim, Norway, 2009.

[60] B. Hagerman and C. Kinnefors, "Efficient adaptive methods for measuring speech reception threshold in quiet and in noise," *Scandinavian Audiology*, vol. 24, no. 1, pp. 71–77, 1995.

[61] N. Prins and F. Kingdom, "Palamedes: Matlab routines for analyzing psychophysical data." http://www.palamedestoolbox.org, 2009.

[62] A. H. Andersen, J. M. De Haan, Z.-H. Tan, and J. Jensen, "Nonintrusive Speech Intelligibility Prediction Using Convolutional Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1925–1939, Oct. 2018.

[63] M. B. Pedersen, A. Heidemann Andersen, S. H. Jensen, and J. Jensen, "A Neural Network for Monaural Intrusive Speech Intelligibility Prediction," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Barcelona, Spain: IEEE, May 2020, pp. 336–340.

[64] R. E. Zezario, S.-W. Fu, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "Deep Learning-Based Non-Intrusive Multi-Objective Speech Assessment Model With Cross-Domain Features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 54–70, 2023.

Femke B. Gelderblom received the B.Sc. degree in applied physics, and the M.Sc. degree in biomedical engineering from Delft University of Technology, Delft, the Netherlands, in 2012. Since then, she has been working as a research scientist with the Acoustics group of SINTEF Digital, Trondheim, Norway. In March 2023, she obtained a Ph.D. degree from the Signal Processing group of the Norwegian University of Science and Technology, Trondheim, Norway. Her research interests include speech enhancement, deep learning, and microphone arrays.

Tron Vedul Tronstad received the M.Sc. degree in acoustics from the Department of Electronics and Telecommunications, Norwegian University of Science and Technology, Trondheim, Norway, and the Ph.D. degree from the Department of Electronic Systems, Norwegian University of Science and Technology, in 2007 and 2018, respectively. He is currently a research scientist with the Acoustics group at SINTEF Digital, Trondheim, Norway. His research interests include hearing and hearing damage.

Torbjørn Svendsen (Senior Member, IEEE) received the Siv.Ing (M.Sc.) and Dr.Ing. degrees from the Norwegian Institute of Technology (NTH), in 1980, and 1985, respectively. He is a Professor with the Department of Electronics and Telecommunications, Norwegian University of Science and Technology's (NTNU). Dr. Svendsen has been a Research Scientist with SINTEF before joining NTH as an Associate Professor in 1988. Since 1995, he has been a Professor of speech processing with NTNU. He has had extended research stays at AT&T Bell Laboratories, Murray Hill, NJ, USA, AT&T Labs, Florham Park, NJ, USA, Griffith University, Brisbane, Australia, Queensland University of Technology, Brisbane, Australia and MIT. His research interests include automatic speech recognition, speech synthesis, speech coding and speech analysis and modeling. He has authored or coauthored more than 90 papers in these areas. Prof. Svendsen is a IEEE Signal Processing Society (SPS) Senior Member and an International Speech Communication Association (ISCA) Board member. He has been a Member of the IEEE SPS Speech Processing Technical Committee.

Tor Andre Myrvoll received the Siv. Ing. and Dr. Ing. degrees in automatic speech recognition from the Department of Electronics and Telecommunications, Norwegian University of Science and Technology, in 1997 and 2002 respectively. He is currently a senior R&D engineer at Kongsberg Maritime, as well as Adjunct Associate Professor at the Department of Electronics and Telecommunications, Norwegian University of Science and Technology. His research interests include statistical signal processing for acoustic arrays and machine learning.