On the Deficiency of Intelligibility Metrics as Proxies for Subjective Intelligibility

Espejo, Ivan Lopez; Edraki, Amin ; Chan, Wai-Yip; Tan, Zheng-Hua; Jensen, Jesper

Published in:
Speech Communication

DOI (link to publication from Publisher):
10.1016/j.specom.2023.04.001

Publication date:
2023

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

Speech Communication 150 (2023) 9–22

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

On the deficiency of intelligibility metrics as proxies for subjective intelligibility

Iván López-Espejo [a,*], Amin Edraki [b], Wai-Yip Chan [b], Zheng-Hua Tan [a], Jesper Jensen [a,c]

[a] Department of Electronic Systems, Aalborg University, Aalborg, 9220, Denmark
[b] Department of Electrical and Computer Engineering, Queen's University, Kingston, K7L 3N6, Canada
[c] Oticon A/S, Smørum, 2765, Denmark

ARTICLE INFO

Keywords:
Speech enhancement
Speech intelligibility
Deep learning
Loss function
Intelligibility test

ABSTRACT

A recent trend in deep neural network (DNN)-based speech enhancement consists of using intelligibility and quality metrics as loss functions for model training with the aim of achieving high subjective speech intelligibility and perceptual quality in real-life conditions. In this study, we analyze a variety of loss functions, including some based on state-of-the-art intelligibility and quality metrics, to train an end-to-end speech enhancement system based on a fully convolutional neural network. The loss functions include perceptual metric for speech quality evaluation (PMSQE), scale-invariant signal-to-distortion ratio (SI-SDR), SI-SDR integrating speech pre-emphasis, short-time objective intelligibility (STOI), extended STOI (ESTOI), spectro-temporal glimpsing index (STGI), and a composite loss function combining STGI and SI-SDR. While DNNs trained with these loss functions produce notable speech intelligibility (and quality) gains according to pertinent objective metrics, we conduct a subjective intelligibility test that contradicts this result, showing no intelligibility improvement. From the results of this study, our conclusion is twofold: (1) subjective intelligibility evaluation is currently not replaceable by objective intelligibility evaluation, and (2) both the development of meaningful intelligibility metrics and DNN-based speech enhancement systems that can consistently improve the intelligibility of noisy speech for human listening remain open problems.

1. Introduction

Speech enhancement, which aims at improving both quality and intelligibility of distorted/noisy speech signals, has a wide range of applications in systems like hearing aids, mobile communication systems and automatic speech recognition systems (Wang and Chen, 2018). As for many other speech processing problems, the advent of deep neural networks (DNNs) was also a major turning point in speech enhancement, since DNNs boosted its performance, particularly when dealing with non-stationary noises and low signal-to-noise ratios (SNRs) (Kolbæk et al., 2017; Wang and Chen, 2018; Fu et al., 2020).

The earliest DNN-based speech enhancement systems typically were designed to work in the linear magnitude spectral domain (Xu et al., 2014; Martín-Doñas et al., 2017; Zhao et al., 2018a; Wang and Chen, 2018; Wang et al., 2021). In these systems, a DNN is responsible for estimating either the target (i.e., clean) speech magnitude spectrum or a time–frequency mask to be applied to the noisy magnitude spectrum to retrieve the target spectrum. Subsequently, the phase of the original noisy signal is appended to synthesize the enhanced signal (Xu et al., 2014; Martín-Doñas et al., 2017; Wang and Chen, 2018; Wang et al., 2021). However, because the use of the noisy signal phase constrains the potential speech enhancement performance, a relatively recent trend consists of the design of end-to-end systems directly estimating the enhanced waveform from the noisy one (Défossez et al., 2020; Wang et al., 2021; Zhang et al., 2021; Xiang et al., 2021). In fact, the end-to-end approach has consistently proven to be superior to approaches operating in the magnitude spectral domain (Wang et al., 2021).

A crucial aspect in any DNN-based speech enhancement system that has recently attracted much attention is the training loss function (Kolbæk et al., 2020; Braun and Tashev, 2021). Early DNN-based speech enhancement systems tended to use the popular mean squared error (MSE) in the linear frequency or time domain as a loss function (Gelderblom et al., 2017, 2019; Tan et al., 2019; Pandey and Wang, 2019; Kolbæk et al., 2020). In these domains, MSE might, arguably, not be the best choice for speech enhancement purposes (Loizou and Kim, 2011; Loizou, 2013), and, actually, some research (Fu et al., 2018; Kolbæk et al., 2020) demonstrates that there is no guarantee that low

* Corresponding author.
E-mail address: ivl@es.aau.dk (I. López-Espejo).

https://doi.org/10.1016/j.specom.2023.04.001
Received 18 January 2023; Received in revised form 10 March 2023; Accepted 14 April 2023
Available online 21 April 2023

MSE implies high perceptual quality and/or intelligibility. This is why a number of recent works has explored the use of perceptually-motivated loss functions (Fu et al., 2018; Zhao et al., 2018b; Martin-Doñas et al., 2018; Zhang et al., 2018; Fu et al., 2020; Kolbæk et al., 2020; Vuong et al., 2021; Li and Yamagishi, 2021; Borgström and Brandstein, 2021). In particular, aiming at maximizing the perceptual quality of the enhanced speech signals, approximations to the well-known perceptual evaluation of speech quality (PESQ) (Rix et al., 2001; ITU-T, 2003) metric are employed as training loss functions in, e.g., Martin-Doñas et al. (2018), Zhang et al. (2018), Fu et al. (2020), Li and Yamagishi (2021). Furthermore, intelligibility metrics (or approximations of them), which serve as proxies for subjective intelligibility, are considered as loss functions in, e.g., Fu et al. (2018), Zhao et al. (2018b), Zhang et al. (2018), Kolbæk et al. (2020), Li and Yamagishi (2021) in order to maximize the estimated intelligibility of the enhanced speech signals. Despite these speech enhancement systems tend to work very well at test time (i.e., when applied to noisy speech signals that are not seen during network training) in terms of the corresponding intelligibility metric considered for training, we are generally agnostic about their utility in real life because no formal intelligibility tests using a panel of subjects are conducted (Zhao et al., 2018b; Zhang et al., 2018; Wang et al., 2020; Kolbæk et al., 2020; Xiang et al., 2021).

In this paper, we study end-to-end monaural speech enhancement considering a variety of perceptually- and non-perceptually-motivated training loss functions: perceptual metric for speech quality evaluation (PMSQE) (Martin-Doñas et al., 2018), scale-invariant signal-to-distortion ratio (SI-SDR) (Roux et al., 2019), SI-SDR integrating speech pre-emphasis, short-time objective intelligibility (STOI) (Taal et al., 2011), extended STOI (ESTOI) (Jensen and Taal, 2016), spectro-temporal glimpsing index (STGI) (Edraki et al., 2021a), and a composite loss function combining STGI and SI-SDR. Our main finding is that, while DNN enhancement systems trained with these loss functions yield substantial improvements according to objective intelligibility (and quality) metrics, these improvements are contradicted by a subjective intelligibility test, which mostly indicates no improvement or even degradation of intelligibility. The main contributions of this manuscript can be summarized as follows:

1. Building upon (Kolbæk et al., 2020), we explore, for the first time, a recently proposed intelligibility metric as a loss function for training a DNN-based speech enhancement system: STGI (Edraki et al., 2021a, 2022). In particular, previous results suggest that STGI is more widely applicable and performs better than solidly-established intelligibility metrics like STOI and ESTOI (Edraki et al., 2021a).
2. Based on our preliminary observations indicating that integrating speech pre-emphasis into a loss function like SI-SDR may improve subjective intelligibility, we conduct a formal analysis to accept or reject this hypothesis.
3. In order to focus on both quality and intelligibility aspects during the enhancement process, we explore a composite loss function combining STGI and SI-SDR. Among all the evaluated loss functions, this composite loss function provides the best performance in terms of a variety of both intelligibility and quality metrics.
4. Differently from Kolbæk et al. (2020), we conduct a subjective intelligibility test, which studies two different types of models: *(1)* general-purpose models trained considering a variety of acoustic conditions, and *(2)* noisy condition-, language- and speaker-matched models trained in the same acoustic conditions as those to be found at test time. Differently from *(1)* that seeks to analyze a *practical scenario*, *(2)* examines an *ideal scenario* to determine an upper bound for speech intelligibility improvement. The test contradicts – by mainly indicating no improvement or even degradation – the intelligibility predicted by intelligibility metrics. Actually, similarly to Kolbæk et al. (2017), we only find statistically significant subjective intelligibility improvements over the unprocessed noisy signal for one noisy condition (−5 dB SNR, speech-shaped noise) in the idealized situation entailed by *(2)*.

While studies have reported subjective intelligibility improvements for hearing-impaired listeners (Healy et al., 2019) and cochlear implant users (Goehring et al., 2017) when using DNN-based speech enhancement, results are fewer for normal-hearing listeners. To the best of our knowledge, Healy et al. (2019) is one of the few works (which follows a time–frequency masking approach) claiming subjective intelligibility improvements for normal-hearing listeners, which, moreover, seems to only be achieved when the training and testing acoustic conditions match. Differently from Healy et al. (2019), in Gelderblom et al. (2017, 2019), Gelderblom et al. reach a similar finding to ours (i.e., an important discrepancy between subjective and objective intelligibility) when studying DNN-based speech enhancement operating in the linear magnitude spectral domain and trained making use of standard MSE as a loss function. The fact that the approach/framework of Gelderblom et al. *substantially differs from ours*, while their experimental results and ours are aligned, points to the following findings which are carefully elaborated throughout the rest of this manuscript:

1. Speech enhancement researchers and practitioners should avoid basing conclusions only on intelligibility metrics, which cannot replace subjective intelligibility tests.
2. Both the development of meaningful intelligibility metrics and DNN-based speech enhancement systems that can consistently improve the intelligibility of noisy speech for human listening remain open problems.

The remainder of this paper is organized as follows. Section 2 presents the end-to-end speech enhancement approach considered in this work, which covers the problem statement and DNN architecture, loss functions and training issues. Then, Section 3 outlines the speech and noise datasets employed for experimental purposes. The rest of the experimental framework and results are presented in Sections 4 and 5. Finally, Section 6 includes a discussion of the results and conclusions.

2. End-to-end speech enhancement

This section is devoted to explaining the end-to-end speech enhancement approach followed in this work. First, Section 2.1 states the problem addressed and describes the DNN architecture for speech enhancement. Second, Sections 2.2 and 2.3 present the different loss functions considered and DNN training issues, respectively.

2.1. Problem statement and architecture

Let $\mathbf{x} \in \mathbb{R}^L$ be a vector comprising L time-domain samples of a clean speech signal. Furthermore, let $\mathbf{n} \in \mathbb{R}^L$ represent an L-dimensional vector containing time-domain samples of a noise signal that contaminates $\mathbf{x}$ in accordance with an additive signal model:

$$\mathbf{y} = \mathbf{x} + \mathbf{n}, \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^L$ is the corresponding noisy speech signal. Then, our objective is to find a function implemented by a DNN, $\mathbf{f}(\cdot|\theta) : \mathbb{R}^L \rightarrow \mathbb{R}^L$, where θ corresponds to the DNN weight set, providing an estimate $\hat{\mathbf{x}}$ of the target/clean speech signal $\mathbf{x}$ from the noisy one $\mathbf{y}$, namely,

$$\hat{\mathbf{x}} = \mathbf{f}(\mathbf{y}|\theta) . \tag{2}$$

Notice that the DNN weight set θ is discriminatively estimated in a training phase using backpropagation in order to optimize any of the loss functions presented in the next subsection.

The non-linear mapping function $\mathbf{f}(\cdot|\theta) : \mathbb{R}^L \rightarrow \mathbb{R}^L$ is implemented in this work as a fully convolutional neural network (FCNN), a popular type of architecture for end-to-end speech enhancement purposes (Pandey and Wang, 2019; Kolbæk et al., 2020; Xu et al., 2021). In particular, to ease the computational load during the training

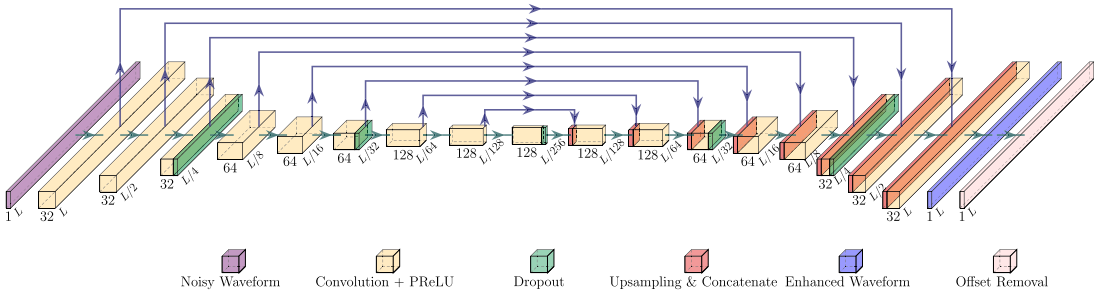


Fig. 1. Architecture of the fully convolutional neural network employed in this study to carry out end-to-end speech enhancement. Note that some of the layers integrate dropout with a rate of 0.2.

phase, we use a lighter variant where the number of feature maps of the FCNN is halved compared to the DNN employed in Kolbæk et al. (2020). The fact that we are able to significantly improve the metrics used as loss functions (see Section 4) proves that the size of this network is sufficient. The architecture employed in this study, which is depicted in Fig. 1, is fed with noisy speech signals comprised of $L = 19{,}712$ samples (corresponding to almost 2 s at a sampling rate of 10 kHz). Arranged in an encoder–decoder fashion (Ronneberger et al., 2015), this architecture primarily consists of 17 convolutional layers with a filter size of 11 samples and parameterized rectified linear unit (PReLU) (He et al., 2015) activations. The number of feature maps can be seen below each layer in Fig. 1. By using a stride of 2, the size of the input signal is repeatedly halved throughout the encoder up to $L/256$. Symmetrically, all the convolutional layers of the decoder apply upsampling by a factor of 2. Skip connections concatenating encoder and decoder feature maps as illustrated in Fig. 1 are also used. Except for the first layer, all the layers of the encoder use a stride of 2. The decoder convolutional layers employ a stride of 1.

The decoder is followed by an output convolutional layer – with a filter size of 1 sample and hyperbolic tangent activation – and, unlike in Kolbæk et al. (2020), a time-domain signal offset removal layer. The reason for adding the offset removal layer is that some of the considered loss functions, e.g., those based on intelligibility metrics like STGI, are offset-invariant and, therefore, they do not ensure offset-free enhanced speech signals. Let subscript l denote the lth element of a vector and let $\tilde{\mathbf{x}} \in \mathbb{R}^L$, $\tilde{\mathbf{x}} = \left(\tilde{x}_0, \ldots, \tilde{x}_l, \ldots, \tilde{x}_{L-1}\right)^\top$, be the input to the offset removal layer. This layer simply implements the following operation:

$$\hat{x}_l = \tilde{x}_l - \frac{1}{L}\sum_{l'=0}^{L-1} \tilde{x}_{l'}, \quad l = 0, 1, \ldots, L-1. \tag{3}$$

The receptive field of the architecture of Fig. 1 is 2,561 samples. This implies that, at a 10 kHz sampling rate,[1] the latency of this DNN-based speech enhancement system is approximately 256 ms. Moreover, the total number of parameters of this architecture is around 1.7M.

2.2. Loss functions

Immediately below, we briefly review the different perceptually- and non-perceptually-motivated training loss functions that we consider in this work for DNN-based speech enhancement. These are PMSQE (to be able to include speech quality aspects), SI-SDR (due to its excellent performance and popularity), SI-SDR with pre-emphasis (to analyze the potential advantages of emphasizing higher speech frequencies), (E)STOI (because of their state-of-the-art performance), STGI (due to its excellent performance in a wider range of conditions than (E)STOI) and a composite loss combining STGI and SI-SDR (to investigate the potential benefits of combining different loss functions targeting at different attributes of the speech signal). To the best of our knowledge, SI-SDR with pre-emphasis, STGI and the aforementioned composite loss are for the first time examined in this paper with the purpose of training DNN-based speech enhancement systems.

2.2.1. Perceptual metric for speech quality evaluation

The perceptual metric for speech quality evaluation (PMSQE) (Martin-Doñas et al., 2018) loss function, $\mathcal{L}_{\text{PMSQE}}$, is a differentiable approximation to the well-known, non-differentiable speech quality metric perceptual evaluation of speech quality (PESQ) (Rix et al., 2001; ITU-T, 2003). Particularly, PMSQE has a negative monotonic relationship with respect to PESQ. Hence, one would expect to maximize PESQ when minimizing $\mathcal{L}_{\text{PMSQE}} \in [0, 3]$. Similarly to PESQ, PMSQE is defined to operate at a sampling rate of either 8 kHz or 16 kHz. In this paper, a sampling rate of 8 kHz is considered when working with a DNN-based speech enhancement system relying on $\mathcal{L}_{\text{PMSQE}}$ (Kolbæk et al., 2020). The reader is referred to Martin-Doñas et al. (2018) for further details about PMSQE.

2.2.2. Scale-invariant signal-to-distortion ratio

The scale-invariant signal-to-distortion ratio (SI-SDR) (Roux et al., 2019), which was proposed as a scale-invariant alternative to the standard signal-to-distortion ratio (SDR) (Févotte et al., 2005), has shown to yield an outstanding performance in terms of different objective metrics when used as a training loss function for DNN-based speech enhancement (Luo and Mesgarani, 2019; Kolbæk et al., 2020). Expressed in dB, this loss function, $\mathcal{L}_{\text{SI-SDR}} \in (-\infty, +\infty)$, can simply be written as the negative of SI-SDR (Kolbæk et al., 2020),

$$\begin{aligned} \mathcal{L}_{\text{SI-SDR}} &= -\text{SI-SDR} \\ &= -10\log_{10}\left(\frac{\left\|\frac{\hat{\mathbf{x}}^\top \mathbf{x}}{\|\mathbf{x}\|^2}\mathbf{x}\right\|^2}{\left\|\frac{\hat{\mathbf{x}}^\top \mathbf{x}}{\|\mathbf{x}\|^2}\mathbf{x} - \hat{\mathbf{x}}\right\|^2}\right) \\ &= -10\log_{10}\left(\frac{\left(\mathbf{x}^\top \hat{\mathbf{x}}\right)^2}{\left(\mathbf{x}^\top \mathbf{x}\right)\left(\hat{\mathbf{x}}^\top \hat{\mathbf{x}}\right) - \left(\mathbf{x}^\top \hat{\mathbf{x}}\right)^2}\right). \end{aligned} \tag{4}$$

2.2.3. Scale-invariant signal-to-distortion ratio with pre-emphasis

Speech signals are defined by a low-pass characteristic for which higher frequency components have a lower dynamic range than the lower frequency components (Borgström and Brandstein, 2021). In spite of this fact, a loss function like SI-SDR does not differentiate between frequency regions, which may result in enhanced speech signals exhibiting poorly-estimated high frequency components in comparison with lower frequency content. In this work, we explore whether compensating for this fact by emphasizing higher frequencies during training can help to improve subjective intelligibility. To do this, we integrate pre-emphasis in SI-SDR as follows. Let $\delta(l)$ denote the unit impulse and let $h_l = \delta(l) - \alpha\delta(l-1)$ – where $\alpha = 0.98$ in this work –

[1] This is the typical operational sampling rate of intelligibility metrics like STOI, ESTOI and STGI.

correspond to the first-order pre-emphasis filter response. Furthermore, let $\bar{x}_l = h_l * x_l$ and $\hat{\bar{x}}_l = h_l * \hat{x}_l$, $l = 0, 1, \ldots, L-1$, denote the pre-emphasized target and enhanced speech signals, respectively, where $*$ refers to the convolution operator. As a result, SI-SDR with pre-emphasis is defined as

$$\mathcal{L}_{\text{SI-SDR+PE}} = -10\log_{10}\left(\frac{\left(\bar{\mathbf{x}}^{\top}\hat{\bar{\mathbf{x}}}\right)^2}{\left(\bar{\mathbf{x}}^{\top}\bar{\mathbf{x}}\right)\left(\hat{\bar{\mathbf{x}}}^{\top}\hat{\bar{\mathbf{x}}}\right) - \left(\bar{\mathbf{x}}^{\top}\hat{\bar{\mathbf{x}}}\right)^2}\right), \tag{5}$$

where, again expressed in dB, $\mathcal{L}_{\text{SI-SDR+PE}} \in (-\infty, +\infty)$.

2.2.4. Short-time objective intelligibility

Short-time objective intelligibility (STOI) (Taal et al., 2011) is still nowadays the most popular intelligibility metric (Kolbæk et al., 2020; Edraki et al., 2021; Xiang et al., 2021; Edraki et al., 2021a). In short, STOI is calculated as an average across time and frequency of normalized cross-correlation coefficients between one-third octave band short-time temporal envelope vectors of the target and degraded signals. Therefore, the STOI score for any given speech signal is a scalar $d_{\text{STOI}} \in [-1, 1]$. Accordingly, the STOI loss function is merely defined as $\mathcal{L}_{\text{STOI}} = -d_{\text{STOI}}$. Notice that, as pointed out by Kolbæk et al. (2020), STOI is differentiable except for the presence of a $\min(\cdot)$ operator whose gradient calculation has a computational complexity that is similar to that of the rectified linear unit (ReLU) activation function.

2.2.5. Extended short-time objective intelligibility

Extended short-time objective intelligibility (ESTOI) was proposed in Jensen and Taal (2016) aiming to improve STOI, which shows a poor performance for modulated noise sources. Similarly to STOI, ESTOI is calculated as an average of normalized cross-correlation coefficients between one-third octave band short-time temporal envelope vectors of the target and degraded signals. And, hence, for any given speech signal, the ESTOI score is also a scalar $d_{\text{ESTOI}} \in [-1, 1]$. In consequence, the ESTOI loss function can be simply defined as $\mathcal{L}_{\text{ESTOI}} = -d_{\text{ESTOI}}$. Notice that, unlike STOI, ESTOI does not comprise a $\min(\cdot)$ operator and is fully differentiable.

2.2.6. Spectro-temporal glimpsing index

Spectro-temporal glimpsing index (STGI) (Edraki et al., 2021a, 2022) is a recently proposed intelligibility metric that is based on the detection of glimpses (Cooke, 2006b) in short-time speech segments in the spectro-temporal modulation domain. Prior work (Edraki et al., 2021a) suggests that STGI can be effective over a broader range of degradation conditions, outperforming solidly-established intelligibility metrics like STOI and ESTOI. This is the reason behind studying STGI as a loss function for the first time in this work.

STGI uses a Gabor spectro-temporal modulation filterbank (Schädler et al., 2012; Edraki et al., 2021) applied to a log-magnitude spectrogram to extract spectro-temporal modulation envelopes. Specifically, $S = 11$ and $R = 4$ spectral and temporal Gabor modulation filters, respectively, are employed, leading to a total of $S \times R = 44$ spectro-temporal modulation envelopes. STGI computes normalized cross-correlation coefficients between the spectro-temporal modulation envelopes of the target and degraded signals $d^k_{\text{STGI}}(s, r) \in [-1, 1]$, where k indexes the short-time speech segment out of the K segments in which the input signal is divided, and s and r denote, respectively, the spectral and temporal modulation indices of the associated filters. The final STGI score, $d_{\text{STGI}} \in [0, 1]$, is obtained as (Edraki et al., 2021a)

$$d_{\text{STGI}} = \frac{1}{KSR}\sum_{k=1}^{K}\sum_{s=1}^{S}\sum_{r=1}^{R}\mathbb{1}_{+}\left(d^k_{\text{STGI}}(s, r) - \beta_{sr}\right), \tag{6}$$

where, in accordance with the glimpsing model of speech perception (Cooke, 2006b), $\mathbb{1}_{+}(\cdot) : \mathbb{R} \rightarrow \{0, 1\}$ is the indicator function, namely,

$$\mathbb{1}_{+}(x) = \begin{cases} 1, & x > 0, \\ 0, & x \leq 0, \end{cases} \tag{7}$$

and $-1 \leq \beta_{sr} \leq 1$ is a glimpsing threshold that is unique for each spectro-temporal index pair (s, r) as determined in Edraki et al. (2021a). Since the derivative of $\mathbb{1}_{+}(x)$ is 0 for $x \neq 0$ and undefined for $x = 0$, we replace the indicator function in Eq. (6) by the sigmoid function in order to define the STGI loss function as

$$\mathcal{L}_{\text{STGI}} = -\frac{1}{KSR}\sum_{k=1}^{K}\sum_{s=1}^{S}\sum_{r=1}^{R}\frac{1}{1 + e^{-\left(d^k_{\text{STGI}}(s,r) - \beta_{sr}\right)}}, \tag{8}$$

where $\mathcal{L}_{\text{STGI}} \in \left[-\left(1 + e^{-2}\right)^{-1}, -\left(1 + e^{2}\right)^{-1}\right]$ is fully differentiable.

2.2.7. Composite loss function

In the context of DNN-based speech enhancement, some papers have studied the possible advantages of training with a combination of different loss functions that target different attributes of the speech signal (e.g., quality and intelligibility) (Zhang et al., 2018; Braun and Tashev, 2021; Li and Yamagishi, 2021). Generally, the use of composite loss functions leads to improvements in terms of various objective metrics of interest when compared with the employment of single loss functions (Braun and Tashev, 2021). Based on the argued potentials of STGI as an intelligibility metric and the outstanding performance of the waveform-level loss function $\mathcal{L}_{\text{SI-SDR}}$ (Luo and Mesgarani, 2019; Kolbæk et al., 2020), we analyze their combination in this work.

Let $d^{\{y\}}_{\text{STGI}} \in [0, 1]$ be the STGI score of a training or validation noisy speech signal $\mathbf{y}$. Then, instead of using a constant loss function combination weight as is typically done in the speech enhancement literature (Zhang et al., 2018; Braun and Tashev, 2021; Li and Yamagishi, 2021), preliminary experiments revealed the benefit of employing an STGI score-dependent combination weight $\omega\left(d^{\{y\}}_{\text{STGI}}\right) = A \cdot \left(d^{\{y\}}_{\text{STGI}}\right)^{\gamma}$ (parameter values $A = 0.02$ and $\gamma = 10$ were chosen by means of a validation dataset) to define our composite loss function $\mathcal{L}_{\text{STGI+SI-SDR}} \in (-\infty, +\infty)$ as

$$\mathcal{L}_{\text{STGI+SI-SDR}} = \mathcal{L}_{\text{STGI}} + \omega\left(d^{\{y\}}_{\text{STGI}}\right)\mathcal{L}_{\text{SI-SDR}}. \tag{9}$$

Notice that the loss function combination weight $\omega\left(d^{\{y\}}_{\text{STGI}}\right) : [0, 1] \rightarrow [0, A]$ follows a power-law expression whose goal is to primarily focus on $\mathcal{L}_{\text{STGI}}$ when intelligibility (as predicted by STGI) is low, while increasing importance is given to $\mathcal{L}_{\text{SI-SDR}}$ when intelligibility is better. Among all the evaluated loss functions in this article, this composite loss function allows us to achieve the best performance in terms of a variety of intelligibility and quality metrics across a wide range of noisy conditions (see Section 4).

2.3. Training issues

STOI, ESTOI and STGI have the same built-in ideal voice activity detector that is employed to only take into account signal segments where speech is present when estimating speech intelligibility (Taal et al., 2011; Jensen and Taal, 2016; Edraki et al., 2021a). This ideal voice activity detection step is omitted by $\mathcal{L}_{\text{STOI}}$, $\mathcal{L}_{\text{ESTOI}}$, $\mathcal{L}_{\text{STGI}}$ and $\mathcal{L}_{\text{STGI+SI-SDR}}$. Instead, ideal voice activity detection is applied to the training and validation data prior to model training in order to discard speech-absent segments. As in Kolbæk et al. (2020), this is primarily done to avoid possible long silent regions for which processing cannot improve speech intelligibility, while easing the training procedure.

Before training, speech data are downsampled to 10 kHz when considering all the loss functions except $\mathcal{L}_{\text{PMSQE}}$ in order to be consistent with what is required by STOI, ESTOI and STGI (Taal et al., 2011; Jensen and Taal, 2016; Edraki et al., 2021a). To shape mini-batches for training – which have a size of 8 utterances –, every utterance is either truncated or zero-padded to have a duration of, approximately, 2 s.[2] In case of dealing with $\mathcal{L}_{\text{PMSQE}}$, speech data are downsampled to

[2] Recall that the FCNN employed in this study is fed with noisy speech signals comprised of $L = 19{,}712$ samples.

Table 1
Learning rates employed for the different loss functions studied in this work.

Loss function	$\mathcal{L}$ PMSQE	$\mathcal{L}$ SI-SDR	$\mathcal{L}$ SI-SDR+PE	$\mathcal{L}$ STOI	$\mathcal{L}$ ESTOI	$\mathcal{L}$ STGI	$\mathcal{L}$ STGI+SI-SDR
Learning rate	$5 \cdot 10^{-4}$	$5 \cdot 10^{-4}$	$5 \cdot 10^{-4}$	10^{-4}	10^{-4}	10^{-4}	10^{-4}

8 kHz, which involves working with, approximately, 2.5-second long speech segments instead.

The architecture of Fig. 1 is discriminatively trained by using Adam (Kingma and Ba, 2015) as an optimizer with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. As in Kolbæk et al. (2020), we use loss function-dependent learning rates to deal with the different nature of the different loss functions and their gradients. On the one hand, for $\mathcal{L}_{\text{PMSQE}}$, $\mathcal{L}_{\text{SI-SDR}}$, $\mathcal{L}_{\text{STOI}}$ and $\mathcal{L}_{\text{ESTOI}}$ we make use of the optimal learning rates found in Kolbæk et al. (2020). On the other hand, for $\mathcal{L}_{\text{SI-SDR+PE}}$, $\mathcal{L}_{\text{STGI}}$ and $\mathcal{L}_{\text{STGI+SI-SDR}}$, additional experiments were conducted at an early stage to choose proper learning rate values. Table 1 shows the loss function-dependent learning rate values employed in this work. During training, we also use a learning rate schedule halving the learning rate when the validation loss does not decrease for 2 epochs. Model training is performed for a maximum of 200 epochs and early-stopping (Gershenfeld, 1988) monitoring the validation loss with a patience of 5 epochs is used for regularization purposes.

Speech enhancement system implementation has been done by means of Keras (Chollet et al., 2015) working as an abstraction layer of TensorFlow (Abadi et al., 2015). Our TensorFlow implementation of the STGI loss function $\mathcal{L}_{\text{STGI}}$ has been made publicly available.[3]

3. Speech and noise databases

In this section, we present the speech databases and noise signals employed to generate synthetic noisy speech that we use for experimental purposes.

3.1. Noise signals

To create synthetic noisy speech data, we consider the same 6 types of noise as in Kolbæk et al. (2020): stationary speech-shaped noise (SSN), non-stationary babble, café, street, pedestrian street and bus. While the café, street, pedestrian street and bus noise types are from the CHiME-3 dataset (Barker et al., 2015, 2017), SSN and babble are synthetically generated as follows. First, SSN is obtained by filtering Gaussian white noise by means of a twelfth-order all-pole filter whose coefficients are derived from linear predictive coding analysis of the concatenation of 100 utterances randomly picked from the TIMIT dataset (Garofolo et al., 1993b). Second, babble noise, originating from 6 different speakers of both genders, is generated by mixing utterances randomly chosen from the TIMIT dataset as well.

While all of these noise types are seen at test time, pedestrian street and bus noises are excluded from training and validation sets. This is done in order to assess the generalization ability of speech enhancement models to types of noise not seen during the training phase. Finally, note that noise realizations do not overlap across training, validation and test sets.

3.2. Speech signals: The Wall Street Journal corpus

The Wall Street Journal (WSJ0) speech corpus (Garofolo et al., 1993a) is used here in a similar fashion as in Kolbæk et al. (2020) to mainly carry out objective performance evaluations in Section 4. Specifically, we create training, validation and test sets by randomly picking, following a sampling-with-replacement scheme, clean speech utterances from the WSJ0 corpus as follows:

- *Training set*: This set is shaped by 60,000 utterances drawn from a subset of the `si_tr_s` part of WSJ0. This subset is made up of 11,613 utterances that were recorded from 44 male and 47 female speakers.
- *Validation set*: This set is formed by 6,000 utterances selected from another subset of the `si_tr_s` part of WSJ0. This subset is comprised of 1,163 utterances that were recorded from 5 male and 5 female speakers.
- *Test set*: This set is generated by picking a total of 2,000 utterances from the `si_et_05` and `si_dt_05` parts of WSJ0. These parts are composed of 1,857 utterances that were recorded from 10 male and 6 female speakers.

First, note that speakers do not overlap across training, validation and test sets. Second, while each clean speech utterance of WSJ0 is selected more than once on average, each utterance in the above training, validation and test sets becomes a unique realization when mixed with acoustic noise in accordance with the following procedure.[4] Given a clean speech utterance $\mathbf{x}$, a noise segment $\tilde{\mathbf{n}}$ with the same length as $\mathbf{x}$ is randomly cut out of a randomly selected noise signal. Furthermore, to ensure a certain SNR level, $\tilde{\mathbf{n}}$ is scaled in accordance with the active speech level of $\mathbf{x}$ in line with the ITU P.56 recommendation (ITU-T, 2011). Finally, a unique noisy signal, $\mathbf{y}$, is obtained as $\mathbf{y} = \mathbf{x} + \mathbf{n}$, where $\mathbf{n}$ represents a version of $\tilde{\mathbf{n}}$, scaled to achieve a particular SNR.

By following the above procedure, the training and validation sets are contaminated with the noises SSN, babble, café and street, where the SNR level is uniformly sampled from $[-10, 10]$ dB. Differently from the training and validation sets, the test set considers the 6 types of noise SSN, babble, café, street, pedestrian street and bus, as well as the set of SNR levels $\{-10, -5, 0, 5, 10, 15, 20\}$ dB. Let us define a noisy condition as a particular combination of a noise type and an SNR level. Then, the 2,000 test clean speech utterances are contaminated a total of 6 noises $\times$ 7 SNRs $= 42$ times by considering every possible noisy condition. This implies a total of $2,000 \times 42 = 84,000$ noisy test utterances. In addition, it is worth noticing that test set speech-absent segments are not removed prior to speech enhancement processing in order to simulate a real use case, where the location of speech-absent segments is unknown.

3.3. Speech signals: The Akustiske Databaser for Dansk

The Akustiske Databaser for Dansk (ADFD) (Nasjonalbiblioteket, 2011) is a speech corpus in Danish that is employed in this work to train noisy condition- and language-matched speech enhancement models[5] to be used primarily to conduct a subjective intelligibility test, in Danish, in Section 5. In this case, we only create training and validation sets, since the associated test set is based on the (Danish) Dantale II dataset (Hansen and Ludvigsen, 2001; Wagener et al., 2003). The reason behind this arrangement is that the Dantale II dataset, which is used to actually generate intelligibility test stimuli, is too small to form training and validation sets in Danish, so ADFD is used as a support.

Similarly to the WSJ0 case, training and validation sets are created by randomly choosing, again following a sampling-with-replacement scheme, clean speech utterances from the ADFD database as follows:

[3] http://ilopez.es.mialias.net/codes/STGI.zip

[4] That being said, keep in mind that clean speech utterances do not overlap across training, validation and test sets.

[5] In this work, a noisy condition- and language-matched speech enhancement model is a model that is trained and tested, exclusively, on speech data from a specific noisy condition (i.e., combination of noise type and SNR level) and language.

- *Training set*: This set is formed by 60,000 utterances drawn from the original ADFD training set. The latter is composed of 24,960 utterances that were recorded from 40 male and 40 female speakers.
- *Validation set*: This set consists of 6,000 utterances chosen from the original ADFD validation set. The latter is made up of 1,000 utterances that were recorded from 10 speakers.

In the context of the ADFD database, we consider the 2 types of noise SSN and café, as well as the SNR levels $\{-10, -5\}$ dB, which define the 2 noises $\times$ 2 SNRs = 4 noisy conditions to which the intelligibility test subjects are exposed (see Section 5). Hence, to train noisy condition- and language-matched speech enhancement models, 4 different noisy condition-dependent training and validation sets are created from the above clean speech training and validation sets by following a speech and noise mixing procedure similar to that described in the previous subsection.

3.4. Speech signals: The Dantale II dataset

The Dantale II dataset (Hansen and Ludvigsen, 2001; Wagener et al., 2003) is a speech corpus, in Danish, that we employ to generate intelligibility test stimuli (see Section 5). This corpus is formed by a total of 16 lists of which we consider 15 comprising 10 sentences each, i.e., 150 utterances. All the utterances were recorded by a *single* female native Danish speaker in clean conditions. Moreover, every utterance consists of a sequence of five mono- and bi-syllabic words following the syntactical structure `name + verb + numeral + adjective + object`. To generate each of these five-word sentences, given a word class, a particular word was randomly chosen from a set of 10 different words. While all the utterances in the Dantale II dataset are syntactically correct, they may make no or little sense. The resulting low word-context predictability helps to prevent the intelligibility test subjects from guessing words based on previous ones (Knudsen, 2021).

Five out of the fifteen lists of the Dantale II dataset that we consider (i.e., 50 utterances) are reserved to fine tune the noisy condition- and language-matched speech enhancement models trained by means of ADFD. The goal of this is to construct noisy condition-, language- and *speaker*-matched models to conduct a subjective intelligibility test in Section 5 in an idealized situation to measure an upper bound for speech intelligibility improvement. Specifically, by again following a sampling-with-replacement scheme, we create 200- and 100-utterance fine tuning training and validation sets from, respectively, 4/15 and 1/15 lists. Similarly to the ADFD case considering SSN and café noises and SNRs $\{-10, -5\}$ dB, 4 different noisy condition-dependent fine tuning training and validations sets are built from the above clean speech fine tuning training and validation sets. Then, the remaining 10/15 lists (i.e., 100 utterances) are saved to produce intelligibility test stimuli by adding noise, as usual, to the clean speech utterances and processing the result by means of various speech enhancement models (see Section 5 for further details).

4. Objective performance evaluation

We evaluate the speech enhancement systems trained with the different loss functions presented in Section 2.2 when WSJ0 is employed for model training and testing as outlined in Section 3.2. This evaluation is performed in terms of various objective intelligibility (STOI (Taal et al., 2011), ESTOI (Jensen and Taal, 2016) and STGI (Edraki et al., 2021a)), waveform-level (SI-SDR (Roux et al., 2019) and SDR (Févotte et al., 2005)) and quality (PESQ (Rix et al., 2001; ITU-T, 2003)) metrics. Tables 2 and 3 report the corresponding results, which are broken down by SNR, averaged across the types of noise *seen* – SSN, babble, café and street – and *not seen* – pedestrian street and bus – during the training phase, respectively. As a reference, metric scores of the original noisy (i.e., unprocessed) speech signals are also shown. Moreover, given an input SNR value and a metric, best results are marked in bold.

From these tables, we can see that, despite PMSQE being an approximation of PESQ (Martin-Doñas et al., 2018), $\mathcal{L}_{\text{PMSQE}}$ clearly yields the worst PESQ results among all the evaluated loss functions, which is consistent with previous findings in Kolbæk et al. (2020). When disregarding $\mathcal{L}_{\text{PMSQE}}$, it can be observed that, particularly at lower SNRs, every loss function tends to achieve, among the rest of them, the best performance in terms of the metric the corresponding loss function is based on. This behavior is obviously expected and serves as a sanity check that networks are trained successfully. Besides, note that the integration of speech pre-emphasis into $\mathcal{L}_{\text{SI-SDR}}$ does not help to improve $\mathcal{L}_{\text{SI-SDR}}$ particularly in terms of objective intelligibility, despite that our preliminary observations indicated that considering pre-emphasis may improve subjective intelligibility. The latter hypothesis is more carefully examined in the next section.

Also from Tables 2 and 3, it is worth noticing the rather poor performance of systems trained with loss functions $\mathcal{L}_{\text{PMSQE}}$, $\mathcal{L}_{\text{STOI}}$, $\mathcal{L}_{\text{ESTOI}}$ and $\mathcal{L}_{\text{STGI}}$ in terms of a waveform-level metric like SI-SDR. The reason behind this behavior is related to the fact that these loss functions do not penalize deviations of the enhanced waveform with respect to the target waveform (since they essentially operate with short-time spectral magnitude information) (Kolbæk et al., 2020). In other words, speech signals enhanced by systems trained using $\mathcal{L}_{\text{PMSQE}}$, $\mathcal{L}_{\text{STOI}}$, $\mathcal{L}_{\text{ESTOI}}$ or $\mathcal{L}_{\text{STGI}}$ may exhibit small time shifts with respect to target speech signals, which is not relevant for intelligibility and quality of the enhanced speech despite time shifts are penalized by waveform-level measures like SI-SDR.

It is interesting to see from Tables 2 and 3 how incorporating SI-SDR in $\mathcal{L}_{\text{STGI}}$ as is done in the composite loss function $\mathcal{L}_{\text{STGI+SI-SDR}}$ allows us to compensate for the aforementioned waveform deviations resultant from the STGI loss function, as reflected by the high SI-SDR scores obtained by the system trained with $\mathcal{L}_{\text{STGI+SI-SDR}}$. In fact, particularly at higher SNRs, this composite loss function tends to achieve the best performance in terms of all the considered metrics. Specifically, $\mathcal{L}_{\text{STGI+SI-SDR}}$ produces the highest PESQ scores for all the evaluated noisy conditions except for babble at an input SNR of -10 dB. This is a clear example of a successful synergy: the combination of an intelligibility metric-based loss with a waveform-level loss brings higher speech quality scores than any of the two loss functions does independently. In conclusion, according to its outstanding objective performance across all the considered noisy conditions in terms of a variety of intelligibility, waveform-level and quality metrics, $\mathcal{L}_{\text{STGI+SI-SDR}}$ is a strong candidate when it comes to the training of a speech enhancement system to be deployed in real life.

5. Subjective versus objective intelligibility

Before deploying any speech enhancement system in real life, it is important to verify, if possible, the performance indicated by objective metrics through subjective tests with relevant end-users. In this work, we specifically focus on speech intelligibility and, consequently, we carry out a subjective intelligibility test. The test evaluates three different loss functions used for network training: $\mathcal{L}_{\text{SI-SDR}}$ (due to its competitive performance in both Kolbæk et al. (2020) and this paper), $\mathcal{L}_{\text{SI-SDR+PE}}$ (to assess the potential benefits of speech pre-emphasis) and $\mathcal{L}_{\text{STGI+SI-SDR}}$ (due to this loss function being the best objectively performing in this work). In addition, this test also measures, as a reference, the intelligibility of the original noisy/unprocessed speech.

The present subjective intelligibility test consists of two different parts: *Part A* and *Part B*. *Part A* (see Section 5.3) aims at evaluating the general-purpose speech enhancement models[6] of Section 4, which

[6] Recall that by general-purpose speech enhancement models we mean that they are trained (in this case, using the WSJ0 corpus) considering a variety of noisy conditions.

Table 2
Results, in terms of various metrics, from the evaluation on WSJ0 of the different loss functions studied in this work. Results are broken down by SNR and averaged across the types of noise *seen* – SSN, babble, café and street – during the training phase. Given an input SNR value and a metric, best results are marked in bold.

SNR (dB)	Metric	Noisy	Processed						
			$\mathcal{L}$ PMSQE	$\mathcal{L}$ SI-SDR	$\mathcal{L}$ SI-SDR+PE	$\mathcal{L}$ STOI	$\mathcal{L}$ ESTOI	$\mathcal{L}$ STGI	$\mathcal{L}$ STGI+SI-SDR
−10	STOI	0.53	0.64	0.69	0.67	**0.72**	0.69	0.69	0.70
	ESTOI	0.21	0.40	0.42	0.38	0.45	**0.47**	0.45	0.45
	STGI	0.56	0.74	0.76	0.76	0.76	0.76	**0.80**	**0.80**
	SI-SDR	−12.32	−23.06	**0.88**	−0.34	−6.74	−6.73	−21.27	0.54
	SDR	−10.64	−5.84	**2.85**	1.52	−0.23	−0.65	−0.15	2.49
	PESQ	1.31	1.30	**1.55**	1.51	1.47	1.37	1.48	**1.55**
−5	STOI	0.64	0.80	0.83	0.81	**0.85**	0.83	0.83	0.84
	ESTOI	0.33	0.61	0.63	0.59	0.65	**0.67**	0.65	0.66
	STGI	0.76	0.92	0.92	0.92	0.92	0.93	**0.94**	**0.94**
	SI-SDR	−7.31	−19.58	**6.23**	4.97	−3.00	−2.58	−18.87	6.06
	SDR	−6.61	−2.15	**7.57**	6.25	4.57	4.61	4.54	7.40
	PESQ	1.38	1.58	1.93	1.87	1.81	1.74	1.89	**2.03**
0	STOI	0.75	0.89	0.90	0.89	**0.91**	**0.91**	0.90	**0.91**
	ESTOI	0.48	0.74	0.76	0.73	0.78	**0.80**	0.77	0.79
	STGI	0.89	**0.97**	0.96	0.96	0.96	0.96	**0.97**	**0.97**
	SI-SDR	−2.31	−18.56	**10.14**	8.86	−1.31	−0.80	−18.10	10.10
	SDR	−1.99	−0.30	**11.18**	9.92	7.85	8.27	7.64	11.17
	PESQ	1.56	2.00	2.36	2.32	2.20	2.19	2.38	**2.55**
5	STOI	0.85	0.93	**0.94**	0.93	**0.94**	**0.94**	0.93	**0.94**
	ESTOI	0.63	0.82	0.85	0.82	0.85	**0.86**	0.84	**0.86**
	STGI	0.95	**0.99**	0.98	0.98	0.98	0.98	0.98	0.98
	SI-SDR	2.69	−18.17	13.33	11.93	−0.49	0.00	−17.95	**13.38**
	SDR	2.87	0.51	14.24	12.91	10.16	10.84	9.69	**14.29**
	PESQ	1.81	2.46	2.77	2.76	2.58	2.62	2.82	**2.98**
10	STOI	0.91	0.95	**0.96**	0.95	**0.96**	0.95	0.95	**0.96**
	ESTOI	0.76	0.87	0.89	0.87	0.89	**0.90**	0.88	**0.90**
	STGI	0.98	**0.99**	0.98	0.98	0.98	0.98	**0.99**	**0.99**
	SI-SDR	7.70	−17.96	16.14	14.46	−0.06	0.38	−17.77	**16.26**
	SDR	7.83	0.81	17.02	15.46	11.79	12.61	11.06	**17.12**
	PESQ	2.13	2.87	3.12	3.11	2.94	2.99	3.16	**3.31**
15	STOI	0.95	0.95	**0.97**	0.96	0.96	0.96	0.96	**0.97**
	ESTOI	0.86	0.89	0.92	0.90	0.92	0.92	0.90	**0.93**
	STGI	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
	SI-SDR	12.70	−17.84	18.58	16.55	0.14	0.55	−17.73	**18.79**
	SDR	12.81	0.92	19.57	17.62	12.80	13.71	11.92	**19.69**
	PESQ	2.53	3.17	3.41	3.39	3.29	3.32	3.43	**3.58**
20	STOI	0.97	0.96	0.97	0.96	0.97	0.97	0.96	**0.98**
	ESTOI	0.92	0.90	**0.94**	0.92	0.93	0.93	0.92	**0.94**
	STGI	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
	SI-SDR	17.70	−17.80	20.43	18.10	0.22	0.63	−17.63	**20.76**
	SDR	17.81	0.96	21.63	19.30	13.32	14.27	12.40	**21.78**
	PESQ	2.99	3.37	3.67	3.62	3.58	3.59	3.64	**3.79**

reflect a *practical scenario*. *Part B* (see Section 5.4) tests noisy condition-, language- and speaker-matched models. This is done to assess the magnitude of the potentially greater gains that the speech enhancement systems may achieve if we could operate in a *close-to-ideal scenario*, where the sources of acoustic variability are minimized. Recall that the models of Part B are trained on the ADFD speech corpus and fine tuned on 5 lists of the Dantale II dataset.

5.1. Intelligibility test data preparation

As introduced in Section 3.3, the present subjective intelligibility test considers 2 types of noise, SSN and café, as well as 2 input SNR values chosen such that the speech intelligibility of the unprocessed noisy signals is reduced: −10 dB and −5 dB. Given a noisy condition and one of the four processing types to be tested (i.e., noisy/unprocessed, $\mathcal{L}_{\text{SI-SDR}}$, $\mathcal{L}_{\text{SI-SDR+PE}}$ and $\mathcal{L}_{\text{STGI+SI-SDR}}$), we generate 6 different stimuli/speech signal realizations. Thus, each of the test parts (i.e., Part A and Part B) is comprised of a total of 2 noises × 2 SNRs×4 processing types×6 sentences = 96 stimuli. To generate these stimuli, for Part A and Part B independently, 96 clean speech signals are drawn (in a semi-random fashion to avoid the same utterance repeating) from the 10 lists (i.e., 100 utterances) of the Dantale II dataset reserved for this purpose. By repeating this procedure multiple times, we create 20 different realizations of this subjective intelligibility test and every test subject is randomly assigned one of them.

By following an analogous procedure, we also generate a set of 8 stimuli intended for a training session for the test subjects to become familiar with the intelligibility test aiming to reduce the inherent learning bias. This training session is purposely kept short to not overwhelm the subjects with an unnecessarily long test that might negatively affect the validity of the results.

5.2. Intelligibility test procedure

The test was conducted in a silent audiometry room with sound-isolating walls, floor and ceiling. One at a time, each test subject was sitting at a small table in the middle of this room wearing Beyerdynamic DT 990 Pro headphones connected to a laptop Lenovo ThinkPad T480s. Running on this laptop, the intelligibility test interface was controlled by the subjects. The test interface (Heidemann Andersen, 2017) presented all candidate words for each of the 5 word classes on the laptop screen (facilitating a closed-vocabulary matrix test). The subject initiated stimulus playback by a mouse click (stimulus was played *only once*). Next, the subject chose the words heard from the matrix of candidate words using mouse clicks. This procedure was repeated

Table 3
Results, in terms of various metrics, from the evaluation on WSJ0 of the different loss functions studied in this work. Results are broken down by SNR and averaged across the types of noise *not seen* – pedestrian street and bus – during the training phase.

SNR (dB)	Metric	Noisy	Processed						
			$\mathcal{L}$ PMSQE	$\mathcal{L}$ SI-SDR	$\mathcal{L}$ SI-SDR+PE	$\mathcal{L}$ STOI	$\mathcal{L}$ ESTOI	$\mathcal{L}$ STGI	$\mathcal{L}$ STGI+SI-SDR
−10	STOI	0.62	0.74	0.77	0.76	**0.80**	0.77	0.77	0.78
	ESTOI	0.29	0.52	0.53	0.51	0.56	**0.59**	0.57	0.56
	STGI	0.74	0.84	0.85	0.87	0.86	0.86	**0.90**	0.89
	SI-SDR	−12.33	−21.44	**3.68**	2.44	−5.33	−5.11	−20.22	3.38
	SDR	−10.65	−3.36	**5.36**	3.97	2.32	2.03	2.28	4.97
	PESQ	1.37	1.62	1.85	1.82	1.78	1.69	1.82	**1.91**
−5	STOI	0.70	0.85	0.87	0.86	**0.88**	0.87	0.87	**0.88**
	ESTOI	0.39	0.67	0.69	0.66	0.71	**0.73**	0.71	0.72
	STGI	0.87	0.95	0.95	0.96	0.95	0.95	**0.97**	0.96
	SI-SDR	−7.32	−19.16	**8.27**	6.86	−2.34	−1.91	−18.58	8.11
	SDR	−6.62	−0.98	**9.54**	8.01	6.37	6.49	6.18	9.35
	PESQ	1.51	1.96	2.26	2.21	2.14	2.08	2.22	**2.36**
0	STOI	0.78	0.91	0.92	0.91	**0.93**	0.92	0.92	**0.93**
	ESTOI	0.51	0.77	0.80	0.77	0.81	**0.82**	0.80	**0.82**
	STGI	0.94	0.98	0.98	0.98	0.98	0.98	**0.99**	**0.99**
	SI-SDR	−2.31	−18.28	**11.79**	10.29	−0.95	−0.44	−18.07	**11.79**
	SDR	−2.00	0.24	**12.91**	11.29	9.32	9.78	8.93	12.88
	PESQ	1.74	2.39	2.69	2.62	2.54	2.53	2.67	**2.83**
5	STOI	0.86	0.94	**0.95**	0.94	**0.95**	**0.95**	0.94	**0.95**
	ESTOI	0.64	0.84	0.86	0.84	0.87	**0.88**	0.86	**0.88**
	STGI	0.97	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
	SI-SDR	2.69	−17.87	14.74	13.21	−0.28	0.18	−17.81	**14.84**
	SDR	2.87	0.75	15.80	14.18	11.31	11.99	10.73	**15.85**
	PESQ	2.05	2.81	3.08	3.01	2.93	2.93	3.06	**3.21**
10	STOI	0.92	0.95	0.96	0.95	0.96	0.96	0.96	**0.97**
	ESTOI	0.76	0.88	0.90	0.88	**0.91**	**0.91**	0.89	**0.91**
	STGI	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
	SI-SDR	7.69	−17.72	17.30	15.68	0.05	0.46	−17.65	**17.52**
	SDR	7.83	0.92	18.41	16.70	12.57	13.39	11.83	**18.52**
	PESQ	2.42	3.16	3.41	3.35	3.27	3.27	3.38	**3.52**
15	STOI	0.95	0.96	**0.98**	0.97	0.97	0.97	0.96	**0.98**
	ESTOI	0.86	0.90	**0.93**	0.91	**0.93**	**0.93**	0.91	**0.93**
	STGI	0.99	**1.00**	**1.00**	0.99	0.99	0.99	**1.00**	**1.00**
	SI-SDR	12.69	−17.70	19.42	17.56	0.19	0.58	−17.55	**19.78**
	SDR	12.81	0.98	20.70	18.72	13.25	14.16	12.43	**20.87**
	PESQ	2.85	3.39	3.67	3.64	3.57	3.55	3.63	**3.77**
20	STOI	**0.98**	0.96	**0.98**	0.97	**0.98**	0.97	0.97	**0.98**
	ESTOI	0.92	0.91	**0.95**	0.93	0.94	0.94	0.93	**0.95**
	STGI	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.99	**1.00**	**1.00**
	SI-SDR	17.69	−17.72	20.94	18.77	0.24	0.64	−17.52	**21.42**
	SDR	17.80	0.99	22.41	20.06	13.56	14.50	12.72	**22.62**
	PESQ	3.29	3.50	3.88	3.86	3.80	3.75	3.81	**3.96**

until the test ended. Keep in mind that the stimuli were presented in a random order.

A total of 26 native Danish speakers, 19 males and 7 females, were recruited as test subjects. Their age ranged from 18 to 30. All of them reported no hearing issue except for one indicating a very slight bilateral tinnitus. The volume at which the stimuli were played was initially adjusted and kept fixed across all subjects since all of them found it comfortable. After the training session, each subject went through Part A or Part B (randomly chosen) before taking a break to prevent listening fatigue that was then followed by the other part (i.e., Part B or Part A). On average, the training session (8 stimuli) took to the subjects 2–3 min, the first part (96 stimuli), 23 min, and, the second part after the break (96 stimuli), 20 min. Test subjects were paid for their participation.

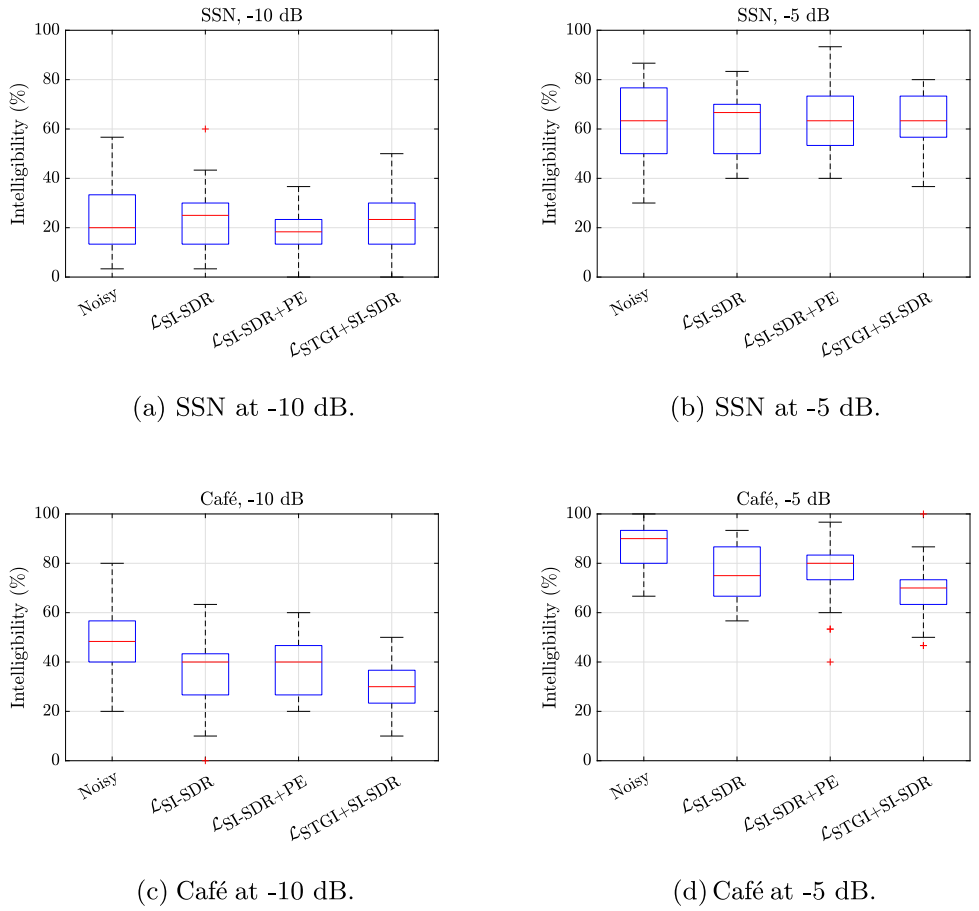
Intelligibility, defined as the percentage of words correctly identified (Allen, 2005), is used as a subjective metric. For a statistical significance test of the subjective intelligibility test results, we consider the Kruskal–Wallis H test (Kruskal and Wallis, 1952), which is a non-parametric version of the classical parametric one-way analysis of variance (ANOVA) (Fisher, 1918). The main reason for using the Kruskal–Wallis H test instead of ANOVA is that the assumption that the sample populations have normal distributions as required by ANOVA does not hold in our case according to the Kolmogorov–Smirnov test (Massey, 1951).

Table 4
p-values corresponding to Part A (use of general-purpose models) of the subjective intelligibility test. p-values below a significance level of 0.05, which indicate statistically significant intelligibility differences, are marked in bold.

p-values	SSN		Café	
Comparison	−10 dB	−5 dB	−10 dB	−5 dB
Noisy - $\mathcal{L}$ SI-SDR	0.990	0.995	0.067	**0.002**
Noisy - $\mathcal{L}$ SI-SDR+PE	0.434	0.998	0.130	0.052
Noisy - $\mathcal{L}$ STGI+SI-SDR	0.995	0.999	**≈0.000**	**≈0.000**
$\mathcal{L}$ SI-SDR - $\mathcal{L}$ SI-SDR+PE	0.271	0.975	0.992	0.738
$\mathcal{L}$ SI-SDR - $\mathcal{L}$ STGI+SI-SDR	0.952	0.992	0.221	0.455
$\mathcal{L}$ SI-SDR+PE - $\mathcal{L}$ STGI+SI-SDR	0.576	0.999	0.125	0.061

5.3. Intelligibility results for Part A: General-purpose models

In this subsection, we present the results from Part A of the subjective intelligibility test. Recall that Part A aims at evaluating the general-purpose enhancement models of Section 4 that are trained on the WSJ0 corpus. Fig. 2 displays intelligibility scores in the form of box plots, where horizontal red lines indicate median (across test subjects) intelligibility. Complementing this figure, Table 4 shows the corresponding p-values obtained by the Kruskal–Wallis H test. In this



(a) SSN at -10 dB. (b) SSN at -5 dB.

(c) Café at -10 dB. (d) Café at -5 dB.

Fig. 2. Box plots corresponding to Part A (use of general-purpose models) of the subjective intelligibility test. Red lines mark median (across test subjects) intelligibility, whereas outliers are indicated by red "plus" markers. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

table, p-values below a significance level of 0.05 are considered to indicate statistically significant intelligibility differences and are marked in bold. Jointly considering Fig. 2 and Table 4, we can see that no speech enhancement processing (using networks optimized towards $\mathcal{L}_{\text{SI-SDR}}$, $\mathcal{L}_{\text{SI-SDR+PE}}$ and $\mathcal{L}_{\text{STGI+SI-SDR}}$) is able to improve the intelligibility of the unprocessed speech (Noisy). This is in contrast to what we reported in Tables 2 and 3, where $\mathcal{L}_{\text{SI-SDR}}$, $\mathcal{L}_{\text{SI-SDR+PE}}$ and $\mathcal{L}_{\text{STGI+SI-SDR}}$ all consistently improve the estimated intelligibility of original noisy speech at the SNRs −10 dB and −5 dB according to the intelligibility metrics STOI, ESTOI and STGI. Furthermore, again jointly considering Fig. 2 and Table 4, we can observe that $\mathcal{L}_{\text{STGI+SI-SDR}}$ worsens, in a statistically significant manner, the intelligibility of the unprocessed speech for noise type café. This is also the case of $\mathcal{L}_{\text{SI-SDR}}$ under the noisy condition café at −5 dB. Besides, note that there are no statistically significant intelligibility differences among $\mathcal{L}_{\text{SI-SDR}}$, $\mathcal{L}_{\text{SI-SDR+PE}}$ and $\mathcal{L}_{\text{STGI+SI-SDR}}$.

In contrast to Section 4, where both the training and test speech data come from the same corpus (i.e., WSJ0), for Part A of the subjective intelligibility test, there is a substantial acoustic mismatch between the training and test speech data, which come from, respectively, the WSJ0 and Dantale II datasets. Therefore, one may argue that this acoustic mismatch could be the cause of the poor subjective intelligibility results presented in this subsection. However, Tables 5 and 6, reporting objective metric scores for the stimuli from Part A, contradict this hypothesis. These tables indicate, in line with Tables 2 and 3, that all $\mathcal{L}_{\text{SI-SDR}}$, $\mathcal{L}_{\text{SI-SDR+PE}}$ and $\mathcal{L}_{\text{STGI+SI-SDR}}$ consistently enhance the intelligibility of the unprocessed speech as measured by the intelligibility metrics STOI, ESTOI and STGI.

To better assess this discrepancy between subjective intelligibility and what is expressed by intelligibility metrics, Fig. 3 depicts scatter plots representing median intelligibility from Part A as a function of the corresponding intelligibility scores produced by STOI, ESTOI and STGI. In these (per noisy condition) plots, each data point corresponds to a different processing type. Furthermore, Kendall's τ (Kendall, 1938) coefficients calculated from the data in each scatter plot are also shown. Note that $\tau \in [-1, 1]$ measures the ordinal association between two quantities and tends to 1 (−1) when they have a positive (negative) monotonic relationship, while $\tau = 0$ when they are statistically independent. Obviously, the key requirement of an intelligibility metric is that it has a strong positive monotonic relationship with subjective intelligibility (Jensen and Taal, 2014). However, from Fig. 3,

Table 5
Evaluation, in terms of various metrics, of the stimuli from Part A (use of general-purpose models) of the subjective intelligibility test. Results are broken down by SNR and correspond to the noise type **SSN**. Given an input SNR value and a metric, best results are marked in bold.

SNR (dB)	Metric	Noisy	Processed		
			$\mathcal{L}$ SI-SDR	$\mathcal{L}$ SI-SDR+PE	$\mathcal{L}$ STGI+SI-SDR
−10	STOI	0.50	0.59	0.53	**0.60**
	ESTOI	0.16	0.34	0.27	**0.37**
	STGI	0.23	0.52	0.44	**0.60**
	SI-SDR	−12.78	**2.58**	0.14	2.25
	SDR	−11.94	**4.59**	1.97	4.27
	PESQ	**1.33**	1.19	1.18	1.20
−5	STOI	0.57	0.72	0.71	**0.76**
	ESTOI	0.28	0.50	0.49	**0.56**
	STGI	0.46	0.74	0.73	**0.82**
	SI-SDR	−7.79	5.19	3.81	**5.87**
	SDR	−7.48	7.34	5.65	**8.08**
	PESQ	1.36	1.42	1.40	**1.47**

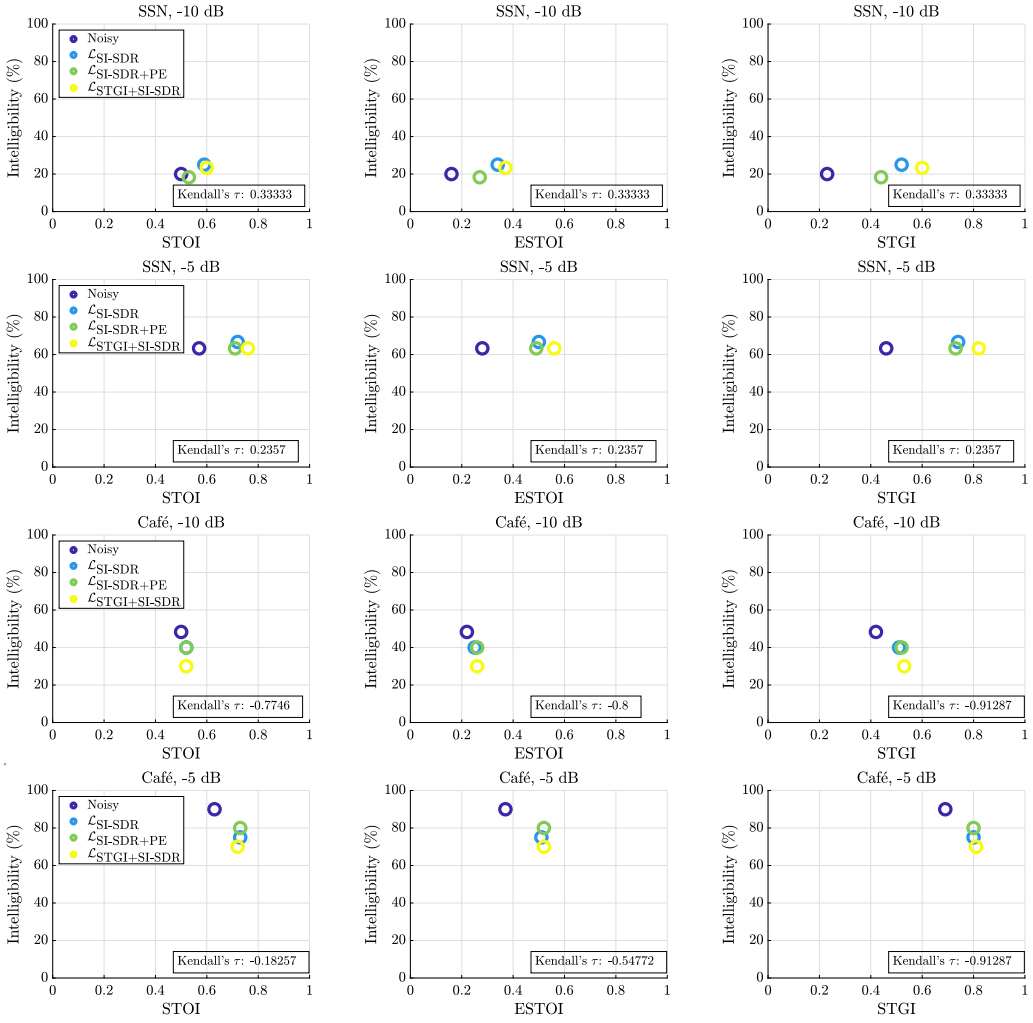


Fig. 3. Scatter plots representing median intelligibility from Part A (use of general-purpose models) of the subjective intelligibility test as a function of the intelligibility scores given by STOI, ESTOI and STGI. Each data point corresponds to a different processing type. Kendall's τ (Kendall, 1938) coefficients calculated from the data in each scatter plot are also shown.

we can see that this requirement is met by none of the employed intelligibility metrics. In fact, for the noise type café, STOI, ESTOI and STGI all exhibit a negative monotonic relationship with subjective intelligibility.

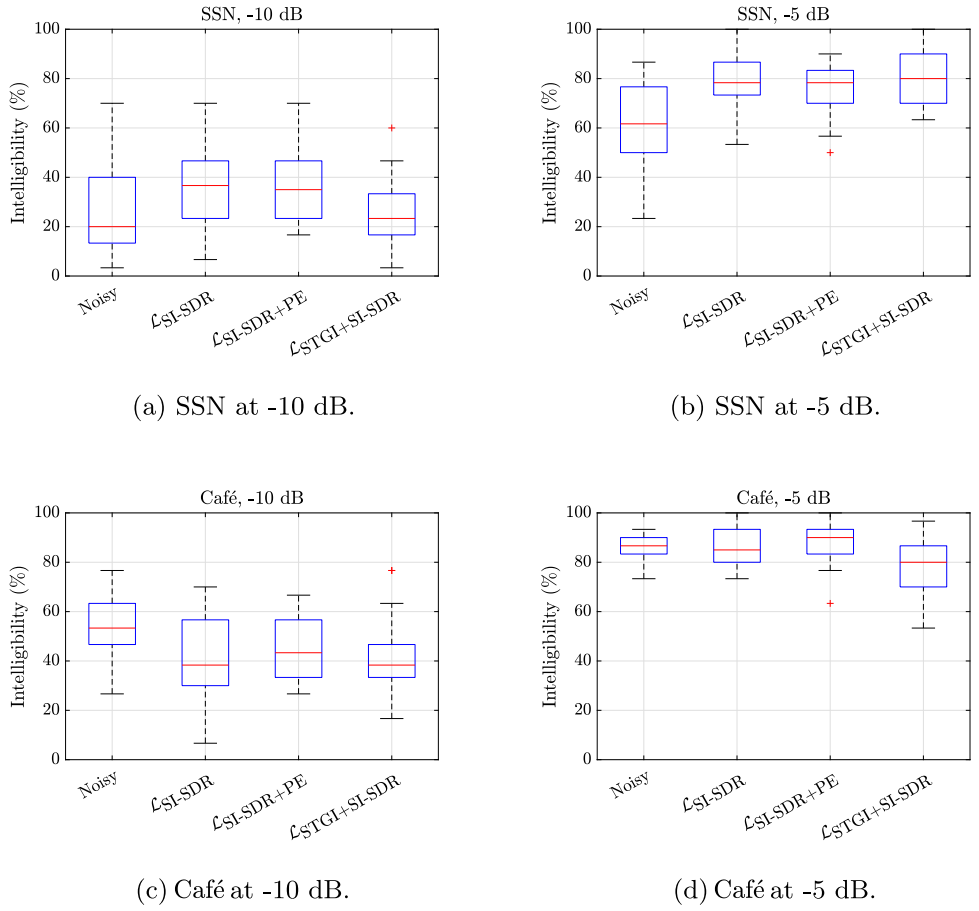
5.4. Intelligibility results for Part B: Noisy condition-, language- and speaker-matched models

In the previous subsection, we have mentioned that, for Part A of the subjective intelligibility test, there is a significant acoustic mismatch between the training and test speech data that may negatively affect subjective intelligibility. Differently from Part A, Part B evaluates noisy condition-, language- and speaker-matched speech enhancement models with the aim of minimizing the sources of acoustic variability. In this way, we can estimate the extent of what we may gain in terms of intelligibility in case we could operate in a close-to-ideal scenario. Similarly to Section 5.3, Fig. 4 shows Part B results in the form of box plots, while Table 7 gathers the associated p-values yielded by the Kruskal–Wallis H test. First of all, comparing the box plots in Figs. 4 and 2, it can be seen at a glance that the results for Noisy are slightly different. This is simply because test subjects listened to different unprocessed speech stimuli in Parts A and B. That being said, from Table 8, which reports p-values corresponding to the comparison between Part A and

Table 6
Evaluation, in terms of various metrics, of the stimuli from Part A (use of general-purpose models) of the subjective intelligibility test. Results are broken down by SNR and correspond to the noise type **café**. Given an input SNR value and a metric, best results are marked in bold.

SNR (dB)	Metric	Noisy	Processed		
			$\mathcal{L}$ SI-SDR	$\mathcal{L}$ SI-SDR+PE	$\mathcal{L}$ STGI+SI-SDR
−10	STOI	0.50	**0.52**	**0.52**	**0.52**
	ESTOI	0.22	0.25	**0.26**	**0.26**
	STGI	0.42	0.51	0.52	**0.53**
	SI-SDR	−12.75	−6.69	**−6.06**	−7.26
	SDR	−12.07	−2.92	−3.11	**−2.74**
	PESQ	**1.23**	1.15	1.21	1.16
−5	STOI	0.63	**0.73**	**0.73**	0.72
	ESTOI	0.37	0.51	**0.52**	**0.52**
	STGI	0.69	0.80	0.80	**0.81**
	SI-SDR	−7.74	**1.86**	1.32	1.78
	SDR	−7.46	**4.70**	3.72	4.48
	PESQ	1.22	1.39	**1.43**	1.36

Part B, it is important to note that there are no statistically significant intelligibility differences for Noisy. On the contrary, noisy condition-, language- and speaker-matched models generally yield statistically significant intelligibility improvements with respect to general-purpose



(a) SSN at -10 dB. (b) SSN at -5 dB.

(c) Café at -10 dB. (d) Café at -5 dB.

Fig. 4. Box plots corresponding to Part B (use of noisy condition-, language- and speaker-matched models) of the subjective intelligibility test. Red lines mark median (across test subjects) intelligibility, whereas outliers are indicated by red "plus" markers. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

Table 7
p-values corresponding to Part B (use of noisy condition-, language- and speaker-matched models) of the subjective intelligibility test. *p*-values below a significance level of 0.05, which indicate statistically significant intelligibility differences, are marked in bold.

p-values	SSN		Café	
Comparison	−10 dB	−5 dB	−10 dB	−5 dB
Noisy - $\mathcal{L}$ SI-SDR	0.089	**0.001**	0.092	0.999
Noisy - $\mathcal{L}$ SI-SDR+PE	0.052	**0.030**	0.142	0.807
Noisy - $\mathcal{L}$ STGI+SI-SDR	0.998	**≈0.000**	**0.012**	0.051
$\mathcal{L}$ SI-SDR - $\mathcal{L}$ SI-SDR+PE	0.996	0.833	0.997	0.737
$\mathcal{L}$ SI-SDR - $\mathcal{L}$ STGI+SI-SDR	0.062	0.991	0.886	0.070
$\mathcal{L}$ SI-SDR+PE - $\mathcal{L}$ STGI+SI-SDR	**0.034**	0.669	0.795	**0.003**

models according to this same table and having in mind Figs. 4 and 2. Notice that *p*-values in Table 8 were estimated by means of the Mann–Whitney *U* test, also known as Wilcoxon rank-sum test, which is equivalent to the Kruskal–Wallis *H* test dealing with two sample populations (Wilcoxon, 1945).

Differently from Part A, Part B speech enhancement models are able to improve, in a statistically significant manner, the intelligibility of the unprocessed speech, although this is *only* under the noisy condition SSN at −5 dB (see Fig. 4 and Table 7). As for Part A, $\mathcal{L}_{\text{STGI+SI-SDR}}$ also harms, in a statistically significant manner, the intelligibility of the unprocessed speech under the noisy condition café at −10 dB. In addition, there are no significant differences among $\mathcal{L}_{\text{SI-SDR}}$, $\mathcal{L}_{\text{SI-SDR+PE}}$ and $\mathcal{L}_{\text{STGI+SI-SDR}}$ except for $\mathcal{L}_{\text{SI-SDR+PE}}$ outperforming $\mathcal{L}_{\text{STGI+SI-SDR}}$ under the noisy conditions SSN at −10 dB and café at −5 dB.

Table 8
p-values corresponding to the comparison between Part A and Part B of the subjective intelligibility test. *p*-values below a significance level of 0.05, which indicate statistically significant intelligibility differences, are marked in bold.

p-values	SSN		Café	
Processing	−10 dB	−5 dB	−10 dB	−5 dB
Noisy	0.733	1.000	0.097	0.637
$\mathcal{L}$ SI-SDR	**0.003**	**≈0.000**	0.192	**≈0.000**
$\mathcal{L}$ SI-SDR+PE	**≈0.000**	**0.002**	0.084	**≈0.000**
$\mathcal{L}$ STGI+SI-SDR	0.490	**≈0.000**	**0.003**	**0.015**

Tables 9 and 10 show objective metric scores for the stimuli from Part B. Once again, these objective scores are highly aligned with those previously reported in Tables 2 and 3, as well as in Tables 5 and 6. This means that, also for Part B, $\mathcal{L}_{\text{SI-SDR}}$, $\mathcal{L}_{\text{SI-SDR+PE}}$ and $\mathcal{L}_{\text{STGI+SI-SDR}}$ invariably improve the estimated intelligibility of the unprocessed speech in terms of STOI, ESTOI and STGI for all the noisy conditions evaluated. Parallel to Fig. 3, Fig. 5 shows scatter plots, which also include Kendall's τ coefficients, representing median intelligibility from Part B as a function of these objective intelligibility scores. From this figure, we can assess how the substantial reduction of the acoustic mismatch between training and testing generally improves the positivity of the association between subjective and objective intelligibility, particularly under the noisy condition SSN at −5 dB. Even so, especially for the type of noise café with some negative Kendall's τ coefficients, all STOI, ESTOI and STGI still exhibit, as for Part A, a low correlation with subjective intelligibility.

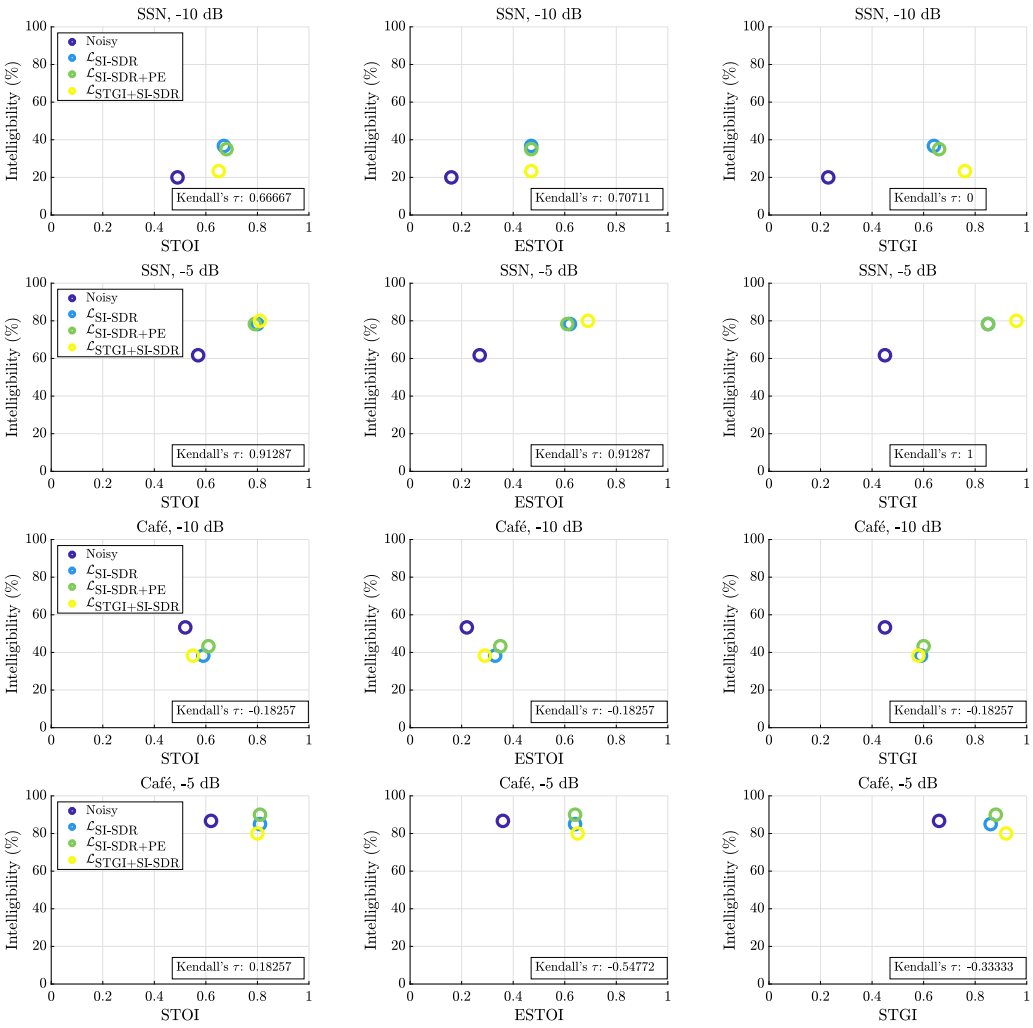


Fig. 5. Scatter plots representing median intelligibility from Part B (use of noisy condition-, language- and speaker-matched models) of the subjective intelligibility test as a function of the intelligibility scores given by STOI, ESTOI and STGI. Each data point corresponds to a different processing type. Kendall's τ (Kendall, 1938) coefficients calculated from the data in each scatter plot are also shown.

Table 9
Evaluation, in terms of various metrics, of the stimuli from Part B (use of noisy condition-, language- and speaker-matched models) of the subjective intelligibility test. Results are broken down by SNR and correspond to the noise type **SSN**. Given an input SNR value and a metric, best results are marked in bold.

SNR (dB)	Metric	Noisy	Processed		
			$\mathcal{L}$ SI-SDR	$\mathcal{L}$ SI-SDR+PE	$\mathcal{L}$ STGI+SI-SDR
−10	STOI	0.49	0.67	**0.68**	0.65
	ESTOI	0.16	**0.47**	**0.47**	**0.47**
	STGI	0.23	0.64	0.66	**0.76**
	SI-SDR	−12.76	**4.97**	4.77	1.45
	SDR	−11.91	**7.42**	7.29	4.18
	PESQ	**1.45**	1.38	1.40	1.28
−5	STOI	0.57	0.80	0.79	**0.81**
	ESTOI	0.27	0.62	0.61	**0.69**
	STGI	0.45	0.85	0.85	**0.96**
	SI-SDR	−7.77	**7.81**	7.78	4.14
	SDR	−7.46	**10.60**	10.54	7.48
	PESQ	1.30	1.65	1.61	**1.76**

Table 10
Evaluation, in terms of various metrics, of the stimuli from Part B (use of noisy condition-, language- and speaker-matched models) of the subjective intelligibility test. Results are broken down by SNR and correspond to the noise type **café**. Given an input SNR value and a metric, best results are marked in bold.

SNR (dB)	Metric	Noisy	Processed		
			$\mathcal{L}$ SI-SDR	$\mathcal{L}$ SI-SDR+PE	$\mathcal{L}$ STGI+SI-SDR
−10	STOI	0.52	0.59	**0.61**	0.55
	ESTOI	0.22	0.33	**0.35**	0.29
	STGI	0.45	0.59	**0.60**	0.58
	SI-SDR	−12.76	0.63	**1.19**	−6.44
	SDR	−12.04	2.88	**3.37**	−3.40
	PESQ	1.23	1.24	**1.27**	1.15
−5	STOI	0.62	**0.81**	**0.81**	0.80
	ESTOI	0.36	0.64	0.64	**0.65**
	STGI	0.66	0.86	0.88	**0.92**
	SI-SDR	−7.80	**6.19**	6.14	3.21
	SDR	−7.54	**8.23**	8.14	6.17
	PESQ	1.24	1.65	1.65	**1.69**

6. Discussion and conclusions

Training DNN-based speech enhancement systems towards maximizing intelligibility and quality metrics in pursuit of improving subjective intelligibility and perceptual quality in real life seems to be a reasonable choice. This motivated us to analyze the performance of, in addition to $\mathcal{L}_{\text{SI-SDR}}$ and its variant integrating speech pre-emphasis $\mathcal{L}_{\text{SI-SDR+PE}}$, a set of five different perceptually-motivated loss functions: $\mathcal{L}_{\text{PMSQE}}$, $\mathcal{L}_{\text{STOI}}$, $\mathcal{L}_{\text{ESTOI}}$, $\mathcal{L}_{\text{STGI}}$ and $\mathcal{L}_{\text{STGI+SI-SDR}}$. While an objective evaluation on the WSJ0 corpus has revealed that these loss functions can

Table 11
Kendall's τ coefficients summarizing processing type-dependent scatter plots representing median intelligibility from Parts A and B of the subjective intelligibility test as a function of the intelligibility scores given by STOI, ESTOI and STGI. In these scatter plots, each data point corresponds to a different noisy condition (i.e., combination of noise type and SNR level).

Processing	Part A			Part B		
	STOI	ESTOI	STGI	STOI	ESTOI	STGI
Noisy	0.913	1.000	1.000	1.000	1.000	0.913
$\mathcal{L}$ SI-SDR	0.667	0.667	0.667	0.667	0.667	0.667
$\mathcal{L}$ SI-SDR+PE	0.667	0.667	1.000	0.667	0.667	0.667
$\mathcal{L}$ STGI+SI-SDR	0.333	0.333	0.333	0.548	0.548	0.548

bring substantial estimated intelligibility and quality gains, particularly, $\mathcal{L}_{\text{STGI+SI-SDR}}$, the gains have been mostly contradicted by a formal intelligibility test with a panel of normal-hearing subjects.

Subjective intelligibility results have pointed out that, generally, DNN speech enhancement processing either maintains or harms intelligibility of the original noisy (i.e., unprocessed) speech. Actually, *only* under one noisy condition, namely, SSN at −5 dB, and operating in a close-to-ideal scenario (i.e., using noisy condition-, language- and speaker-matched models) we have found that speech enhancement is able to provide statistically significant intelligibility improvements. Furthermore, no significant differences have been generally observed among $\mathcal{L}_{\text{SI-SDR}}$, $\mathcal{L}_{\text{SI-SDR+PE}}$ and $\mathcal{L}_{\text{STGI+SI-SDR}}$. This implies that, contradicting our hypothesis supported by preliminary observations, the integration of speech pre-emphasis into a loss function like $\mathcal{L}_{\text{SI-SDR}}$ to give an increased importance to higher frequency components does not help to enhance intelligibility. Besides, we have also seen that all $\mathcal{L}_{\text{SI-SDR}}$, $\mathcal{L}_{\text{SI-SDR+PE}}$ and $\mathcal{L}_{\text{STGI+SI-SDR}}$ perform even worse when dealing with a non-stationary type of noise like café than when facing a stationary noise type (i.e., SSN), which comes as no surprise.

An interesting finding of this study is the low correlation existing – when noise type and SNR are fixed – between subjective intelligibility and intelligibility as estimated by various state-of-the-art speech intelligibility predictors including STGI, which has been shown to outperform many well-established intelligibility metrics across a wide range of degradation conditions (Edraki et al., 2021a). For STOI, ESTOI and STGI, noisy condition-dependent Kendall's τ coefficients have demonstrated a rather weak positive or directly negative association between subjective and objective intelligibility, yet the most important requirement of an intelligibility metric is that it has a strong positive monotonic relationship with subjective intelligibility (Jensen and Taal, 2014). At least when it comes to normal-hearing listeners, this disappointing finding is in line with previous DNN-based speech enhancement work (Gelderblom et al., 2017, 2019). In this way, within the context of a speech enhancement system comparison just based on intelligibility metrics, one might end up choosing a deficient solution to be deployed in real life. On the other hand, it is also important to bear in mind that speech intelligibility metrics such as ESTOI, STGI and to some extent STOI are able to predict the effect of input SNR (Jensen and Taal, 2016; Edraki et al., 2021a; Taal et al., 2011). This statement is in part endorsed by Table 11, which shows Kendall's τ coefficients summarizing processing type-dependent scatter plots (not included due to space limitations) similar to those of Figs. 3 and 5. In these processing type-dependent scatter plots, each data point corresponds to a different noisy condition (namely, combination of noise type and SNR level).

From the above observations, two main conclusions can be drawn. First, unfortunately, subjective evaluation of speech intelligibility cannot systematically be replaced by objective intelligibility evaluation. Secondly, both the development of meaningful intelligibility metrics and DNN-based speech enhancement systems that can consistently enhance the intelligibility of noisy speech for human listening remain open problems. In relation to intelligibility metrics, one of their main issues to be overcome is their poor generalization to speech data not used during the development of these metrics (Kuyk et al., 2018) as a result of their lack of reflection of human perception (Fu et al., 2018). Counting in future on intelligibility metrics highly correlated with subjective intelligibility will be crucial to avoid the need to conduct formal subjective intelligibility tests altogether, since these tests tend to be time-consuming, laborious and expensive. Importantly, this will also contribute towards the development of effective DNN-based speech enhancement systems that are trained employing intelligibility metric-based loss functions.

CRediT authorship contribution statement

Iván López-Espejo: Software, Validation, Investigation, Data curation, Writing, Visualization, Conceptualization, Methodology, Formal analysis. **Amin Edraki:** Scientific input, Writing, Revision. **Wai-Yip Chan:** Methodology, Scientific input, Supervision, Writing, Revision. **Zheng-Hua Tan:** Scientific input, Resources, Supervision, Writing, Revision. **Jesper Jensen:** Conceptualization, Methodology, Scientific input, Resources, Supervision, Writing, Revision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

We would like to thank Dr. Morten Kolbæk for providing us with part of the basis experimental framework as well as for his valuable support during the development of this study. Furthermore, we also want to express our special appreciation to the test subjects who voluntarily contributed to this study by participating in the intelligibility test.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org. URL http://tensorflow.org/.

Allen, J.B., 2005. Articulation and Intelligibility. In: Synthesis Lectures on Speech and Audio Processing, Morgan and Claypool Publishers.

Barker, J., Marxer, R., Vincent, E., Watanabe, S., 2015. The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines. In: Proceedings of ASRU 2015 – IEEE Automatic Speech Recognition and Understanding Workshop. December 13–17, Scottsdale, USA, pp. 504–511.

Barker, J., Marxer, R., Vincent, E., Watanabe, S., 2017. The third 'CHiME' speech separation and recognition challenge: Analysis and outcomes. Comput. Speech Lang. 46, 605–626.

Borgström, B.J., Brandstein, M.S., 2021. Speech enhancement via attention masking network (SEAMNET): An end-to-end system for joint suppression of noise and reverberation. IEEE/ACM Trans. Audio Speech Lang. Process. 29, 515–526.

Braun, S., Tashev, I., 2021. A consolidated view of loss functions for supervised deep learning-based speech enhancement. In: Proceedings of TSP 2021 – 44th International Conference on Telecommunications and Signal Processing. July 26–28, Brno, Czech Republic, pp. 72–76.

Chollet, F., et al., 2015. Keras. https://keras.io.

Cooke, M., 2006b. A glimpsing model of speech perception in noise. J. Acoust. Soc. Am. 119, 1562–1573.

Défossez, A., Synnaeve, G., Adi, Y., 2020. Real time speech enhancement in the waveform domain. In: Proceedings of INTERSPEECH 2020 – 21st Annual Conference of the International Speech Communication Association. October 25–29, Shanghai, China, pp. 3291–3295.

Edraki, A., Chan, W.-Y., Jensen, J., Fogerty, D., 2021. Speech intelligibility prediction using spectro-temporal modulation analysis. IEEE/ACM Trans. Audio Speech Lang. Process. 29, 210–225.
Edraki, A., Chan, W.-Y., Jensen, J., Fogerty, D., 2021a. A spectro-temporal glimpsing index (STGI) for speech intelligibility prediction. In: Proceedings of INTERSPEECH 2021 – 22nd Annual Conference of the International Speech Communication Association. August 30-September 3, Brno, Czechia, pp. 206–210.
Edraki, A., Chan, W.-Y., Jensen, J., Fogerty, D., 2022. Spectro-temporal modulation glimpsing for speech intelligibility prediction. Hear. Res. 426, 1–10.
Févotte, C., Gribonval, R., Vincent, E., 2005. BSS_EVAL toolbox user guide revision 2.0. Publication interne n° 1706.
Fisher, R., 1918. The correlation between relatives on the supposition of Mendelian inheritance. Trans. R. Soc. Edinb. 52, 399–433.
Fu, S.-W., Liao, C.-F., Tsao, Y., 2020. Learning with learned loss function: Speech enhancement with Quality-Net to improve perceptual evaluation of speech quality. IEEE Signal Process. Lett. 27, 26–30.
Fu, S.-W., Wang, T.-W., Tsao, Y., Lu, X., Kawai, H., 2018. End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. 26, 1570–1584.
Garofolo, J.S., Graff, D., Paul, D., Pallett, D., 1993a. CSR-I (WSJ0) complete. https://catalog.ldc.upenn.edu/LDC93s6a.
Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L., Zue, V., 1993b. TIMIT acoustic-phonetic continuous speech corpus. https://catalog.ldc.upenn.edu/LDC93s1.
Gelderblom, F.B., Tronstad, T.V., Viggen, E.M., 2017. Subjective intelligibility of deep neural network-based speech enhancement. In: Proceedings of INTERSPEECH 2017 – 18th Annual Conference of the International Speech Communication Association. August 20–24, Stockholm, Sweden, pp. 1968–1972.
Gelderblom, F.B., Tronstad, T.V., Viggen, E.M., 2019. Subjective evaluation of a noise-reduced training target for deep neural network-based speech enhancement. IEEE/ACM Trans. Audio Speech Lang. Process. 27, 583–594.
Gershenfeld, N., 1988. An experimentalist's introduction to the observation of dynamical systems. In: Directions in Chaos — Volume 2. World Scientific, pp. 310–353.
Goehring, T., Bolner, F., Monaghan, J.J., van Dijk, B., Zarowski, A., Bleeck, S., 2017. Speech enhancement based on neural networks improves speech intelligibility in noise for cochlear implant users. Hear. Res. 344, 183–194.
Hansen, M., Ludvigsen, C., 2001. Dantale II: Danske hagerman sætninger. https://audiologi.dk/wp-content/uploads/2011/05/Dantale-II-rapport1.pdf.
He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In: Proceedings of ICCV 2015 – IEEE International Conference on Computer Vision. December 7–13, Santiago, Chile, pp. 1026–1034.
Healy, E.W., Delfarah, M., Johnson, E.M., Wang, D., 2019. A deep learning algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker and reverberation. J. Acoust. Soc. Am. 145, 1378–1388.
Heidemann Andersen, A., 2017. Speech Intelligibility Prediction for Hearing Aid Systems (Ph.D. thesis). Aalborg University.
ITU-T, 2003. Mapping Function for Transforming P.862 Raw Result Scores to MOS-LQO. Recommendation P.862.1, International Telecommunication Union, Geneva.
ITU-T, 2011. Objective Measurement of Active Speech Level. Recommendation P.56, International Telecommunication Union, Geneva.
Jensen, J., Taal, C.H., 2014. Speech intelligibility prediction based on mutual information. IEEE/ACM Trans. Audio Speech Lang. Process. 22, 430–440.
Jensen, J., Taal, C.H., 2016. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. IEEE/ACM Trans. Audio Speech Lang. Process. 24, 2009–2022.
Kendall, M.G., 1938. A new measure of rank correlation. Biometrika 30, 81–93.
Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization. In: Proceedings of ICLR 2015 – 3rd International Conference on Learning Representations. May 7–9, San Diego, USA.
Knudsen, T.P., 2021. Predictability-Based Objective Evaluation of Sound (Master's thesis). Aalborg University, Denmark.
Kolbæk, M., Tan, Z.-H., Jensen, J., 2017. Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems. IEEE/ACM Trans. Audio Speech Lang. Process. 25, 153–167.
Kolbæk, M., Tan, Z.-H., Jensen, S.H., Jensen, J., 2020. On loss functions for supervised monaural time-domain speech enhancement. IEEE/ACM Trans. Audio Speech Lang. Process. 28, 825–838.
Kruskal, W.H., Wallis, W.A., 1952. Use of ranks in one-criterion variance analysis. J. Amer. Statist. Assoc. 47, 583–621.
Kuyk, S.V., Kleijn, W.B., Hendriks, R.C., 2018. An evaluation of intrusive instrumental intelligibility metrics. IEEE/ACM Trans. Audio Speech Lang. Process. 26, 2153–2166.
Li, H., Yamagishi, J., 2021. Multi-metric optimization using generative adversarial networks for near-end speech intelligibility enhancement. IEEE/ACM Trans. Audio Speech Lang. Process. 29, 3000–3011.
Loizou, P.C., 2013. Speech Enhancement: Theory and Practice, second ed. CRC Press.
Loizou, P.C., Kim, G., 2011. Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions. IEEE Trans. Audio Speech Lang. Process. 19, 47–56.
Luo, Y., Mesgarani, N., 2019. Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation. IEEE/ACM Trans. Audio Speech Lang. Process. 27, 1256–1266.
Martin-Doñas, J.M., Gomez, A.M., Gonzalez, J.A., Peinado, A.M., 2018. A deep learning loss function based on the perceptual evaluation of the speech quality. IEEE Signal Process. Lett. 25, 1680–1684.
Martín-Doñas, J.M., Gomez, A.M., López-Espejo, I., Peinado, A.M., 2017. Dual-channel DNN-based speech enhancement for smartphones. In: Proceedings of MMSP 2017 – 19th International Workshop on Multimedia Signal Processing. October 16–18, Luton, UK.
Massey, F.J., 1951. The Kolmogorov–Smirnov test for goodness of fit. J. Amer. Statist. Assoc. 46, 68–78.
Nasjonalbiblioteket, 2011. Akustiske databaser for dansk. https://www.nb.no/sbfil/dok/nst_taledat_dk.pdf.
Pandey, A., Wang, D., 2019. TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain. In: Proceedings of ICASSP 2019 – 44th IEEE International Conference on Acoustics, Speech and Signal Processing. May 12–17, Brighton, UK, pp. 6875–6879.
Rix, A., Beerends, J., Hollier, M., Hekstra, A., 2001. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In: Proceedings of ICASSP 2001 – 26th IEEE International Conference on Acoustics, Speech and Signal Processing. May 7–11, Salt Lake City, USA, pp. 749–752.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. In: Proceedings of MICCAI 2015 – 18th Medical Image Computing and Computer-Assisted Intervention. October 5–9, Munich, Germany, pp. 234–241.
Roux, J.L., Wisdom, S., Erdogan, H., Hershey, J.R., 2019. SDR - half-baked or well done? In: Proceedings of ICASSP 2019 – 44th IEEE International Conference on Acoustics, Speech and Signal Processing. May 12–17, Brighton, UK, pp. 626–630.
Schädler, M.R., Meyer, B.T., Kollmeier, B., 2012. Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition. J. Acoust. Soc. Am. 131, 4134–4151.
Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J., 2011. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. IEEE/ACM Trans. Audio Speech Lang. Process. 19, 2125–2136.
Tan, K., Zhang, X., Wang, D., 2019. Real-time speech enhancement using an efficient convolutional recurrent network for dual-microphone mobile phones in close-talk scenarios. In: Proceedings of ICASSP 2019 – 44th IEEE International Conference on Acoustics, Speech and Signal Processing. May 12–17, Brighton, UK, pp. 5751–5755.
Vuong, T., Xia, Y., Stern, R.M., 2021. A modulation-domain loss for neural-network-based real-time speech enhancement. In: Proceedings of ICASSP 2021 – 46th IEEE International Conference on Acoustics, Speech and Signal Processing. June 6–11, Toronto, Canada, pp. 6643–6647.
Wagener, K., Josvassen, J.L., Ardenkjær, R., 2003. Design, optimization and evaluation of a Danish sentence test in noise: Diseño, optimización y evaluación de la prueba Danesa de frases en ruido. Int. J. Audiol. 42, 10–17.
Wang, D., Chen, J., 2018. Supervised speech separation based on deep learning: An overview. IEEE/ACM Trans. Audio Speech Lang. Process. 26, 1702–1726.
Wang, Z.-Q., Wichern, G., Roux, J.L., 2021. On the compensation between magnitude and phase in speech separation. IEEE Signal Process. Lett. 28, 2018–2022.
Wang, Y., Yu, G., Wang, J., Wang, H., Zhang, Q., 2020. Improved relativistic cycle-consistent GAN with dilated residual network and multi-attention for speech enhancement. IEEE Access 8, 183272–183285.
Wilcoxon, F., 1945. Individual comparisons by ranking methods. Biom. Bull. 1, 80–83.
Xiang, X., Zhang, X., Chen, H., 2021. A convolutional network with multi-scale and attention mechanisms for end-to-end single-channel speech enhancement. IEEE Signal Process. Lett. 28, 1455–1459.
Xu, Y., Du, J., Dai, L.-R., Lee, C.-H., 2014. A regression approach to speech enhancement based on deep neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. 23, 7–19.
Xu, Z., Jiang, T., Li, C., Yu, J., 2021. An attention-augmented fully convolutional neural network for monaural speech enhancement. In: Proceedings of ISCSLP 2021 – 12th International Symposium on Chinese Spoken Language Processing. January 24–27, Hong Kong, China.
Zhang, Z., Li, X., Li, Y., Dong, Y., Wang, D., Xiong, S., 2021. Neural noise embedding for end-to-end speech enhancement with conditional layer normalization. In: Proceedings of ICASSP 2021 – 46th IEEE International Conference on Acoustics, Speech and Signal Processing. June 6–11, Toronto, Canada, pp. 7113–7117.
Zhang, H., Zhang, X., Gao, G., 2018. Training supervised speech separation system to improve STOI and PESQ directly. In: Proceedings of ICASSP 2018 – 43rd IEEE International Conference on Acoustics, Speech and Signal Processing. April 15–20, Calgary, Canada, pp. 5374–5378.
Zhao, Y., Wang, Z.-Q., Wang, D., 2018a. Two-stage deep learning for noisy-reverberant speech enhancement. IEEE/ACM Trans. Audio Speech Lang. Process. 27, 53–62.
Zhao, Y., Xu, B., Giri, R., Zhang, T., 2018b. Perceptually guided speech enhancement using deep neural networks. In: Proceedings of ICASSP 2018 – 43rd IEEE International Conference on Acoustics, Speech and Signal Processing. April 15–20, Calgary, Canada, pp. 5074–5078.